








POTHER: Patch-Voted Deep Learning-Based Chest X-ray Bias Analysis for COVID-19 Detection

Tomasz Szczepański¹ , Arkadiusz Sitek² , Tomasz Trzciniński^{1,3,4} ,
and Szymon Płotka^{1,2}  

¹ Warsaw University of Technology, Warsaw, Poland

² Sano Centre for Computational Medicine, Cracow, Poland
s.plotka@sanoscience.org

³ Jagiellonian University, Cracow, Poland

⁴ Tooploox, Wroclaw, Poland

Abstract. A critical step in the fight against COVID-19, which continues to have a catastrophic impact on peoples lives, is the effective screening of patients presented in the clinics with severe COVID-19 symptoms. Chest radiography is one of the promising screening approaches. Many studies reported detecting COVID-19 in chest X-rays accurately using deep learning. A serious limitation of many published approaches is insufficient attention paid to explaining decisions made by deep learning models. Using explainable artificial intelligence methods, we demonstrate that model decisions may rely on confounding factors rather than medical pathology. After an analysis of potential confounding factors found on chest X-ray images, we propose a novel method to minimise their negative impact. We show that our proposed method is more robust than previous attempts to counter confounding factors such as ECG leads in chest X-rays that often influence model classification decisions. In addition to being robust, our method achieves results comparable to the state-of-the-art. The source code and pre-trained weights are publicly available at (<https://github.com/tomek1911/POTHER>).

Keywords: COVID-19 · Deep learning · Explainable AI

1 Introduction

The SARS-CoV-2 outbreak has claimed the lives of millions of people, and despite its onset in 2019, it remains a serious concern and a threat to public health. The gold standard for diagnosing COVID-19 disease is an RT-PCR test; however, it is expensive, necessitates specialised laboratories and requires the patient to wait relatively long for the outcome. For this reason, computer scientists and radiologists become interested in the computer-aided diagnosis (CAD) capabilities of chest X-rays (CXR). The automatic diagnosis of COVID-19 using chest X-ray images is challenging due to the high intra-class variations, superimposition of anatomical

structures, or implanted electronic devices [1]. Moreover, a significant limitation in developing reliable models for detecting pneumonia and COVID-19 is the lack of precisely annotated and rigorously collected data. These limitations may result in models learning various confounding factors in the CXRs. We investigate this phenomenon in our work.

Most of the approaches developed for the classification of the CXR images rely on global features [2, 11, 12, 14]. However, these features may not accurately represent the complex nature of CXR images [22]. It should be verified when working with medical images, especially when the model's accuracy is very high [13, 21], the reasons why algorithms perform so well to prevent developing algorithms which base their decision on confounding factors rather than medical pathology. Wang *et al.* [7] propose COVID-Net, the first lightweight capacity neural network dedicated for COVID-19 detection and introduce a novel COVIDx dataset. Authors create a feature extraction method tailored for COVID-19 and provide results of the model's decisions analysis with the use of GSInquire [24]. They create a map of lung areas important for COVID-19 detection and claim a production-ready solution. Considering the need for a critical approach to chest classification and analysis of biases [20, 22, 23], we carry out further research into the analysis of the deep learning model decisions using the COVIDx dataset.

Recently, a new patch-based learning technique [3, 10, 25] emerged as a successful method for robust model learning and generalization. Li *et al.* [3] propose a multi-resolution patch-based CNN approach for lung nodule detection on CXR image. The method achieves high accuracy by leveraging the local context of CXR images. Roy *et al.* [25] use the model's local training method for classification on the ICIAR-2018 breast cancer histology image dataset, which allows them to achieve state-of-the-art results on this dataset. Oh *et al.* [10], inspired by their statistical analysis of the potential imaging biomarkers of the CXRs, explore patch-based learning for COVID-19 detection. They propose random patch cropping, from which the final classification result is obtained by majority voting from inference results at multiple patch locations. Authors analyse model decisions using proposed probability gradient-weighted class activation map (GradCAM) [18] as a part of explainable AI (XAI) method [4], upon which they conclude their results are correlated with radiological findings. Similar to Oh *et al.*, our proposed method POTHER explores a patch-based learning approach in CXR images for reliably detecting COVID-19. In contrast, we do not use segmented lungs as input; instead, we propose a multi-task model that leverages segmentation task to extract valuable features. In addition, we limit the area from which we draw patches and reduce their size.

We use XAI methods to demonstrate that model decisions may rely on *confounding factors* rather than medical pathology. Degraeve *et al.* [20] and Cabrera *et al.* [22] call them shortcuts, while Maguolo *et al.* [23] biases. An analysis of shortcuts in CXRs, which we call *confounding biases* (CBs), based on open-source datasets and global learning methods, is presented by [20]. It demonstrates the detrimental influence on models decisions caused by laterality tokens, i.e. L or R letter, meaning the left or right side of the image, the position of the clavicles and

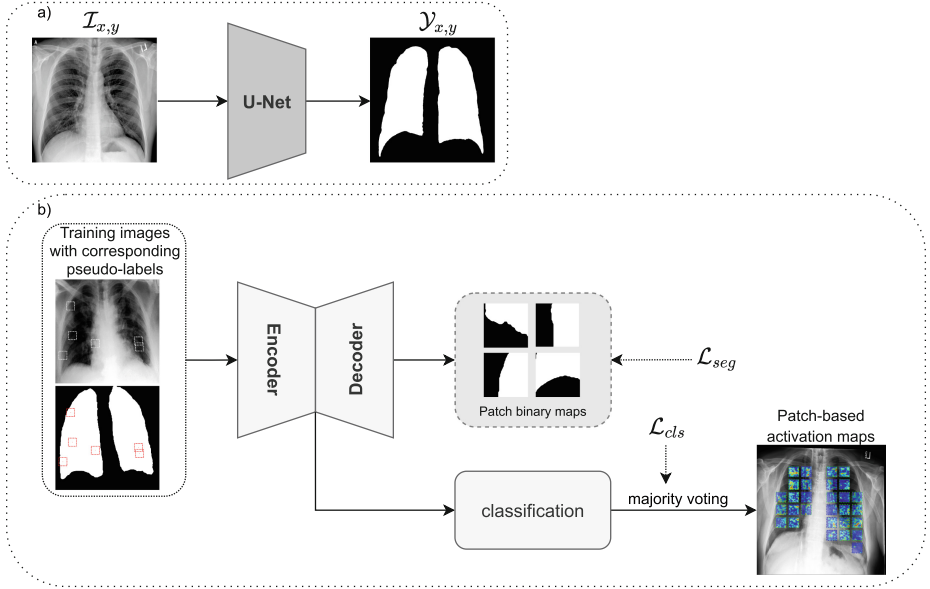


Fig. 1. The overview of the proposed POTHER framework for chest X-ray bias analysis. (a) Pre-trained segmentation network to generate pseudo-labels, (b) an encoder-decoder patch-based multi-task learning network to classify CXR images. We adopt majority patch-based voting for classification and employ patch-based activation maps to explain the results.

the presence of arms in the upper parts of the image. Authors present that machine learning (ML) models trained on CXR images may generalise poorly and owe the majority of their performance to the learning of shortcuts that may be consistently detected in both internal and external domains - making external validation alone insufficient to detect poorly behaved models. We analyse CBs in the frequently cited COVIDx dataset [7] and propose a multi-task learning COVID-19 detection method called POTHER that improves robustness to CBs.

The contributions of this paper are threefold. Firstly, we propose a novel multi-task patch-voted learning-based method called POTHER for chest X-ray image analysis for COVID-19 detection. Secondly, we analyse activation maps of currently available methods and sources of confounding biases in CXRs from the COVIDx dataset. Our analysis reveals a significant number of CBs. These biases should be considered when using the COVIDx dataset in future research. The main CBs are ECG leads, laterality tokens and hospital markings. Finally, we demonstrate that deep learning models learn how to classify pneumonia manifestations and lung morphology, focusing on the confounding mentioned above factors rather than the actual manifestation of the disease or lack thereof on lung parenchyma. In our methods, to counter CBs, we used segmentation as a helper task which allowed for an efficient feature extraction method that performs comparably to the state of the art. The paper is organized as follows. Section 2 details

our proposed method. Next, Sect. 3 presents details about our experimental setup and results. We follow with Sect. 4 which discusses the results. Finally, Sect. 5 concludes the paper.

2 Method

Figure 1 presents an overview of our proposed method. First, we pre-train the U-Net lung fields segmentation network to generate pseudo-labels. Second, we use patches to train a multi-task U-Net-based neural network for CXR image bias analysis. Finally, for classification, we use a majority patch-based voting.

Pre-training Segmentation Network. We train a segmentation neural network to extract lung fields from CXR images. For this task, we adopt the U-Net encoder-decoder to perform semantic segmentation [5]. With the pre-trained model on [8, 9] datasets, we generate pseudo labels for unlabelled COVIDx dataset [7]. Before feeding those pseudo masks to the multi-task neural network, we use custom pre-processing algorithms.

Multi-task Neural Network. Inspired by [15–17], we use an encoder-decoder based convolutional neural network (CNN) for simultaneous classification and segmentation of the lungs in the CXRs. Our network is U-Net-based with ImageNet pre-trained ResNet-50 as backbone encoder [6]. We use an encoder part to extract high-level X-ray image features. Every encoder block forwards feature maps and concatenates them with the corresponding decoder part. We employ an attention mechanism to the skip connections to learn salient maps suppressing irrelevant lung vicinity that may be a source of CBs. We extend an attention gate mechanism by aggregating multi-scale feature maps from the decoder to learn the local context of the lung feature maps representation. Each feature map is fed through an inception module that leverages convolutional filters of multiple kernels (i.e. 1×1 , 3×3 , 5×5 , and 7×7) and stride sizes ($S = 1$, $S = 2$), which does not increase the number of parameters significantly. The amplitudes of features from the deeper layers are smaller than the shallow ones. To prevent shallow layers from dominating deeper ones, we normalise the weights of the features from multiple scales with the L2 norm before concatenation.

Patch-Based Learning. Motivated by [10], we adopt a patch-based learning method to train our multi-task neural network. Unlike Oh, we do not cut out the lungs with masks to avoid inductive bias. We use the lung masks instead as pseudo-labels for segmentation training. This helper task requires the model to learn new features necessary to recognise lung tissue boundaries. A pre-processed CXR image and corresponding mask are resized to 1024×1024 . Then using a draw area based on the scaled-down, with a ratio of 0.9, whole lung mask, we randomly choose a patch centre. Its coordinates are drawn with a uniform distribution of non-zero pixels from the draw area mask. Figure 1(b) shows training images with corresponding pseudo-labels in the context of the entire CXR image,

i.e. exemplary drawn image patches are marked with white squares in the CXR image and pseudo-labels with red squares on the corresponding mask. In our method, we use patches of size 80×80 , in contrast to Oh uses patches of size 224×224 . A region of the image and the corresponding mask are cut off using determined patch coordinates, resulting in an image patch and pseudo-label pair. Then the image patch and its mask are interpolated to 224×224 . This pair is used to train a segmentation head, while the image patch and its corresponding class represent a training pair for the classification head. Thanks to the reduction of the draw area and the small size of the patch, the model’s input never contains laterality tokens and hospital markings. At the same time, if chosen close to the lung edge, a patch covers a small lung boundary fragment, allowing the model to recognise lung tissue based on the segmentation task.

Majority Patch-Based Voting. Only one patch per image, per batch, is used in the training phase. However, we repeat the random draw multiple times for the inference to cover the whole lung field as the area of interest. Each time, based on the patch, the model makes a single classification called a single vote. The majority vote result, i.e. the class chosen based on the majority of the patches, is the final classification for that image.

3 Experiments and Results

In this section, we evaluate our methods on the COVIDx dataset. Next, we present gradient-weighted class activation maps that compare the decision bases of our proposed method with existing methods. We employ GradCAM to provide visual explanations of our method decision bases that focuses its attention on lung morphology and is less sensitive to CBs. Finally, we show quantitative results on the COVIDx test set.

3.1 Datasets

To pre-train a neural network for the segmentation of lungs, we use data from [8,9]. The datasets consist of 6380 2D CXR images. The annotations for each image provide a manual segmentation mask rendered as polygons, including the retrocardiac region. To evaluate our methods, we use an open-source **COVIDx dataset** [7]. In total, the dataset consists of 13970 2D CXR images (8806 normal, 5551 pneumonia and 353 COVID-19 cases). The authors constantly expand this dataset, and in order to compare with their published results, we use the same version of the dataset as they used at the time of publication. We create a validation set by splitting the training set with a 70:30 ratio, and the test dataset is built using the script provided by the COVIDx authors.

3.2 Data Pre-processing

Lung masks. We remove unnecessary objects from lungs masks like electronic devices, which are labelled with a shade of grey. We perform mask filtering to

improve the masks or remove them from the dataset when the correction is not possible. We present a detailed algorithm in the supplement¹. Finally, we resize the original image and the mask to a size of 1024×1024 pixels using linear interpolation. All images are resized to square regardless of their original aspect ratio.

COVIDx Dataset. The images for the training of our model are pre-processed with histogram equalisation. Soft augmentation methods consist of offset, scaling without preserved aspect ratio, rotation, horizontal flip, Contrast Limited Adaptive Histogram Equalisation (CLAHE) with random clip range and random grid range, brightness and contrast adjustment, sharpening and embossing.

3.3 Implementation Details

We implement our model in PyTorch deep learning framework and train on a workstation with a single GPU NVIDIA Titan RTX 24 GB until convergence over 100 epochs, with a mini-batch size of 16, an initial learning rate of 1×10^{-4} and a weight decay factor of 1×10^{-4} . We set Rectified Adam (RAdam) as the optimiser to minimise the loss function. To prevent overfitting, we apply various soft data augmentation techniques. During training, we perform the following transformations: horizontal flip, sharpen, emboss and CLAHE with $p = 0.5$. We also apply a random-weighted sampler. The weights are computed as an inverse class frequency. As loss function, we use Dice loss for segmentation task:

$$\mathcal{L}_{dice} = 1 - \frac{2 \sum_i^N p_i g_i + \epsilon}{\sum_i^N p_i^2 + \sum_i^N g_i^2 + \epsilon}, \quad (1)$$

where p_i is the prediction pixel value, g_i is the ground truth, and ϵ is a numerical stability to avoid divide by zero errors, and Weighted Cross-Entropy (WCE) for classification task:

$$\mathcal{L}_{WCE} = -\frac{1}{N} \sum_{n=1}^N w r_n \log(p_n) + (1 - r_n) \log(1 - p_n), \quad (2)$$

where w is the class weight, p_n is the Softmax probability for the i^{th} class, and r_n is the ground truth value of $\{0, 1\}$. Finally, the overall multi-task loss function can be formulated as the sum of both loss functions:

$$\mathcal{L} = \mathcal{L}_{Dice} + \mathcal{L}_{WCE}. \quad (3)$$

3.4 Gradient-Weighted Class Activation Mapping Results

We analyse probabilistic class activation maps for gradient weighted models that classify CXR images into three classes: normal, pneumonia and COVID-19 in the COVIDx dataset. It turns out that the globally [7] and the locally

¹ <https://cutt.ly/rIB1JFQ>.

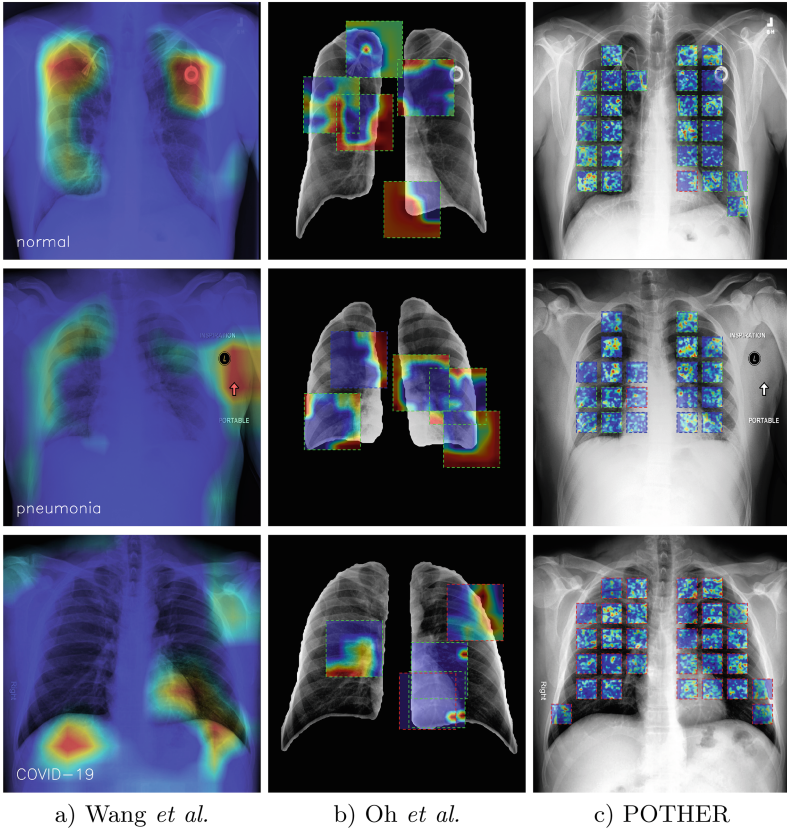


Fig. 2. The activation maps of: **(a)** cropped image global training, **(b)** segmented lung patch-based local training, **(c)** our multi-task patch-based local training method - **POTHER**. The globally trained model focuses on artificial electronics, textual information, and other non-disease-related features such as shoulder position. The model trained on the segmented lung is spared textual information but uses lung contour for classification. Our proposed model does not use these spurious features for classification but lung morphology instead.

[10] trained model focus on non-disease related elements. In contrast, POTHER focuses on the morphological structure of lung tissue. The globally trained model concentrates on laterality tokens and ECG-leads artifacts even though input images are cropped. The locally trained model lacks textual information, but instead, it uses the lung mask’s contour, supported by an edge’s strong gradient because of the black background, to classify the image. The globally trained ResNet-50 model was fed with images pre-processed and augmented, as described in Wang’s work. The pre-processing consists of cropping the top 8% of the image, and according to the authors, it is to mitigate commonly-found embedded textual information, which, as presented in Fig. 2, is not enough because there are texts localised too close to the lung to be cropped.

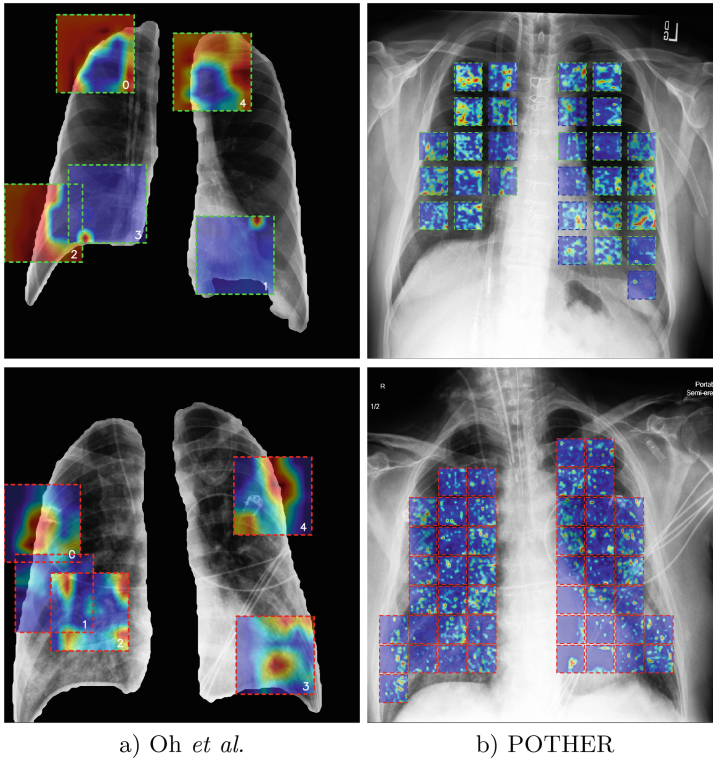


Fig. 3. Patch learning activation maps comparison: (a) Oh *et al.* train the model using big patches which, if placed close to the lung edge, provide a model with awareness of lung contour, allowing it to learn that, (b) **POTHER** model focuses on fine-grained lung markings like, e.g. interstitial opacities.

Figure 3 compares activation maps of two models trained with a patch-based learning approach. We can see that Oh’s patches cover a considerable percentage of the image, allowing the model to focus on the lung’s corner or ECG leads. On the contrary, we train POTHER on the whole unsegmented images. It does not use pixels outside the lung because of small patches and reduced mask areas inside which patch centres can be located. Though some of the POTHER’s patches are confused by cable on the healthy patient’s image, thanks to their relatively small area, the vast majority of patches vote for a proper class during majority voting.

We analyse activation maps of the POTHER model based on the visible manifestations of pneumonia and COVID-19 on CXRs. Visible with an untrained eye, pneumonia and COVID-19 pneumonia signs are related to increased lung density, which may be seen as whiteness in the lungs, which obscures the lung markings that are typically seen depending on the severity of pneumonia [19]. During COVID-19, markings are partially obscured by the increased whiteness. These can be subtle and require a confirmation from a radiologist. COVID-19 pneumonia changes, like

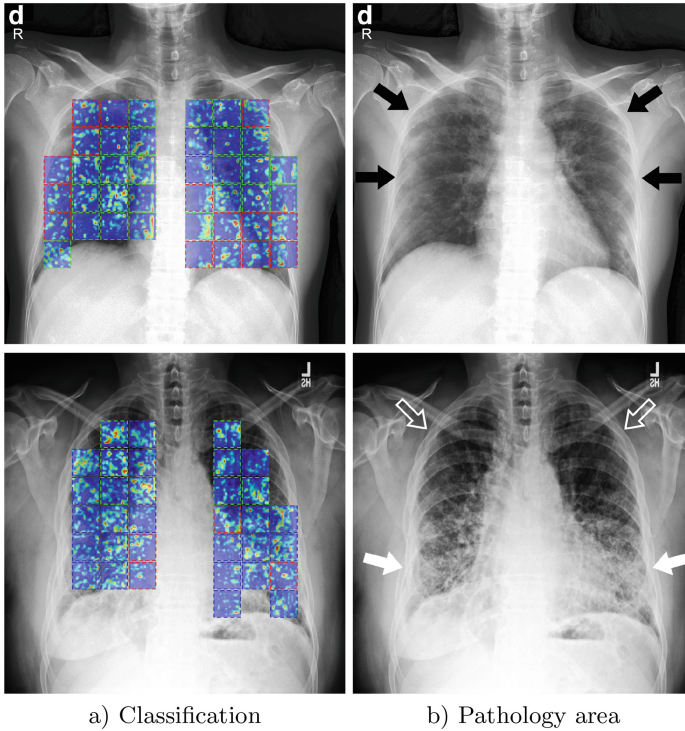


Fig. 4. POTHER activation maps and per-patch classification results - (a) the model output class is coded with patch bounding box colour: *green* - *normal lung*, *blue* - *pneumonia*, *red* - *COVID-19*. (b) black arrows point to the area with white opacity characteristic of COVID-19, white arrows point to lower lobes opacity characteristic to different types of pneumonia, outlined arrows point to the area free of opacity. The areas indicated by the arrows and the POTHER classifications overlap, indicating that the model may consider manifestations specific to the classified diseases when making decisions. (Color figure online)

lung involvement, are mostly bilateral on chest radiograph. Areas of lungs - Fig. 4, where increased whiteness can be found are pointed with blue (pneumonia) and red (COVID-19) arrows, and the region free of it is marked with green arrows. It can be seen that POTHER's decisions about patches are correlated with them, which may suggest a classification based on the disease symptoms seen on the CXR.

The locations and neighbourhoods of ECG lead elements and other non-anatomical artifacts are analysed in more detail. Figure 5 presents generated activation maps of the analysed models, strongly indicating that electronic devices' presence in the image focuses heavily the model's attention and can be an important factor in decision-making. Using POTHER, we classify the exact locations,

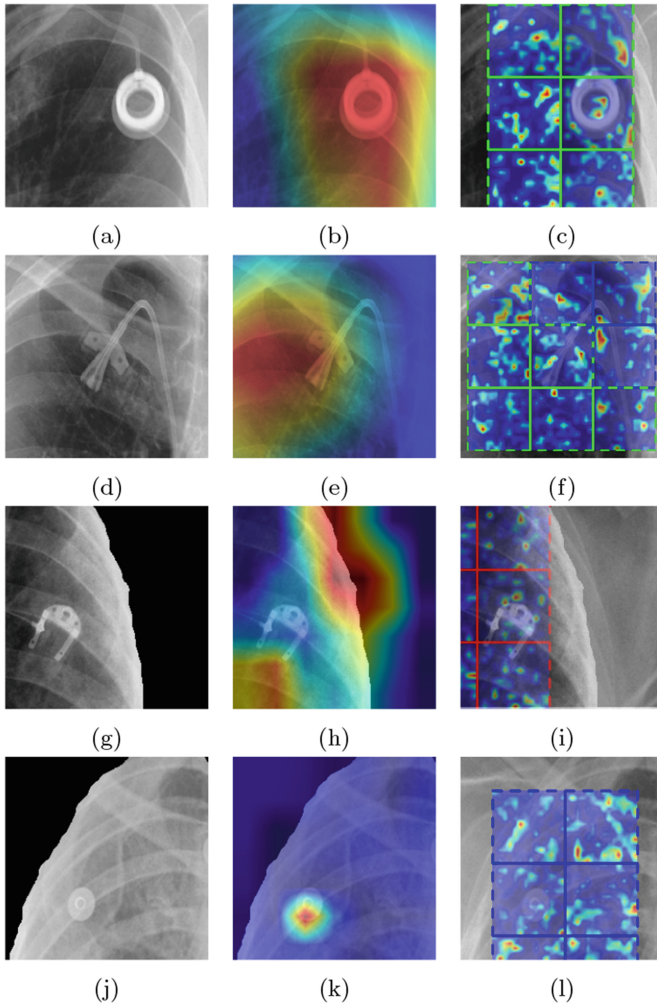


Fig. 5. Analysis of activation maps in the vicinity of elements influencing the model decision. Comparison of the models: (b, e) trained globally, (h, k) segmented-lung patch-based and (c, f, i, l) with our **POTHER** method. The left column of figures (a, d, g, j) shows fragments of the original CXR image, figures (b, e, h, k) a fragment with superimposed activation map and right column (c, f, i, l) activation maps of individual **POTHER** patches and their color-coded votes.

and its activation maps show no significant focus on artificial elements. Figure 5f shows that 3 out of 9 **POTHER**'s patches change their output near the cable; however, thanks to the majority voting it does not influence the final decision.

3.5 Quantitative Analysis

To evaluate results, we use precision, recall, F1 score and accuracy. As a test set, we use a COVIDx introduced by Wang *et al.* [7]. It consists of 300 images equally distributed into three classes. A globally trained model’s inference results in a single classification per image, while inference of a locally trained one requires majority voting before making the final decision. When we train Oh’s patch-based learning model, we follow the same pre-processing and augmentations as the author and the only differences are the model used for lungs segmentation and our mask filtration algorithm. We compare the performance of four models that are trained on the COVIDx dataset. The results presented in Table 1 are comparable in terms of accuracy, which oscillates around 90%. However, COVID-Net’s accuracy, as well as its precision of COVID-19 classification, stands out. Our method achieves the highest F1 score of 0.974 for the COVID-19 class. Oh’s model performs similarly to the results reported in their work, where it scored an accuracy of 0.889 on their dataset.

Table 1. The performance comparison on COVIDx test set

Method	Class	Precision	Recall	F1	Accuracy
ResNet-50 [7]	Normal	0.882	0.970	0.924	0.906
	Pneumonia	0.868	0.920	0.893	
	COVID-19	0.988	0.830	0.902	
COVID-Net [7]	Normal	0.905	0.950	0.927	0.933
	Pneumonia	0.913	0.940	0.926	
	COVID-19	0.989	0.910	0.948	
Patch learning [10]	Normal	0.815	0.970	0.886	0.886
	Pneumonia	0.914	0.813	0.860	
	COVID-19	0.963	0.867	0.912	
Ours (POTHER)	Normal	0.790	0.980	0.875	0.903
	Pneumonia	0.963	0.780	0.862	
	COVID-19	1.000	0.950	0.974	

4 Discussion

In this work, we propose a novel learning method specifically designed to address difficulties inherent in CXR images. In addition, we analyse the model’s biases when it is making classification decisions. The improvement of diagnostic capabilities and the reduction of the negative influence of the numerous CBs in the COVIDx dataset guide us during model development. Many researchers have achieved very high scores on CXR images when trying to improve COVID-19 classification, but unfortunately, very few have analysed what allows the model to achieve such high scores.

When trying to improve COVID-19 classification based on CXRs, great attention should be paid to understanding what features allow the model to achieve good classification. Otherwise, we may obtain a model with little diagnostic value, though it achieves seemingly high scores utilising CBs to classify on dataset that it was trained and tested. We reveal potential pitfalls in the CXR medical imaging domain, i.e. some types of clues irrelevant to disease recognition, like hospital markings and ECG leads, causing the model to miss the point of the disease symptoms. The list of CBs is not exhaustive as it cannot be guaranteed that the model will not shift its attention to something similarly undesired if known CBs are addressed. Therefore, a dataset specifically designed to eliminate as many detected CBs as possible is desirable. Annotation or segmentation of the pathological lung area by a radiologist visible in the CXR would also be valuable. According to [19], a chest radiograph can look normal in up to 63% of people infected with COVID-19, especially in the early stages. Using COVID-19 labelled radiograph with no noticeable disease-related changes in the lung image is very likely confusing to the model and encourages learning of undesirable confounding features.

5 Conclusions

In this paper, we proposed a novel multi-task patch-voted learning-based method called POTHER for CXR-image-based COVID-19 detection. Our model learns pneumonia manifestations and lung morphology features. We performed the activation map analysis and showed that the deep learning methods are susceptible to learning features unrelated to the pathology. The COVIDx dataset contains many images that can confuse the learned model, such as ECG leads, laterality tokens or other hospital-specific markings. To this end, POTHER's patches include the area of lung fields and only its closes vicinity to eliminate some of the sources of undesirable features. Using segmentation task with an attention mechanism provides an efficient feature extraction, allowing satisfactory results despite training with limited lung fragments. Thanks to training with small patches, our method is less sensitive to the CBs. In future work, we will apply our method with Vision Transformer (ViT) [26] on a large-scale CXR dataset to improve the model generalisation with more complex lung disease classes.

Acknowledgements. This work is supported in part by the European Union's Horizon 2020 research and innovation programme under grant agreement Sano No. 857533 and the International Research Agendas programme of the Foundation for Polish Science, co-financed by the EU under the European Regional Development Fund. This research was funded by Foundation for Polish Science (grant no POIR.04.04.00-00-14DE/18-00 carried out within the Team-Net program co-financed by the European Union under the European Regional Development Fund), National Science Centre, Poland (grant no 2020/39/B/ST6/01511).

References

1. Çalli, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G., Murphy, K.: Deep learning for chest X-ray analysis: a survey. *Med. Image Anal.* **72**, 102125 (2021). <https://doi.org/10.1016/j.media.2021.102125>
2. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., Jamalipour Soufi, G.: Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning. *Med. Image Anal.* **65**, 101794 (2020). <https://doi.org/10.1016/j.media.2020.101794>
3. Li, X., et al.: Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. *Artif. Intell. Med.* **103**, 101744 (2020). <https://doi.org/10.1016/j.artmed.2019.101744>
4. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans. Neural Netw. Learning Syst.* **32**, 4793–4813 (2021). <https://doi.org/10.1109/TNNLS.2020.3027314>
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, Las Vegas, NV, USA (2016). <https://doi.org/10.1109/CVPR.2016.90>
7. Wang, L., Lin, Z.Q., Wong, A.: COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **10**, 19549 (2020). <https://doi.org/10.1038/s41598-020-76550-z>
8. Kermany, D., Zhang, K., Goldbaum, M.: Labeled optical coherence tomography (OCT) and chest X-ray images for classification, vol. 2 (2018). <https://doi.org/10.17632/rschjbr9sj.2>
9. Cohen, J.P., Morrison, P., Dao, L., Roth, K., Duong, T.Q., Ghassemi, M.: COVID-19 image data collection: prospective predictions are the future. *arXiv:2006.11988* [cs, eess, q-bio] (2020)
10. Oh, Y., Park, S., Ye, J.C.: Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans. Med. Imaging* **39**, 2688–2700 (2020). <https://doi.org/10.1109/TMI.2020.2993291>
11. Al-Waisy, A.S., et al.: COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images. *Soft Comput.*, 1–16 (2020). <https://doi.org/10.1007/s00500-020-05424-3>
12. Rajaraman, S., Siegelman, J., Alderson, P.O., Folio, L.S., Folio, L.R., Antani, S.K.: Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-Rays. *IEEE Access* **8**, 115041–115050 (2020). <https://doi.org/10.1109/ACCESS.2020.3003810>
13. Pham, T.D.: Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning? *Health Inf. Sci. Syst.* **9**(1), 1–11 (2020). <https://doi.org/10.1007/s13755-020-00135-3>
14. Ucar, F., Korkmaz, D.: COVIDiagnosis-Net: deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Med. Hypotheses* **140**, 109761 (2020). <https://doi.org/10.1016/j.mehy.2020.109761>

15. Płotka, S., Włodarczyk, T., Klasa, A., Lipa, M., Sitek, A., Trzciniński, T.: FetalNet: multi-task deep learning framework for fetal ultrasound biometric measurements. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds.) ICONIP 2021. CCIS, vol. 1517, pp. 257–265. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-92310-5_30
16. Płotka, S., et al.: Deep learning fetal ultrasound video model match human observers in biometric measurements. *Phys. Med. Biol.* **67**, 045013 (2022). <https://doi.org/10.1088/1361-6560/ac4d85>
17. Amyar, A., Modzelewski, R., Li, H., Ruan, S.: Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: classification and segmentation. *Comput. Biol. Med.* **126**, 104037 (2020). <https://doi.org/10.1016/j.compbiomed.2020.104037>
18. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE, Lake Tahoe, NV (2018). <https://doi.org/10.1109/WACV.2018.00097>
19. Cleverley, J., Piper, J., Jones, M.M.: The role of chest radiography in confirming Covid-19 pneumonia. *BMJ* **370**, m2426 (2020). <https://doi.org/10.1136/bmj.m2426>
20. DeGrave, A.J., Janizek, J.D., Lee, S.-I.: AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021). <https://doi.org/10.1038/s42256-021-00338-7>
21. Nayak, S.R., Nayak, D.R., Sinha, U., Arora, V., Pachori, R.B.: Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: a comprehensive study. *Biomed. Signal Process. Control* **64**, 102365 (2021). <https://doi.org/10.1016/j.bspnc.2020.102365>
22. López-Cabrera, J.D., Orozco-Morales, R., Portal-Díaz, J.A., Lovelle-Enríquez, O., Pérez-Díaz, M.: Current limitations to identify Covid-19 using artificial intelligence with chest X-ray imaging (part ii). The shortcut learning problem. *Health Technol.* **11**(6), 1331–1345 (2021). <https://doi.org/10.1007/s12553-021-00609-8>
23. Maguolo, G., Nanni, L.: A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Inf. Fusion* **76**, 1–7 (2021). <https://doi.org/10.1016/j.inffus.2021.04.008>
24. Lin, Z.Q., Shafiee, M.J., Bochkarev, S., Jules, M.S., Wang, X.Y., Wong, A.: Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. [arXiv:1910.07387](https://arxiv.org/abs/1910.07387) [cs] (2019)
25. Roy, K., Banik, D., Bhattacharjee, D., Nasipuri, M.: Patch-based system for classification of breast histology images using deep learning. *Comput. Med. Imaging Graph.* **71**, 90–103 (2019). <https://doi.org/10.1016/j.compmedimag.2018.11.003>
26. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) [cs] (2021)