DeepCOVIDExplainer: Explainable COVID-19 Diagnosis from Chest X-ray Images

Md. Rezaul Karim*†, Till Döhmen*†, Michael Cochez‡, Oya Beyan*†, Dietrich Rebholz-Schuhmann§, Stefan Decker*†
* Fraunhofer Institute for Applied Information Technology FIT, Germany
† Information Systems and Databases, RWTH Aachen University, Germany
‡ Information Center for Life Sciences (ZB MED), German National Library of Medicine, Germany
§ Department of Computer Science, Vrije Universiteit Amsterdam, the Netherlands

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313304

***Abstract*—In this paper[1], we proposed an explainable deep neural networks (DNN)-based method for automatic detection of COVID-19 symptoms from chest radiography (CXR) images, which we call *'DeepCOVIDExplainer'*. We used 15,959 CXR images of 15,854 patients, covering normal, pneumonia, and COVID-19 cases. CXR images are first comprehensively preprocessed and augmented before classifying with a neural ensemble method, followed by highlighting class-discriminating regions using gradient-guided class activation maps (Grad-CAM++) and layer-wise relevance propagation (LRP). Further, we provide human-interpretable explanations for the diagnosis. Evaluation results show that our approach can identify COVID-19 cases with a positive predictive value (PPV) of 91.6%, 92.45%, and 96.12%, respectively for normal, pneumonia, and COVID-19 cases, respectively, outperforming recent approaches.**

***Index Terms*—COVID-19, Biomedical imaging, Deep learning, Explainability, Grad-CAM, Layer-wise relevance propagation.**

I. Introduction

The ongoing coronavirus pandemichas already created a devastating impact on the health and well-being of the global population [1,2]. Recent studies show that COVID-19, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), often, but by no means exclusively, affects elderly persons with pre-existing medical conditions [3–7]. While hospitals are struggling with scaling up capacities to meet the rising number of patients, it is crucial to make use of the screening methods at hand to identify COVID-19 cases and discriminate them from other conditions [1]. The definitive test for COVID-19 is the reverse transcriptase-polymerase chain reaction (RT-PCR) test, which requires specialized laboratories. COVID-19 patients, however, show several unique clinical and para-clinical features, e.g., presenting abnormalities in medical chest imaging with commonly bilateral involvement. Such features that are observable on CXR images and CT scans [6] are only moderately characteristic to the human eye [7] and not easy to distinguish from pneumonia.

AI-based techniques have been utilized in numerous scenarios, including automated diagnoses and treatment in clinical settings. Deep neural networks (DNNs) have been employed for the diagnosis of COVID-19 from medical images, leading to promising results [1,6–9]. However, many current approaches are "black box" methods without providing insights into the decisive image features. Let's imagine a situation where resources are scarce, e.g., a hospital runs out of confirmatory tests or necessary radiologists are occupied, where AI-assisted tools could potentially help less-specialized general practitioners to triage patients, by highlighting critical chest regions to lead automated diagnosis decision [1].

Although COVID-19 diagnosis approaches [1,3,4,8–11] proposed recently looks promising when compared to expert radiologist a lower sensitivity in most of the cases, the reliability can be questioned for three main reasons. The datasets used are severely biased due to a deficient number of COVID-19 cases. Moreover, some results are not statistically reliable and lack of decision biases as the diagnoses were mostly based on a single model. Nevertheless, less accurate localization and visualization of critical chest regions. To overcome the shortcomings and as a quick step towards an AI-based COVID-19 diagnosis, we propose *'DeepCOVIDExplainer'*, a novel diagnosis approach based on neural ensemble method. Based on the following hypotheses, DeepCOVIDExplainer focuses on fairness, algorithmic transparency, and explainability:

- Based on majority voting from a panel of independent radiologists (i.e., ensemble), we get the final prediction fair and trustworthy than a single radiologist.
- By localizing class-discriminating regions with Grad-CAM++ and LRP, we not only can mitigate the opaqueness of the black-box model by providing more human-interpretable explanations of the predictions, but also identify the critical regions on patients chest.

II. Methods

The pipeline of DeepCOVIDExplainer starts with a comprehensive preprocessing of CXR images, followed by training of DenseNet, ResNet, and VGGNet architectures, and creating respective model snapshots. To incorporate the trained model into an ensemble, both Softmax class posterior averaging (SCPA) and prediction maximization (PM) are utilized. Finally, class-discriminating attention maps are generated using Grad-CAM++ and LRP to provide explanations and to identify critical regions on patients chest.

A. Preprocessing

Since radiographs usually have dark edges, images with such distinctly darker regions impact the classification. We

[1] Read longer version of this paper: https://arxiv.org/pdf/2004.04582.pdf

978-1-7281-6215-7/20/$31.00 ©2020 IEEE

perform global contrast enhancement, edge enhancement, and noise elimination on entire CXR images with histogram equalization and unsharp masking edge enhancement [12].

B. Network construction and training

We train VGG-16/19, ResNet-18/34, and DenseNet-161/201 architectures and create their snapshots during a single run with cyclic cosine annealing (CAC)[13], followed by combining their predictions to an ensemble prediction [14,15]. We do not initialize network weights with pretrained (as general object like shapes are not present in CXR images [11]) models. We set number of epochs (NE), learning rate (LR), number of cycles (NC), and current epoch number. CAC starts with a large LR and rapidly decreases to a minimum value before it dramatically increases systematically over epochs to produce different network weights [14]:

$$\alpha(t) = \frac{\alpha_0}{2}\left(\cos\left(\frac{\pi \bmod (t-1, \lceil T/C \rceil)}{\lceil T/C \rceil}\right) + 1\right), \quad (1)$$

where $\alpha(t)$ is the LR at epoch t, α_0 is the maximum LR, T is the total epoch, C is the number of cycles and mod is the modulo operation. After training a network for C cycles, best weights at the bottom of each cycle are saved as a model snapshot (m), giving M model snapshots, where $m \leq M$.

C. Model ensemble

When a single radiologist makes a COVID-19 diagnosis, the chance of a false diagnosis is high. Therefore, it is reasonable to ask for a second or third opinions. We employ neural ensemble to combine the 'expertise' of multiple models into a consolidated prediction [14], as neural ensemble method by combining several deep architectures is more effective than structures solely based on a single model [14,15]. We apply both SCPA and PM of best-performing snapshot models, ensemble their predictions, and propagate them through the Softmax layer, where the class probability of the ground truth j for a given image x is inferred as [16]:

$$P(y=j|\mathbf{x}) = \frac{\exp\left[\sum_{m=1}^{M} \hat{P}_m(y=j|\mathbf{x})\right]}{\exp\left[\sum_{k=1}^{K}\sum_{m=1}^{M} \hat{P}_m(y=k|\mathbf{x})\right]}, \quad (2)$$

where m is the last snapshot model from M, K is the number classes, and $\hat{P}_m(y=j|\mathbf{x})$ is the probability distribution.

D. Decision visualizations

To improve diagnosis transparency, critical chest regions are localized with Grad-CAM [17], Grad-CAM++ [18], and LRP [19]. The idea is to explain where the model provides more attention for the classification in terms of heatmaps, indicating the relevance for the classification decision.

III. Experiments

During the training[2], we set the NE to 200, maximum LR to 1.0, and NC to 20, giving 20 snapshots for each model and 120 snapshot models in total, on which we construct the ensemble model. The best snapshot model, which is used for the decision visualizations is chosen using WeightWatcher [20].

[2] Source code: https://github.com/rezacsedu/DeepCOVIDExplainer

A. Datasets

We used 15,959 CXR images[3] that are categorized into normal (8,066 images), pneumonia (5,538 images), and COVID-19 cases (i.e., 358 CXR images) covering 15,854 patients.

B. Performance of individual model

As summarized in table I, VGG-19 and DenseNet-161 performed best on both balanced and imbalanced datasets, albeit i) VGG-19 outperforms VGG-16, and ii) ResNet-18 performed better than ResNet-34. DenseNet-161 outperforms other models, giving precision, recall, and F1 scores of 0.952, 0.945, and 0.945, respectively, on balanced CXR images. On imbalanced dataset, both DenseNet-161 and ResNet-18 perform consistently. Although VGG-19 and ResNet-18 show competitive results on balanced dataset, the misclassification rate for normal and pneumonia samples are slightly elevated than DenseNet-161, which poses a risk for clinical diagnosis. While DenseNet-161 is found to be resilient against class imbalanced, making it better suited for the clinical setting.

C. Model ensemble

We perform the ensemble on following top-3 models: VGG-19, ResNet-18, and DenseNet-161. As demonstrated in table II, the ensemble based on the SCPA method moderately outperforms the ensemble based on the PM method. The reason is that the PM approach appears to be easily influenced by outliers with high scores. For the SCPA-based ensemble, the combination of VGG-19 and DenseNet-161 outperforms other ensemble combinations. Results show that a majority of samples were classified correctly, with precision, recall, and F1 scores of 0.937, 0.926, and 0.931, respectively, using the PM ensemble method. The SCPA-based ensemble yields slightly higher precision, recall, and F1 of 0.946, 0.943, and 0.945, respectively. We report the class-specific measures in table III.

D. Quantitative analysis

Since we primarily want to limit the number of missed COVID-19 instances, a recall of 90.5% is still an acceptable metric compared to 91% by Wang et al. [1]. To determine how many of all infected persons would be diagnosed positively, we calculate the PPV, where out of 129 COVID-19 samples, only 3 were misclassified as pneumonia and two as normal, giving a PPV of 96.12% for COVID-19 cases. This is still an acceptable metric compared to 98.9% by Wang et al. [1].

In a setting with high COVID-19 prevalence, the likelihood of false-positives is expected to reduce in favor of correct COVID-19 predictions. Our results are backed up by i) a larger test set, ii) better localization and explanation capability, which contributes to the reliability of our evaluation results, given the fact that in healthcare predicting something with high confidence only is not enough, but requires trustworthiness.

[3] Refer to https://github.com/rezacsedu/DeepCOVIDExplainer for the detail.

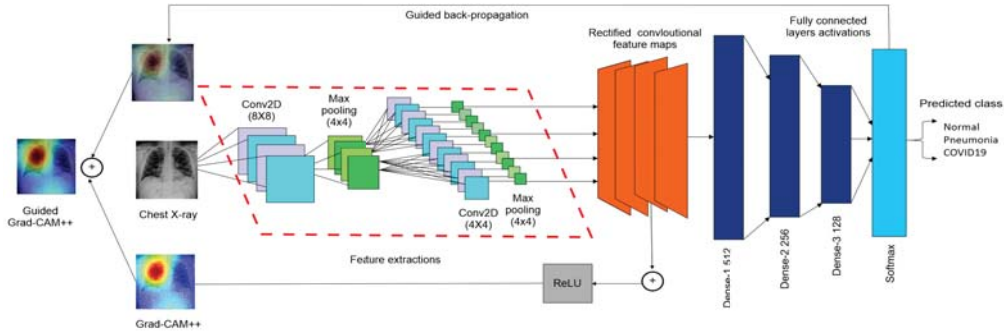

Fig. 1: Classification and decision visualization with CNN-based approach

TABLE I: Classification results of each model on balanced and imbalanced datasets

	Balanced dataset			Imbalanced dataset		
Network	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
VGG-16	0.783	0.771	0.761	0.734	0.753	0.737
ResNet-34	0.884	0.856	0.861	**0.852**	**0.871**	**0.851**
DenseNet-201	0.916	0.905	0.905	0.805	0.773	0.826
ResNet-18	0.924	0.925	0.921	0.873	0.847	0.852
VGG-19	**0.943**	**0.935**	**0.925**	0.862	0.848	0.845
DenseNet-161	**0.952**	**0.945**	**0.948**	**0.893**	**0.874**	**0.883**

TABLE II: Classification results for ensemble methods on balanced dataset

	Prediction maximization			Softmax posterior averaging		
Architecture combination	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ResNet-18+DenseNet-161	0.915	0.924	0.928	0.925	0.94	0.933
VGG-19+DenseNet-161	**0.937**	**0.926**	**0.931**	**0.946**	**0.943**	**0.945**
VGG-19+ResNet-18	0.917	0.923	0.912	0.923	0.945	0.934
DenseNet-161+VGG-19+ResNet-18	0.926	0.901	0.901	0.924	0.937	0.935

TABLE III: Classwise classification results of ensemble model on chest x-rays

	Balanced dataset			Imbalanced dataset		
Infection type	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Normal	0.942	0.927	0.935	0.906	0.897	0.902
Pneumonia	0.916	0.928	0.922	0.864	0.853	0.858
COVID-19	0.904	0.905	0.905	0.877	0.881	0.879

E. COVID-19 predictions and explanations

Critical regions of some CXR images of COVID-19 cases are highlighted in fig. 2, fig. 3, and fig. 4, where class-discriminating areas within the lung are localized. As seen, HM generated by Grad-CAM and Grad-CAM++ are fairly consistent and alike, but those with Grad-CAM++ are more accurately localized, i.e., instead of certain parts, Grad-CAM++ highlights conjoined features more precisely. Although LRP highlights regions much more precisely, it fails to provide attention to critical regions. It turned out that Grad-CAM++ generates the most reliable HM when compared to Grad-CAM and LRP. Let's consider the following examples:

- **Example 1**: CXR image is classified to contain a confirmed COVID-19 case with a probability of 58%, the true class is COVID-19, as shown in fig. 2.
- **Example 2**: CXR image is classified to contain a confirmed COVID-19 case with a probability of 58%, the true class is COVID-19, as shown in fig. 3.
- **Example 3**: CXR image is classified to contain COVID-19 case with a classification score of 10.5, the true class is COVID-19, as shown in fig. 4.

F. Discussion and diagnosis recommendations

Even if a specific approach does not perform well, an ensemble of several models still may outperform individual models. However, models trained on imbalanced training data may provide distorted or wrong predictions. In this case, even a high accuracy score can be achieved without predicting minor classes, hence might be uninformative. Thirdly, the risk resulting from a pneumonia diagnosis is much lower than for a COVID-19 diagnosis. Hence, it is more reasonable to make a decision based on the maximum score from individual model predictions. Fourthly, decision visualizations cannot be provided based on ensemble models, albeit their usage contributes to decision reliability. Therefore, it is recommended to pick the single best model as a basis and to employ Grad-CAM++ for providing the most reliable localization.

IV. Conclusion and Outlook

In this paper, we proposed *'DeepCOVIDExplainer'* to leverage explainable COVID-19 prediction based on CXR images. Evaluation results show that our approach can identify COVID-19 with a PPV of 96.12% and recall of 94.3%, outperforming a recent approach. We would argue *'Deep-*

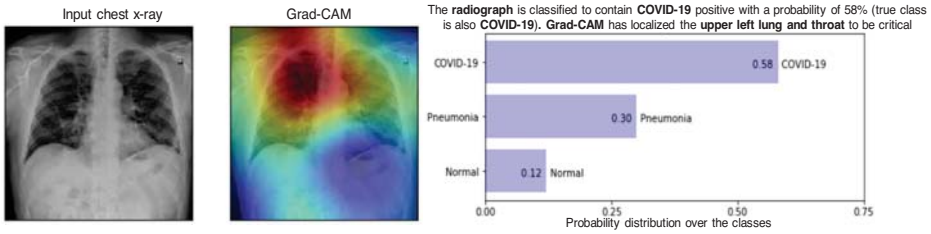

Fig. 2: Classification of CXR with DenseNet-161, decision visualization (Grad-CAM), and human-interpretable explanation

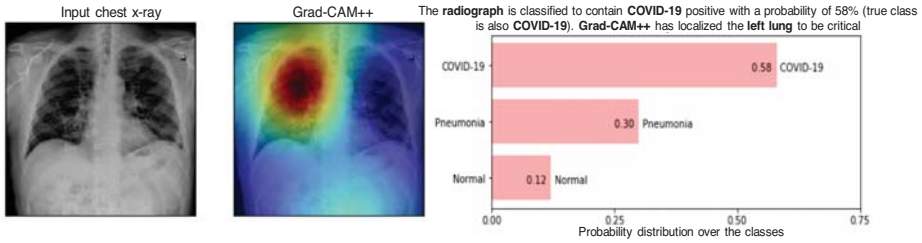

Fig. 3: Classification of CXR with DenseNet-161, decision visualization (Grad-CAM++), and human-interpretable explanation

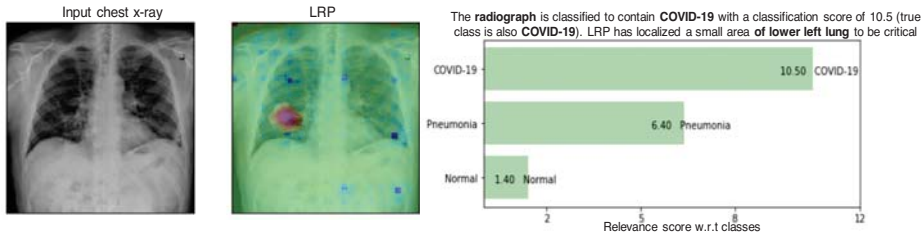

Fig. 4: Classification of CXR with DenseNet-161, decision visualization (LRP), and human-interpretable explanation

COVIDExplainer' with no means a replacement for a human radiologist. On a serious note: due to limited number of CXR images used to train the models, it would be unfair to claim that we can rule out overfitting for our models. Besides, we were yet not been able to verify the diagnosis and localization accuracies with radiologists. In future, we intend to overcome these limitations with a multimodal learning (i.e., based on CT scans, CXR, and clinical notes) outcomes.

REFERENCES

[1] 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:2003.09871*, 2020.

[2] O. Gozes and E. Siegel, "Rapid AI development cycle for coronavirus pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis," *arXiv:2003.05037*, 2020.

[3] K. M. Yee, "X-ray may be missing COVID cases found with CT," *Korean Journal of Radiology*, pp. 1–7, 2020.

[4] Y. Fang, H. Zhang, and W. Ji, "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," *Radiology*, p. 200432, 2020.

[5] T. Ai, Z. Yang, H. Hou, C. Zhan, and L. Xia, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," *Radiology*, p. 200642, 2020.

[6] C. Huang, Y. Wang, X. Li, L. Ren, and X. Gu, "Clinical features of patients infected with novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[7] M.-Y. Ng, E. Y. Lee, J. Yang, and P.-L. Khong, "Imaging profile of COVID-19 infection: radiologic findings and literature review," *Cardiothoracic Imaging*, vol. 2, no. 1, 2020.

[8] T. Ozturk and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Computers in Biology and Medicine*, p. 103792, 2020.

[9] S. Tabik, A. Gómez-Ríos, M. Valero-González *et al.*, "COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images," *arXiv preprint arXiv:2006.01409*, 2020.

[10] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease using X-rays and cnn," *arXiv:2003.10849*, 2020.

[11] M. R. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, and O. Beyan, "DeepCOVIDExplainer: Explainable COVID-19 Diagnosis Based on Chest X-ray Images," *arXiv:2004.04582*, 2020.

[12] S. S. Pathak, P. Dahiwale, and G. Padole, "A combined effect of local and global method for contrast image enhancement," in *International Conference on Engineering & Technology*. IEEE, 2015, pp. 1–5.

[13] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," *arXiv:1608.03983*, 2016.

[14] G. Huang, Y. Li, G. Pleiss, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get m for free," *arXiv:1704.00109*, 2017.

[15] M. R. Karim, S. Decker, and O. Beyan, "A Snapshot Neural Ensemble Method for Cancer-type Prediction Based on Copy Number Variations," *Neural Computing and Applications*, pp. 1–19, 2019.

[16] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach," *Scientific reports*, vol. 8, no. 1, p. 1727, 2018.

[17] R. R. Selvaraju, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE ICCV*, 2017, pp. 618–626.

[18] A. Chattopadhay and A. Sarkar, "Grad-CAM++: Generalized gradient-based visual explanations for convolutional networks," in *Applications of Computer Vision(WACV)*. IEEE, 2018, pp. 839–847.

[19] B. K. Iwana, R. Kuroki, and S. Uchida, "Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation," *arXiv:1908.04351*, 2019.

[20] C. H. Martin and M. W. Mahoney, "Traditional and heavy-tailed self regularization in neural network models," *arXiv:1901.08276*, 2019.