LIME-Enabled Investigation of Convolutional Neural Network Performances in COVID-19 Chest X-Ray Detection

Eduardo Gasca Cervantes and Wai-Yip Chan
Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada,
{eddie.gascacervantes, chan}@queensu.ca

2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) | 978-1-6654-4864-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCECE53047.2021.9569029

***Abstract*—The Coronavirus Disease (COVID-19) has caused millions of casualties across the globe. One inexpensive and noninvasive screening method for COVID-19 is the analysis of chest X-ray (CXR) images for pathological features in the lungs. These features are difficult to detect by humans, but convolutional neural networks (CNN) have proven effective at extracting them. This paper uses four ImageNet-pre-trained CNNs: VGG16, DenseNet201, ResNet50, and EfficientNetB3 to perform transfer learning to a task of COVID-19 CXR image detection on a dataset containing COVID-19, healthy, and viral pneumonia CXR images. We compare the performance of the retrained CNNs using standard measures and investigate the features they use for their predictions using local interpretable model-agnostic explanations (LIME). The networks are retrained on two classification tasks: Task 1 consists of classifying healthy and COVID-19 CXR images and task 2 consists of classifying viral pneumonia and COVID-19 CXR images. We find that DenseNet201 and VGG16 achieve higher accuracies than ResNet50 and EfficientNetB3 in both tasks. However, the LIME explanations reveal that VGG16 does not learn disease-relevant features in the lungs, while DenseNet201, ResNet50, and EfficientNetB3 use regions in the lungs to make their predictions. This observation is reinforced by comparing LIME explanations with ground-truth lung regions on an unseen dataset. The prospect of using "black box" deep neural networks for automatic screening of CXRs for COVID-19 can be improved with LIME-enabled investigations of model performance.**

***Index Terms*—COVID-19, chest X-rays, machine learning, explainability**

I. Introduction

The Coronavirus pandemic has taken over 24,000 lives in Canada alone and continues to spread at alarming rates throughout the world despite the roll-out of several vaccines, public health safety measures, and rapid tests [1]. The Coronavirus disease (henceforth COVID-19, or COVID) causes mild respiratory illness in most infected people, but some develop life-threatening cases of pneumonia.

The spread of COVID-19 is currently being controlled in large part through the use of clinical screening tests. The most common test is the reverse transcription polymerase chain reaction (PCR) test, which uses respiratory tissue samples to screen for the disease [2]. However, this test can be painful, invasive, and in short supply. Instead, less intrusive methods such as diagnostic radiology [3] can be used to diagnose COVID-19 non-invasively. This method includes the analysis of chest X-Ray (CXR) images for features of COVID-19 in the lungs. However, these features are shared by other types of viral pneumonia, so it is critical to be able to distinguish the CXR images of patients with COVID-19 from healthy patients as well as patients with other respiratory diseases.

Automatic feature recognition through the use of deep learning techniques is now widely used in biomedical applications. For medical imaging applications, convolutional neural networks (CNNs) have proven to be extremely effective at extracting subtle features that are not readily perceptible by humans. Many researchers have recently published papers using CNNs to detect COVID-19 in CXR images due to two main factors: the increasing availability of CXR image datasets and the broad access of free machine learning (ML) software and tools [4], [5]. The transfer learning technique is also a major contributor to the successful results in many such papers.

Transfer learning allows researchers to train only the final (top) layers of a very deep CNN using a relatively small dataset while retaining the predictive power of the rest of the pre-trained network. This technique was used by Chowdhury et al. to fine-tune several pre-trained ImageNet models to detect COVID-19 in CXR images [6]. The same concept was also applied by Apostolopoulos and Mpesiana to classify COVID-19 in CXR images among both healthy patients and several types of viral and bacterial pneumonia [7]. Image augmentation is also used to compensate for relatively small COVID-19 datasets. Techniques such as random rotations and translations of the training images were used by Alazab et al. to artificially increase the size of their dataset from 98 to 1,000 training images and achieve competitive COVID-19 detection results [8]. Many papers report validation accuracies of over 95% [9].

Although there has recently been a large amount of research on detecting COVID-19 using CXR images, the interpretability of the models has often been neglected. This aspect of artificial intelligence is of paramount importance in the medical field, since healthcare professionals do not generally trust black-box models [10]. In this paper, we aim to interpret the performances of four popular pre-trained ImageNet CNNs using Local Interpretable Model-agnostic Explanations (LIME) [11]. We compare the performance of VGG16 [12], DenseNet201 [13], ResNet50 [14], and EfficientNetB3 [15] for the detection of COVID-19 cases in CXR images and discuss the validity

978-1-6654-4864-2/21/$31.00 ©2021 IEEE

of a common model training methodology using explained predictions of both seen and unseen data.

II. Methodology

We used transfer learning with four very deep CNNs (VGG16, DenseNet201, ResNet50, and EfficientNetB3) pre-trained on the ImageNet dataset in order to alleviate the need to have a large CXR dataset and long training times [16].

A. Dataset and classification tasks

We used the COVID-19 Radiography Database [4] for the classification tasks in this study. This dataset contains 3,616 COVID-19 positive, 10,192 healthy, and 1,345 viral pneumonia CXRs. Two sub-datasets were created from the COVID-19 Radiography Database for two different classification tasks.

Dataset 1 contained only the COVID-19 and healthy CXRs and was used to train the CNNs to classify patients as either healthy or having COVID-19. This set used 3,616 COVID-19 positive CXRs and a random sample of 3,616 healthy CXRs. Dataset 2 contained the COVID-19 and viral pneumonia CXRs and was used to train the CNNs to classify a patient suffering from COVID-19 or another viral pneumonia. This dataset used 1,345 viral pneumonia CXRs and a random sample of 1,345 COVID-19 positive CXRs. The two classification tasks are shown in Fig. 1.

The same validation scheme was used for both classification tasks: 80% of the data was used in training and 20% of the data was held out for validation. The models were trained using the Adam optimizer with accuracy as the main validation metric and binary cross entropy as the loss function. The performance metrics are discussed further in section II-C. All models and experiments used the same image preprocessing steps: images were all resized to dimensions 256×256 and the proper pixel normalization was done according to the requirements of each CNN.

B. CNN models

Four pre-trained CNNs were used for transfer learning in this paper: VGG16, DenseNet201, ResNet50, and EfficientNetB3. The model architectures and ImageNet weights were included in the Tensorflow Python package. The transfer learning architectures are shown in Fig. 1.

We removed the top classification layer from the pre-trained networks and froze the model weights from the remaining layers during training. A new fully connected layer with two hidden units was appended to the top of the networks with a softmax activation function for the binary classification tasks described in the following section.

C. Performance metrics

The performance of the CNNs in detecting COVID-19 pneumonia was evaluated using three performance metrics: accuracy, sensitivity, and specificity. These are defined in (1), (2), and (3). True positives (TP) refer to the correctly classified COVID-19 cases, false negatives (FN) refer to the incorrectly classified COVID-19 cases, true negatives (TN) refer to the correctly classified healthy (or viral pneumonia) cases, and false positives (FP) refer to the incorrectly classified healthy (or viral pneumonia) cases.

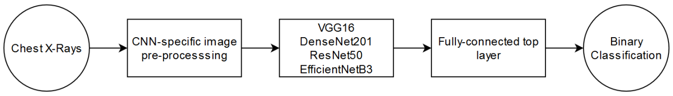


Fig. 1: CNN architecture for the both the healthy vs. COVID-19 and viral pneumonia vs. COVID-19 binary classification tasks.

It was important to measure not only accuracy, but also sensitivity and specificity because of the nature of the classification tasks. For example, it may be advantageous to have a conservative classifier that has a high sensitivity for COVID-19 detection at the expense of specificity. It could also be argued that a classifier with a low specificity (and thus more FP) could result in unnecessary follow-up screenings.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

$$Specificity = \frac{TN}{TN+FP} \tag{3}$$

D. Interpretable explanations of CNN predictions

The predictions of the CNNs were explained using Local Interpretable Model-Agnostic Explanations (LIME) [11]. LIME is a method for training a simple, interpretable linear model to approximate the decision function of any black box model such as a CNN. For image classifiers, LIME begins by creating an interpretable representation of the input images in the form of superpixels, which are collections of pixels that share similar properties such as pixel intensity.

To explain a given instance, LIME samples instances similar to the original instance and gets predictions for them using the original black box model. It then uses the sampled instances and predictions as a new training dataset to which an interpretable linear model (explainer) is fit. The explainer is then used to generate explanations in the form of saliency maps depicting regions of an image that contribute to a given prediction.

LIME explainers were fit to five instances for each model using SP-LIME [11], an algorithm that picks the most representative and least redundant explanations to show the user. The resulting explanations were analyzed to determine the trustworthiness and expected generalizability of each model.

Additionally, we used the RNSA Pneumonia Detection Competition dataset [17] to validate that the CNNs learned important features in the CXR images when detecting COVID-19 and viral pneumonia. This dataset consists of 8,964 pneumonia, 8525 healthy, and 11,500 non-pneumonia disease CXRs. The pneumonia CXR images also included ground truth bounding boxes for the affected regions. We used the

TABLE I: Performance metrics for the CNNs trained on the two classification tasks.

Task	Model	Accuracy	Sensitivity	Specificity
Healthy vs. COVID	VGG16	95.71	92.39	**99.03**
	DenseNet201	**96.47**	**96.4**	96.54
	ResNet50	93.15	90.59	95.71
	EfficientNetB3	92.39	89.62	95.15
Viral vs. COVID	VGG16	**100**	**100**	**100**
	DenseNet201	99.25	98.51	**100**
	ResNet50	96.84	94.79	98.88
	EfficientNetB3	92	84.01	**100**

pneumonia CXR images as the input to our models trained on both classification tasks and compared their LIME explanations with the ground truth regions. While it was difficult to make a performance metric comparison of the models using this unseen dataset since it did not include the COVID-19 CXR images, the ground truth regions in the pneumonia CXR images provided a way to verify the trustworthiness of the predictions generated by the pre-trained models using our common transfer learning methodology.

III. Results

Two different classification tasks were performed using four pre-trained CNN models. Task 1 consisted of classifying healthy and COVID-19 CXR images and task 2 consisted of classifying viral pneumonia and COVID-19 CXR images.

A. Results of healthy vs. COVID-19 classification task

The performance of the CNNs for the healthy vs. COVID-19 classification task on the validation dataset is shown in Table I. All models achieved relatively high classification accuracy, sensitivity, and specificity, but DenseNet201 scored higher than the other CNNs in both accuracy and sensitivity. VGG16 achieved a similar accuracy to DenseNet201, but scored significantly worse in sensitivity and subsequently better on specificity. ResNet50 and EfficientNetB3 underperformed on the validation metrics. Interestingly, the specificity of VGG16, ResNet50, and EfficientNetB3 was considerably greater than their sensitivity, indicating a bias toward predicting patients as healthy.

Although the high accuracy scores were encouraging, we proceeded to train LIME explainers on several instances of healthy and COVID-19 CXR images. Fig. 2 shows representative samples of explanations for one correct COVID-19 prediction and one correct healthy prediction from the two models with the best performance: VGG16 and DenseNet201. The left column shows the original images being explained, the middle column shows the LIME explanation superpixels that contributed to predicting the correct class, and the right column shows the superpixels that contributed against predicting the correct class.

Fig. 2a and Fig. 2b showed that VGG16 prioritized superpixels of high pixel intensity in its predictions. The areas that contributed to both predictions were not centered around the lungs, as one would expect for this type of classification, but were instead located at the bones around the shoulders and lower abdomen. This provided evidence against the claim that the VGG16 model learned relevant features in the lungs that could help distinguish a COVID-19 CXR from a healthy one.

DenseNet201 used more areas in the thorax to make its predictions, as shown in Fig. 2c and Fig. 2d. Both the COVID-19 and healthy predictions were made with contributions from superpixels in the lungs. However, there were also superpixels present in the shoulder and lower abdomen. DenseNet201 had more convincing explanations showing that it may have learned some relevant lung features, but there were also unexpected superpixels in the explanations that detracted from this conclusion.

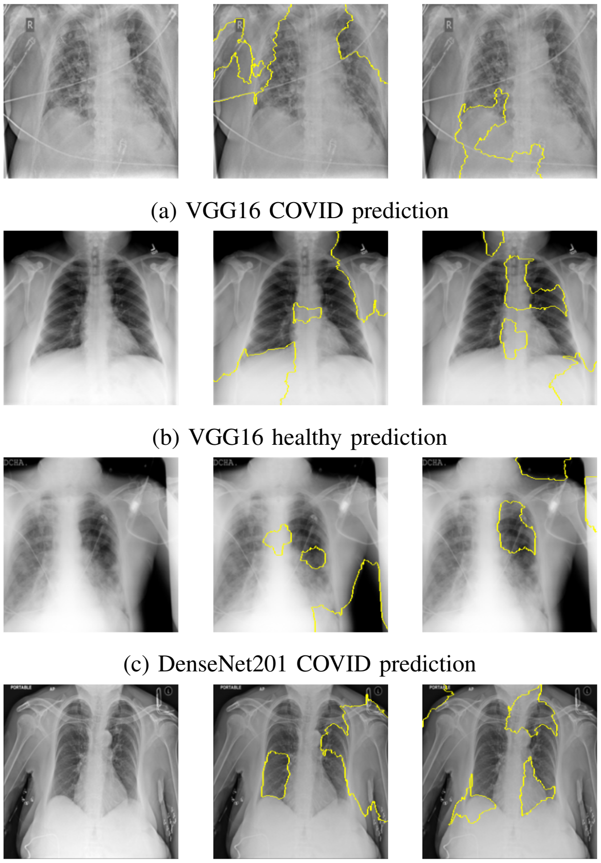
While VGG16 and DenseNet201 performed similarly in terms of accuracy, sensitivity, and specificity, the explanations provided by LIME allowed us to examine the potential generalizability of the models. Since VGG16 used the brightest white superpixels in the CXR images to make its predictions, it may not achieve good performance on a dataset with different dynamic range or pixel intensity characteristics. The DenseNet201 explanations contained superpixels in the lungs, suggesting that it was more successful in learning lung features that could be used to classify unseen datasets.

B. Results of COVID-19 vs. viral pneumonia classification task

The performance of the CNNs on the viral pneumonia vs. COVID-19 classification task on the validation dataset is shown in Table I. Again, VGG16 and DenseNet201 clearly surpassed ResNet50 and EfficientNetB3 in validation accuracy and sensitivity. The very high accuracy achieved by VGG16 and DenseNet201, as well as the high specificity achieved by all the models, may be attributed to the smaller size of the task 2 dataset.

The LIME explanations for the top two top performing models are shown in Fig. 3. VGG16 did not use superpixels containing the lungs, according to Fig. 3a and 3b. The COVID-19 prediction was made using the regions in the sternum, abdomen, and part of the CXR background. The viral pneumonia prediction contained some superpixel regions in the lungs, but also in the shoulder, abdomen, and background. This was similar to the VGG16 behavior in Task 1, where the model clearly did not learn features in the area of the lungs. This provided evidence that the model would not generalize well to unseen data and therefore its predictions could not be trusted.

The explanations of DenseNet201 shown in Fig. 3c and Fig. 3d appeared to be more reasonable. There were more superpixels that contributed to the correct class prediction from the lungs than VGG16, although there were also some superpixels in unexpected areas such as the clavicle and abdomen. This reinforced that DenseNet201 was better able to extract important features in the lungs to classify viral pneumonia and COVID-19. However, the small training dataset limited the effectiveness of the model, so more experiments with a larger training set would be required for more conclusive results.



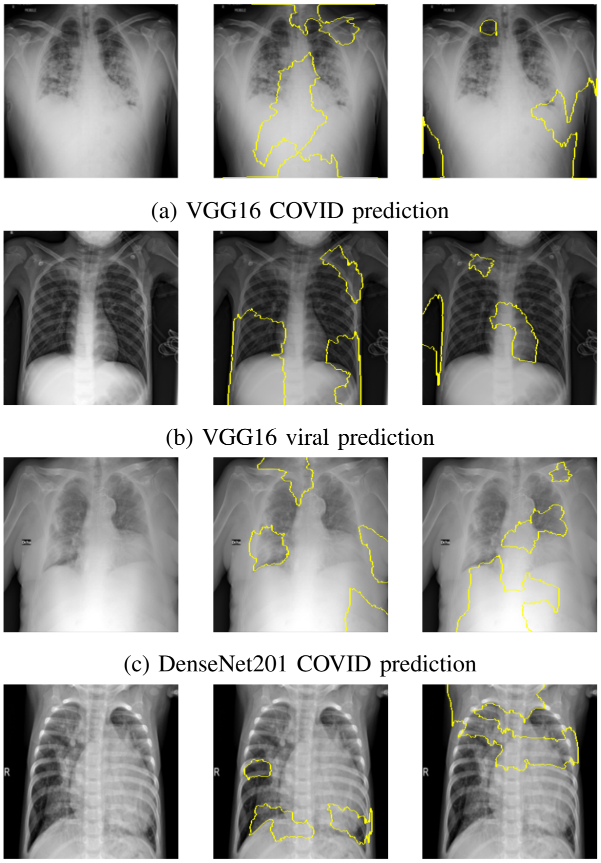
(a) VGG16 COVID prediction

(b) VGG16 healthy prediction

(c) DenseNet201 COVID prediction

(d) DenseNet201 healthy prediction

Fig. 2: LIME explanations for Task 1: Original image (left); superpixels that contributed toward predicting the correct class (middle); superpixels that contributed toward predicting the incorrect class (right)



(a) VGG16 COVID prediction

(b) VGG16 viral prediction

(c) DenseNet201 COVID prediction

(d) DenseNet201 viral prediction

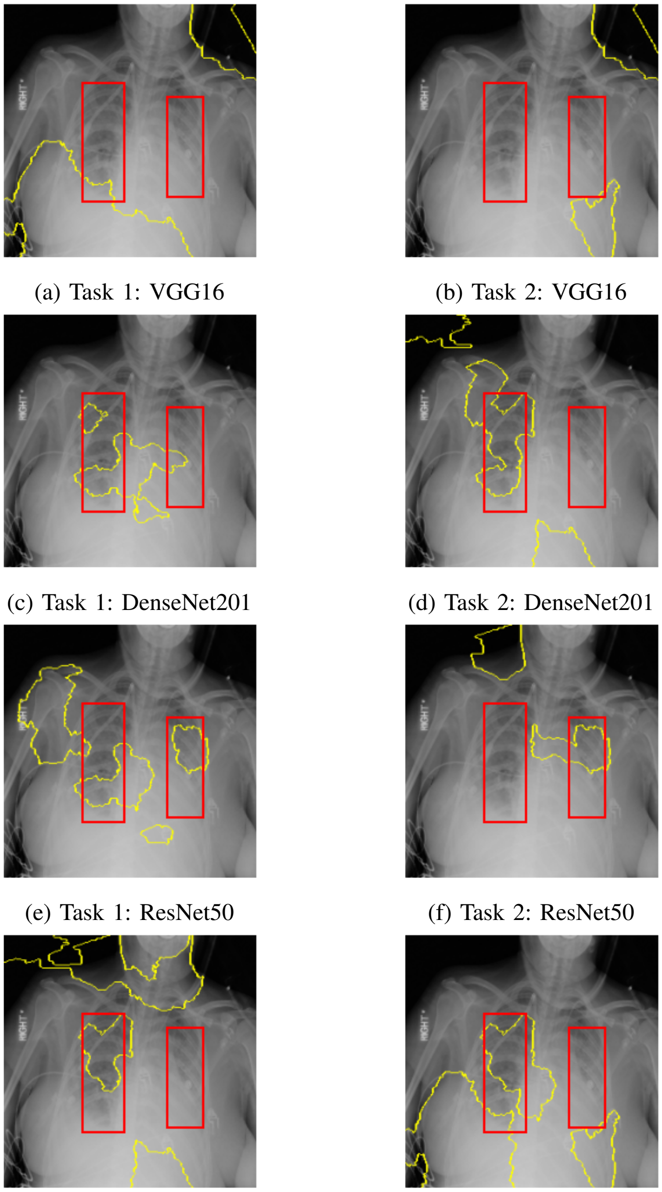
Fig. 3: LIME explanations for Task 2: Original image (left); superpixels that contributed toward predicting the correct class (middle); superpixels that contributed toward predicting the incorrect class (right)

C. *Comparison of ground truth pneumonia regions and model explanations*

To assess the validity of LIME explanations, we predicted the class of viral pneumonia CXR images from the unseen RNSA dataset described in section II using the four CNNs trained on the classification tasks. We then generated LIME explanations showing the superpixel regions that contributed to those predictions and overlayed ground truth bounding boxes representing the location of pneumonia in the lungs, as illustrated in Fig. 4. Although the model classification tasks were not directly compatible with the unseen dataset because it did not contain COVID-19 CXRs, predicting the class of pneumonia CXRs as per the CNN model allowed us to compare the superpixels used for prediction with the ground truth regions.

The models trained for task 1 all predicted that the CXR in the left column of Fig. 4 contained COVID-19. This demonstrated that every model was capable of correctly detecting the presence of a disease in the CXR. However, it is clear that the LIME explanation of the VGG16 prediction does not align with the ground truth pneumonia regions. The explanation revealed that VGG16 used a region in the lower abdomen and the black background of the CXR next to the patient's head in order to make its prediction, further reinforcing the results found in the previous analysis of VGG16 lime explanations: the model did not learn any important features in the lungs contributing to the pneumonia pathology. The explanation regions of the DenseNet201 prediction align much better to the ground truth region, indicating that the model was able to extract relevant lung features. ResNet50 and EfficienNetB3 also had LIME explanation superpixels contained within the ground truth regions, but there were also extraneous superpixel regions in irrelevant parts of the CXR. This suggests overfitting to non-pathological features similar to VGG16.

The models trained for task 2 also predicted that the CXR in the right column of Fig. 4 contained COVID-19. This could be considered a misclassification, since task 2 entails classifying viral pneumonia and COVID-19 CXR images. However, the cases of viral pneumonia in the training dataset differed from those found in the unseen dataset, further complicating the direct comparison of model performance on the unseen dataset. Nevertheless, the characteristics of the LIME explanations



(a) Task 1: VGG16

(b) Task 2: VGG16

(c) Task 1: DenseNet201

(d) Task 2: DenseNet201

(e) Task 1: ResNet50

(f) Task 2: ResNet50

(g) Task 1: EfficientNetB3

(h) Task 2: EfficientNetB3

Fig. 4: Comparison of ground truth pneumonia regions (red boxes) and model explanation superpixels (yellow regions)

were similar to those of the task 1 models. VGG16 again failed to learn features in the lungs, as evidenced by the lack of superpixels in the ground truth regions. DenseNet201, ResNet50, and EfficientNetB3 all showed a combination of superpixels inside and outside the ground truth regions. Due to the viral pneumonia type mismatch, the three models were less effective at extracting the features in the unseen CXR that contained signs of pneumonia.

IV. Conclusion

This paper analyzed the performance and interpretability of transfer learning using four pre-trained deep CNNs. The popular CNN architectures VGG16, DenseNet201, ResNet50, and EfficientNetB3 were trained and validated on two classification tasks: classifying healthy and COVID-19 CXR images and classifying viral pneumonia and COVID-19 CXR images. All models achieved similar results to previous work and DenseNet201 and VGG16 outperformed ResNet50 and EfficientNetB3 in terms of validation accuracy, sensitivity, and specificity in both tasks.

The LIME explanations for the models revealed deficiencies in their feature extraction processes. In both classification tasks, the VGG16 explanations showed that the model used parts of the CXR image irrelevant to pneumonia pathology, such as the shoulder or lower abdomen, to make its predictions. This casted doubt on the model's ability to perform well on unseen datasets. The DenseNet201 explanations were more promising, showing several regions in the lungs that contributed to its predictions. This indicated that the model extracted important features in the lungs corresponding to COVID-19 and viral pneumonia pathology.

In order to verify that the models with better explanations were truly learning COVID-19 and viral pneumonia features, each model predicted the class of unseen pneumonia CXR images with ground truth bounding boxes containing the location of pneumonia in the lungs. Comparing the superpixel regions from the LIME explanations with the ground truth regions demonstrated an important observation: a model that achieves a high validation accuracy may not have learned relevant features which will allow it to generalize to unseen data. VGG16 and DenseNet201 achieved the highest accuracies in both classification tasks, but the experiment revealed that only DenseNet201 had actually extracted features corresponding to the regions of pneumonia in the lungs.

The transfer learning methodologies and classification tasks presented in this work are common in recent literature. Therefore, the insights regarding the interpretability and trustworthiness of the CNNs in this paper have wider applicability. Our work demonstrated that using the machine learning interpretability method LIME to investigate the features used by the machine learning models helped to identify more reliable and trustworthy models.

References

[1] World Health Organization *et al.*, "Novel coronavirus (2019-nCoV): situation report, 3," 2020.

[2] W. Wang, Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, and W. Tan, "Detection of SARS-CoV-2 in different types of clinical specimens," *Jama*, vol. 323, no. 18, pp. 1843–1844, 2020.

[3] W. E. Brant and C. A. Helms, *Fundamentals of diagnostic radiology*. Lippincott Williams & Wilkins, 2012.

[4] T. Rahman, "Covid-19 radiography database," Mar 2021. [Online]. Available: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[5] 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.

[6] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi *et al.*, "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[7] I. D. Apostolopoulos and T. A. Mpesiana, “Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks,” *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, 2020.

[8] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana, and S. Alhyari, “Covid-19 prediction and detection using deep learning,” *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 12, pp. 168–181, 2020.

[9] H. Mohammad-Rahimi, M. Nadimi, A. Ghalyanchi-Langeroudi, M. Taheri, and S. Ghafouri-Fard, “Application of machine learning in diagnosis of COVID-19 through X-Ray and CT images: A scoping review,” *Front. Cardiovasc. Med.*, vol. 8, p. 185, 2021.

[10] F. Wang, R. Kaushal, and D. Khullar, “Should health care demand interpretable artificial intelligence or accept ”black box” medicine?” *Annals of internal medicine*, vol. 172, no. 1, p. 59—60, 2020.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, “”Why should i trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144.

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

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[14] 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.

[15] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105–6114.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.