

Explaining transfer learning models for the detection of COVID-19 on X-ray lung images

Abd Al-Rahman Odeh¹, Ahmad Mustafa²

¹School of Engineering and Computing, American International University, Al-Jahra, Kuwait

²Faculty of Computer and Information Technology, Jordan University of Science and Technology, Irbid, Jordan

Article Info

Article history:

Received Feb 7, 2024

Revised Mar 10, 2024

Accepted Mar 15, 2024

Keywords:

Computer aided diagnosis

COVID-19

Deep learning

Explainable artificial intelligence

Medical decision support

ABSTRACT

Amidst the coronavirus disease 2019 (COVID-19) pandemic, researchers are exploring innovative approaches to enhance diagnostic accuracy. One avenue is utilizing deep learning models to analyze lung X-ray images for COVID-19 diagnosis, complementing existing tests like reverse transcription polymerase chain reaction (RT-PCR). However, trusting these models, often viewed as black boxes, presents a challenge. To address this, six explainable artificial intelligence (XAI) techniques: local interpretable model agnostic explanations (LIME), Shapley additive explanations (SHAP), integrated gradients, smooth-grad, gradient-weighted class activation mapping (Grad-CAM), and Layer-CAM are applied to interpret four transfer learning models. These models: VGG16, ResNet50, InceptionV3, and DenseNet121 are analyzed to understand their workings and the rationale behind their predictions. Validating the results with medical experts poses difficulties due to time and resource constraints, alongside the scarcity of annotated X-ray datasets. To address this, a voting mechanism employing different XAI methods across various models is proposed. This approach highlights regions of lung infection, potentially reducing individual model biases stemming from their structures. If successful, this research could pave the way for an automated system for annotating infection regions, bolstering confidence in predictions and aiding in the development of more effective diagnostic tools for COVID-19.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Abd Al-Rahman Odeh

School of Engineering and Computing, American International University

Al-Jahra, Kuwait

Email: a.odeh@aiu.edu.kw

1. INTRODUCTION

The wide spread of coronavirus disease 2019 (COVID-19) [1], which is an infectious respiratory system disease caused by the SARS-CoV-2 virus, has been an active area of research and the talk of the world since the start of the pandemic around December 2019 till now, after causing the death of over 6 million people around the globe. In addition to threatening the lives of the patients, it affected communities in social, economic, and educational aspects. Early detection of the infection is necessary to decide on an early treatment plan for the patient and this is not possible due to the extensive pressure on the health facilities and the lack of nucleic acid amplification tests and antigen test equipment that are used to identify COVID-19 infections. Another method is required to make up for this lack, thus the usage of lung X-ray images.

The diagnoses of X-ray images still require the intervention of medical experts, which does not solve the pressure on the facilities' problem. Hence, a computer-aided approach that would diagnose the

patients instead of medical experts could reduce the overhead on the health facilities and an example of this is deep learning [2]. Research has already been conducted to detect the infection of several respiratory system diseases like pneumonia, lung opacity, and COVID-19 [3] while other research is conducted to specifically detect COVID-19 [4]. The results of these studies have proven that deep learning could perform well and help doctors and medical staff in their work.

Explainable artificial intelligence (XAI) has always been an obstacle in the machine and deep learning solutions [5]. Furthermore, most of the methods that achieved better results used deep learning approaches such as transfer learning which are considered black-box models. Black-box models usually have a complex architecture, and they lack transparency, that is they do not show how they work from the inside, making them hard to trust and deploy in high-stakes situations such as healthcare. Machine learning models, on the other hand, are simpler and easier to explain when compared to deep learning [6] making them a better choice for some organizations, but usually require more processing and do not perform as well as black-box models. To overcome this interpretability issue in the medical sector and utilize the better performance provided by them, authors have surveyed various approaches to explain deep learning methods and applications in diagnosing diseases such as Alzheimer's, skin diseases, and breast cancer, along with other diseases [7], [8].

In study [4], four transfer learning models were fine-tuned, hyperparameters were optimized, and several experiments were conducted to get the best performance of these models to detect COVID-19 from lung X-ray images. Moreover, to make these models more trustful and transparent, explaining these models is set to be one of the goals of this study. omni explainable artificial intelligence (OmniXAI) [9], which is an open-source Python library for explainable artificial intelligence, is utilized in this research as it offers the user different interpretation methods and techniques such as Integrated gradients (IG) [10], model agnostic counterfactual explanations (MACE) [11], contrastive explanations (CE) [12], local interpretable model-agnostic explanations (LIME) [13], Shapley additive explanations (SHAP) [14], and other approaches that are explored and discussed in this study.

Requesting assistance from medical experts to evaluate the interpretability results of the model can be a time-consuming and expensive process. In most cases, comprehending the results by providing images of the explanation without some kind of medical diagnosis is proven to be difficult. Thus, another goal of this work is to research a voting approach to examine the possibility of raising confidence in the prediction and interpretations without the need to refer to a medical expert. This approach proposes using more than one model (four models in this study) and applying various XAI approaches that are discussed later (six approaches) and then comparing the results across different models. This method assumes that the interpretation methods of the different models should all highlight similar areas of the lung X-ray, and these areas would contain actual signs of the infection. By using different models, it is assumed that a model that is wrongly trained will not affect the results and by using different explanation methods it is assumed that false explanations will not affect the conclusion. Furthermore, biases that are caused by a certain model because of its architecture can be detected when compared to other models' results.

2. RELATED WORK

We discuss the related work that is conducted in the field of explainable artificial intelligence for detecting COVID-19 on X-ray images. A customized convolutional neural network was introduced by the authors in [15] to identify COVID-19, pneumonia, and tuberculosis in chest X-rays. They used different explanatory techniques and three datasets with a total of 7,132 X-ray images divided into four classifications. Resizing and scaling were involved in image preparation. The suggested lightweight convolutional neural network (CNN) architecture included softmax for classification, five 2D-convolutional layers, dropout, and max pooling. LIME, SHAP, and Grad-CAM were used to explain the model's predictions once it had been trained and its performance measured. They worked with medical professionals to validate explanations and achieved great accuracy. Despite efforts to enhance the data, the dataset was small and unbalanced. For X-ray pictures, the use of horizontal flips for augmentation has been questioned. Notably, the lack of hyperparameter optimization and transfer learning faced further restrictions, raising questions regarding model complexity and parameter settings.

In a different study, described in [16], the goal was to use transfer learning models to identify disorders of the pulmonary system using X-rays. They used two publicly accessible datasets to train five transfer learning models and an additional ensemble model. From these datasets, a sample of 1,302 photos were randomly chosen and divided into various classifications. Numerous interpretation techniques were used, both locally and regionally, using six different approaches and neuron activation profiles. Some techniques showed difficulties with certain models. By annotating X-ray pictures, medical professionals improved interpretation and concluded that residual network (ResNet) models were the most comprehensible. Participation of medical experts and a variety of network usage were advantages. However, there was

considerable misunderstanding due to the usage of various XAI techniques. The paper has several issues, including imbalances, insufficient dataset size, confusing explanations of results, and a lack of linked work. Concerns were also raised by the lack of information on model optimization and hyper parameterization.

Researchers created an explainable deep learning strategy for quick lung disease and COVID-19 detection from X-ray images in the study described in [17], to accelerate the diagnostic process in comparison to conventional samples like sputum or blood. The three steps of the methodology were as follows: first, training a model to distinguish between healthy X-ray images and those with pulmonary diseases; second, training a second model to classify specific diseases if the image was deemed unhealthy; and third, using Grad-CAM to illustrate the model's focal points in the images. In each of the initial steps, a refined VGG16 transfer learning model was used. To identify infected lung regions and link them to the model's weightings for classification, the scientists worked with a radiologist. Concerns were expressed when using the VGG16 transfer learning model because more sophisticated models were available. Furthermore, freezing every layer of the model without hyperparameter optimization went against the norm for lung image analysis and medical best practices. Even though the study provided good debugging and explainability, the reliance on a single example and interpretability method prevented thorough comprehension, particularly when the model flagged unnecessary parts that were not covered.

From what is discussed so far and other studies like [18]–[21], it is noticed that there are some problems during the training process of most of the works such as the size of the data, the pre-processing, or the experiments conducted to train the model. Another gap is not building specific models for COVID-19, where a number of the related work trained multi-classification models and applied them to detect COVID-19. Furthermore, Grad-CAM is the only explainable approach that was applied in most of the literature even though several other methods exist such as SHAP, LIME, and MACE. In our research, the main goal is to provide exploitability of all the models that are trained and used by applying several interpretability methods and comparing the results of the models with each other to see if the models agree on the main infection area of the lung to strengthen the confidence in the models.

3. METHOD

The method we suggest in this paper is based on an earlier work [4] where various models were trained to identify COVID-19 in an X-ray image of the lung, the results were compared, and then the optimum configuration for each model was selected. In this paper, several explainable artificial intelligence (XAI) approaches that are discussed in the following subsection are applied to the trained models to provide insights and explanations of how the models work and why their decisions are made. So, when both the previous and current work approaches are combined, the final approach is reached which is shown in Figure 1. Our approach is divided into five phases, where the first three phases are conducted in the previous work [4] and this work introduces the final two phases:

- Phase 1 shows how to gather data and do the pre-processing necessary to alter image formats and sizes.
- Phase 2 is the model training phase, during which the dataset is used to train the model.
- Phase 3 is the outcomes of the executed tests are displayed and compared.
- Phase 4 illustrates applying the XAI methods to the models after choosing specific samples of the dataset that cover all cases of true positives, true negatives, false positives, and false negatives.
- Phase 5 is to evaluate the interpretation results and the performance of the models.

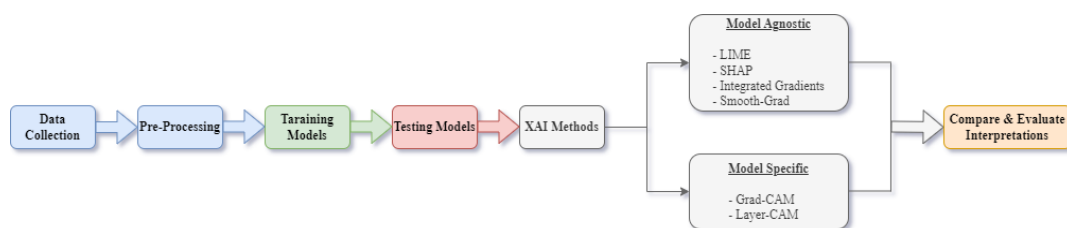


Figure 1. Methodology overview

3.1. Models overview

In this subsection of the methodology, a brief overview of how the models that are going to be interpreted in our paper were trained in the previous work. Two datasets from Kaggle have been utilized with a combined number of X-rays of 23,311 images and classified into four classes: normal, COVID-19,

pneumonia, and lung opacity. After that, the data is pre-processed by excluding the other classes from the analysis as our focus is solely on the detection of COVID-19 as shown in Table 1. Distribution of images across datasets, resizing, and converting all images into grayscale as they come from different datasets. Each network has its pre-processing function that is applied to the data during the training process. Moreover, the data is split into 3,500 training samples for COVID-19, 5,000 training samples for normal, 200 validation samples, and 626 testing samples after trying several splits and finding this one to be the best suited for the data. As for the training process, four models are trained: VGG16, ResNet50, InceptionV3, and DenseNet121 after finetuning their architecture according to the state-of-the-art recommendations and running experiments using different values for the hyperparameters illustrated in Table 2. Hyperparameter values of experiments to get the best performance out of the models.

Table 1. Distribution of images across datasets

Class	Dataset 1 [22]	Dataset 2 [23]	Full Dataset
Normal	10,192	711	10,903
Covid	3,616	711	4,327
Total	13,808	1,422	15,230

Table 2. Hyperparameters values of experiments

Parameter	Value
Learning rate	0.001/0.0001
Decay	Yes/No
Dropout rate	0.5/0.2
Batch size	128/64/32

3.2. XAI methods

As stated before, neural networks are considered black box models, thus they cannot be interpreted easily like decision trees or linear regression. There are two ways to interpret such models, which are model-agnostic and model-specific methods. Furthermore, some approaches can be used to explain the model globally, while others are used to explain it locally. Methods that explain a particular prediction made by a model are referred to as local interpretation while methodologies that explain the model as a whole are referred to as global interpretation. The main focus of this paper is to get an explanation of a specific prediction so that it could be used later in real life. Thus, all the methods that are applied are local interpretations of the models.

Model Agnostic methods are techniques that are independent of the architecture or algorithm of the model and can be applied to any machine or deep learning model to explain its predictions. Four model-agnostic methods have been chosen to be applied in this work: LIME, SHAP, Smooth-Grad, and Integrated Gradients. On the contrary, model-specific methods are specialized to a particular machine learning model and are intended to give justifications for its predictions or conclusions. These techniques may not be easily adaptable to various kinds of models because they frequently depend on the model's design or training methodology. Two techniques are applied in this work that is based on the concept of the class activation maps (CAM) approach which is Grad-CAM and Layer-CAM.

Different tools and libraries are developed by companies to ease the process of interpreting deep learning models. In this paper, two libraries are utilized on the selected models and this subsection explores them. The first library is called OmniXAI [12] which is a Python library for XAI that seeks to deal with the difficulties of really explaining the choices made by machine learning models. That being said, OmniXAI has failed in applying smooth-Grad explanation on the models because of implementation issues in the library, therefore another library is required to apply this method. Thus, tensorflow-keras-visualizer (tf-keras-vis) [24], which is another Python library, is used and has succeeded in applying the methods that OmniXAI has failed to apply.

4. RESULTS AND DISCUSSION

It can be noticed that ResNet50 and DenseNet121 have achieved the best performance due to the complexity of the models' architecture (residual blocks of ResNet and dense blocks of DenseNet). Furthermore, the common principle that the deeper the network, the better the results are violated in this study as 50 layers of the ResNet achieved better results than 121 layers of the DenseNet. ResNet50 was chosen to be the best model in this work even though the testing loss of ResNet and DenseNet is similar, but when the confusion matrix is taken into consideration, ResNet has misclassified 7 testing samples, whereas

DenseNet misclassified 8. The results of VGG16 and InceptionV3 are close, but the false positive and false negative rates are higher when compared to the other two models. Table 3 shows the best hyperparameter values and the performance of each model in terms of accuracy and loss for training, validation, and testing.

Table 3. Performance of the trained models

Model	Training Accuracy	Training Loss	Testing Accuracy	Testing Loss
VGG16	99.56 %	0.01	98.88 %	0.04
ResNet50	99.8 %	0.01	99.44 %	0.02
InceptionV3	99.54 %	0.01	98.8 %	0.03
DenseNet	100 %	0	99.36 %	0.02

4.1. XAI results and discussion

In this subsection, the results of interpreting the previously mentioned models are provided and discussed. Four samples that cover all the models' prediction cases are taken into consideration and six XAI methods (three model-agnostic and three models specific) are applied to the results of the testing to interpret how the models work. Only the results of LIME as the model-agnostic method and Grad-CAM as the model-specific method are included in this subsection. The results of the rest methods are provided in the code files and discussed in each of the following subsections. In LIME, the green areas of the images highlight the regions that have the highest effect on the model to make the prediction, whereas the red areas highlight the regions that have caused the model to think the image would belong to the other class. Grad-CAM, on the other hand, highlights the areas with the most effect on the prediction as red, followed by yellow and blue for areas that have no effect. When the image is passed to the model for testing, the model returns the percentage that the image would fall into the corresponding class and these percentages can be used to measure how well the model performs.

A true positive case is the case where the model has correctly classified patient's X-ray as COVID-19, indicating the ability to detect the disease, and to prepare a treatment plan and take the required precautions to limit the spread of the disease. The sample image "Covid_1036" in Figure 2(a) has been correctly classified by the models ResNet50 in Figure 2(b), DenseNet121 in Figure 2(c), InceptionV3 in Figure 2(d), and VGG16 in Figure 2(e). Upon observing the results of LIME in Figure 2(a), the green areas show the reason why the models have classified the image as covid. It is perceived that the left lung, especially the bottom part of it, is the main reason why this patient is classified as covid. Moreover, Grad-CAM shown in Figure 3(a), Figure 3(b) for ResNet50, Figure 3(c) for DenseNet121, Figure 3(d) for InceptionV3, and Figure 3(e) for VGG16, the same conclusion could be aligned. From the results, we conclude that the left lung may contain signs that indicate COVID-19 infection, and thus the models have made their decision.

A true negative indicates that the models have correctly classified an X-ray as non-covid infection. The sample image ("Normal_10012" in Figure 4(a)) has been successfully classified as non-covid by the four models. The interpretation images of ResNet50 and VGG16 show that the models are looking outside of the lung as no interesting information is found inside of it. On the other hand, DenseNet121 and InceptionV3's focus is on both lungs, but they also show that the area outside the lung had some contribution in the prediction which resulted in "Negative". From the results, if there are no visible signs of COVID-19 infection inside the lung, the model would either look outside of it and make the prediction or look into some areas and find them clear of infection. LIME samples are provided in Figure 4(b) for ResNet50, Figure 4(c) for DenseNet121, Figure 4(d) for InceptionV3, and Figure 4(e) for VGG16. Grad-CAM samples are provided in Figure 5(a), Figure 5(b) for ResNet50, Figure 5(c) for DenseNet121, Figure 5(d) for InceptionV3, and Figure 5(e) for VGG16.

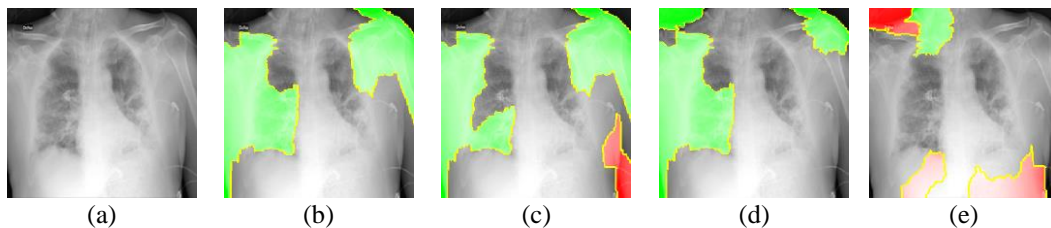


Figure 1. Model-agnostic method (LIME) interpretation for a true positive sample where (a) is the original sample, (b) ResNet50, (c) DenseNet121, (d) InceptionV3, and (e) VGG16 models

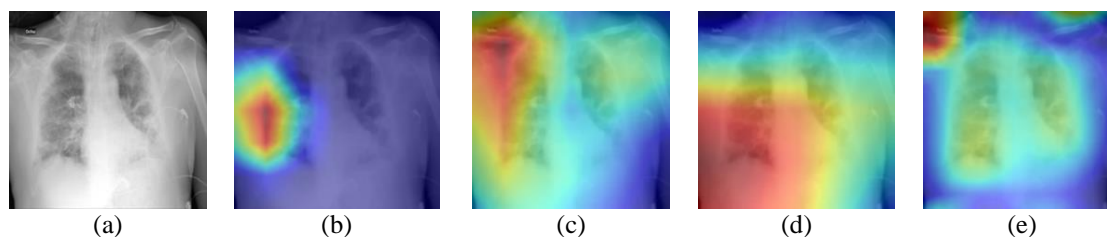


Figure 2. Model-specific method (Grad-CAM) interpretation for a true positive sample where (a) is the original sample, (b) ResNet50, (c) DenseNet121, (d) InceptionV3, and (e) VGG16 models

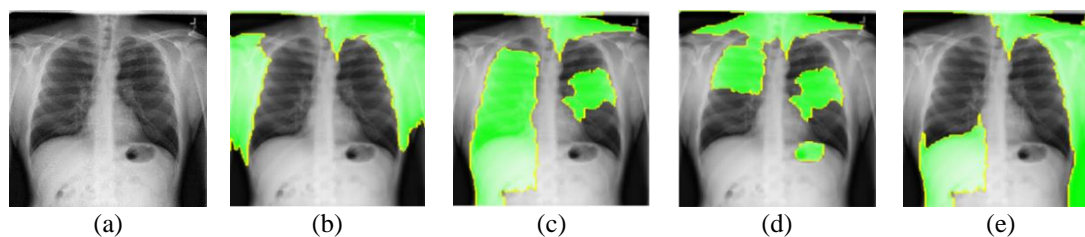


Figure 3. Model-agnostic method (LIME) Interpretation for a true negative sample where (a) is the original sample, (b) ResNet50, (c) DenseNet121, (d) InceptionV3, and (e) VGG16 models

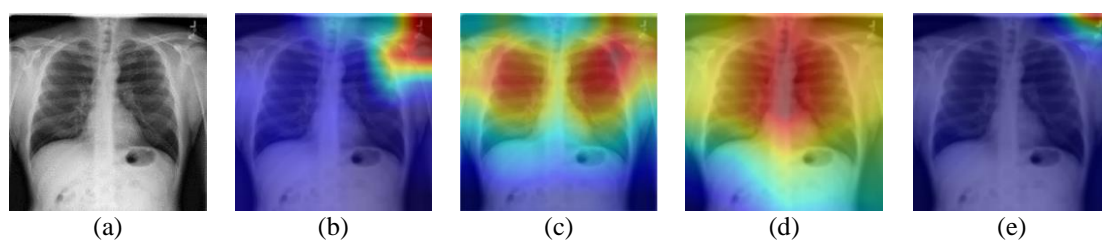


Figure 4. Model-specific method (Grad-CAM) interpretation for a true negative sample where (a) is the original sample, (b) ResNet50, (c) DenseNet121, (d) InceptionV3, and (e) VGG16 models

False positive indicates that the models have incorrectly classified a non-Covid image as “Covid”. Image “Normal_1623” in Figure 6(a) is misclassified by ResNet50, interpreted in Figure 6(b) and 6(c). Image “Normal_8793” in Figures 7(a) and 8(a) has been misclassified classified by DenseNet121 in Figures 7(b) and 8(b), InceptionV3 in Figures 7(c) and 8(c), and VGG16 in Figures 7(d) and 8(d). The green areas of the LIME interpretation highlight signs of COVID-19 infection, while the red areas are the parts of the image that led the models to believe that the image is negative. The results of LIME are not clear as each model focuses on different areas of the image because it is a model-agnostic approach. On the other hand, Grad-CAM results demonstrate that the left lung has misled the models into the wrong prediction.

False negative indicates that the models have failed to detect a covid infection and classified it as “negative”. Studying false negative cases is important in our work because it could lead to more cases due to not taking the necessary precautions such as quarantine. Image “Covid_1551” in Figure 9(a) is misclassified by VGG16 (interpreted in Figure 9(b) by LIME and Figure 9(c) by Grad-CAM). Image “Normal_1379” in Figures 10(a) and 11(a) has been misclassified classified by ResNet50 in Figure 10(b) and 11(b), DenseNet121 in Figures 10(c) and 11(c), and InceptionV3 Figures 10(d) and 11(d). Unlike the false positive cases, the red areas of LIME interpretation highlight signs of COVID-19 infection, while the green areas are the parts of the image that led the models to believe that the image is normal. ResNet50 and InceptionV3 have assigned a high percentage for covid unlike VGG16 and DenseNet121 because the red areas highlight relevant parts of the lung that contain an infection, but the effect of the green areas is higher and misled the models. Also, as perceived from the Grad-CAM results, the throats of the patients have the most effect on the models. The samples covid 1,379, 3,867, 389, and 634 are misclassified by three models, and covid 3,662 is misclassified by four models, so they may be considered as noise due to their bad quality.

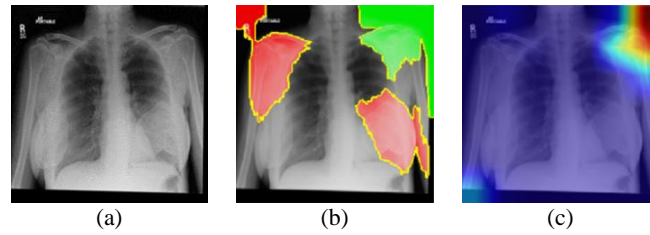


Figure 5. ResNet50 false positive sample (Normal_1623) where (a) is the original image, (b) is the LIME, and (c) is the Grad-CAM interpretation of the sample

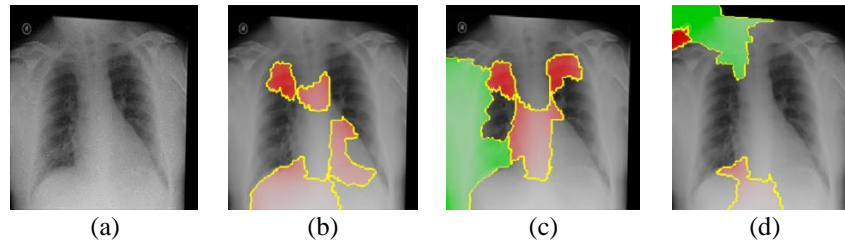


Figure 6. Model-agnostic method (LIME) interpretation for false positive sample (Normal_8793) where (a) is the original image, (b) DenseNet121, (c) InceptionV3, and (d) VGG16 models

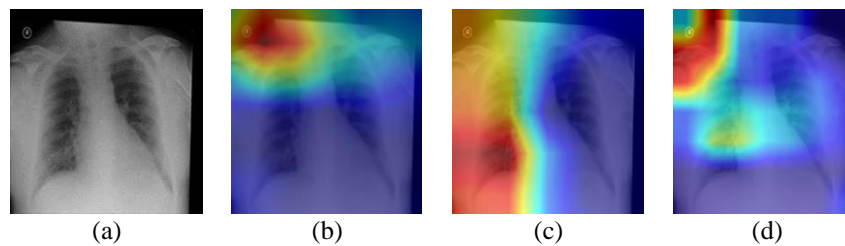


Figure 7. Model-specific method (Grad-CAM) interpretation for false positive sample (Normal_8793) where (a) is the original image, (b) DenseNet121, (c) InceptionV3, and (d) VGG16 models

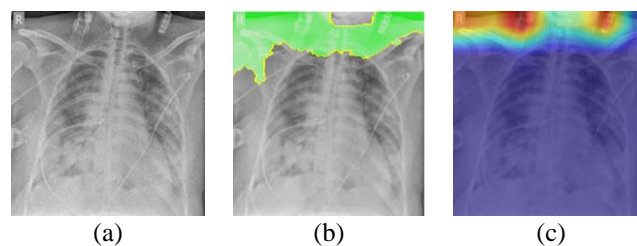


Figure 8. VGG16 false negative sample (Covid_1551) where (a) is the original image, (b) is the LIME, and (c) is the Grad-CAM interpretation of the sample

Overall, our approach has shown good results that can be observed by the figures. For example, integrated gradient has often shown weird patterns across different cases and models, so it is useful to use more than one XAI method (agnostic and specific) unlike most of the related work. Moreover, using more than one model has also proven to come in handy when looking at the false positive and false negative samples where one model classified a sample correctly and the other three have misclassified it. This could help us measure the confidence in the model's predictions and know when to trust them or not. On the other hand, when taking the true positive and true negative cases into consideration, it is noticed that most models focus on the same part of the lung and some of them focus more on specific parts than others. Nevertheless, the same parts are always assigned weight values for attribution in the prediction, which is assumed at the

beginning of the study. For reference, all of the XAI methods results and figures are available on GitHub as a part of this paper [25].

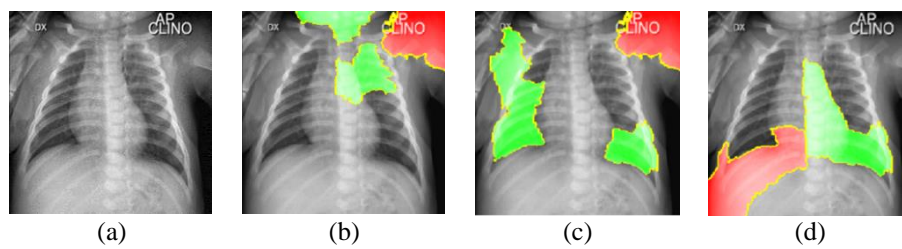


Figure 9. Model-agnostic method (LIME) interpretation for false negative sample (Covid_1379) where (a) is the original image, (b) ResNet50, (c) DenseNet121, and (d) InceptionV3 models

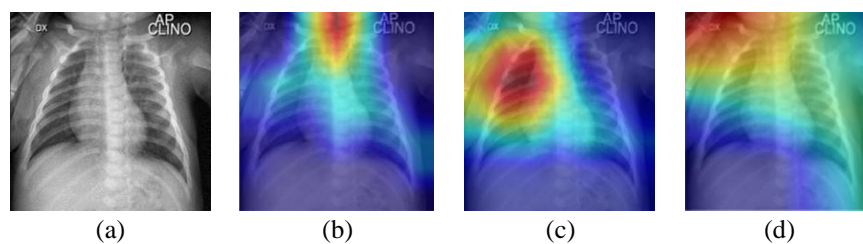


Figure 10. Model-specific method (Grad-CAM) interpretation for false negative sample (Covid_1379) where (a) is the original image, (b) ResNet50, (c) DenseNet121, and (d) InceptionV3 models

5. CONCLUSION AND FUTURE WORK

To increase trust in classification models in the medical field, it is important to provide explanations for their predictions. In this study, six methods for XAI (three model-agnostic and three model-specific) were applied to four previously trained models as a reference for experts in the field. In future we aim to estimate the certainty of predictions by utilizing the interpretability results of different models.

ACKNOWLEDGEMENTS

This work has received funding from the American International University for publication.




REFERENCES

- [1] WHO, "Coronavirus," World Health Organization (WHO), https://www.who.int/health-topics/coronavirus#tab=tab_1 (accessed Dec. 27, 2022).
- [2] H. Schulz and S. Behnke, "Deep learning: Layer-wise learning of feature hierarchies," *KI - Kunstliche Intelligenz*, vol. 26, no. 4, pp. 357–363, May 2012, doi: 10.1007/s13218-012-0198-z.
- [3] E. Khan, M. Z. U. Rehman, F. Ahmed, F. A. Alfouzan, N. M. Alzahrani, and J. Ahmad, "Chest X-ray classification for the detection of COVID-19 using deep learning techniques," *Sensors*, vol. 22, no. 3, p. 1211, Feb. 2022, doi: 10.3390/s22031211.
- [4] A. A. R. Odeh, A. Alomar, and S. Aljawarneh, "Detection of COVID-19 using deep learning on X-ray lung images," *PeerJ Computer Science*, vol. 8, p. e1082, Sep. 2022, doi: 10.7717/PEERJ-CS.1082.
- [5] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): a survey," *arxiv.org/abs/2006.11371*, Jun. 2020.
- [6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
- [7] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," *Multimedia Systems*, vol. 28, no. 6, pp. 2335–2355, Jun. 2022, doi: 10.1007/s00530-022-00960-4.
- [8] B. Hu, B. Vasu, and A. Hoogs, "X-MIR: EXplainable medical image retrieval," in *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, Jan. 2022, pp. 1544–1554, doi: 10.1109/WACV51458.2022.00161.
- [9] W. Yang, H. Le, T. Laud, S. Savarese, and S. C. H. Hoi, "OmniXAI: A library for explainable AI," *arxiv.org/abs/2206.01612*, Jun. 2022.
- [10] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, PMLR 70, 2017.
- [11] W. Yang, J. Li, C. Xiong, and S. C. H. Hoi, "MACE: An efficient model-agnostic framework for counterfactual explanation," *arxiv.org/abs/2205.15540*, May 2022.
- [12] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and




- the GDPR,” *SSRN Electronic Journal*, 2017, doi: 10.2139/ssrn.3063289.
- [13] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’ Explaining the predictions of any classifier,” in *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 2016, pp. 97–101. doi: 10.18653/v1/n16-3020.
- [14] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.
- [15] M. Bhandari, T. B. Shahi, B. Siku, and A. Neupane, “Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI,” *Computers in Biology and Medicine*, vol. 150, Art. no. 106156, Nov. 2022, doi: 10.1016/j.compbiomed.2022.106156.
- [16] S. Chatterjee *et al.*, “Exploration of interpretability techniques for deep COVID-19 classification using Chest X-ray images,” *Journal of Imaging*, vol. 10, no. 2, Jun. 2024, doi: 10.3390/jimaging10020045.
- [17] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, “Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays,” *Computer Methods and Programs in Biomedicine*, vol. 196, Art. no. 105608, Nov. 2020, doi: 10.1016/j.cmpb.2020.105608.
- [18] K. Y. Win, N. Maneerat, S. Sreng, and K. Hamamoto, “Ensemble deep learning for the detection of COVID-19 in unbalanced chest X-ray dataset,” *Applied Sciences (Switzerland)*, vol. 11, no. 22, p. 10528, Nov. 2021, doi: 10.3390/app112210528.
- [19] L. V. de Moura, C. Mattjie, C. M. Dartora, R. C. Barros, and A. M. Marques da Silva, “Explainable machine learning for COVID-19 pneumonia classification with texture-based features extraction in chest radiography,” *Frontiers in Digital Health*, vol. 3, Jan. 2022, doi: 10.3389/fgdh.2021.662343.
- [20] Y. E. Almalki *et al.*, “A novel method for COVID-19 diagnosis using artificial intelligence in chest X-ray images,” *Healthcare (Switzerland)*, vol. 9, no. 5, p. 522, Apr. 2021, doi: 10.3390/healthcare9050522.
- [21] T. Rahman *et al.*, “Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization,” *IEEE Access*, vol. 8, pp. 191586–191601, 2020, doi: 10.1109/ACCESS.2020.3031384.
- [22] P. Viradiya, “COVID-19 radiography dataset,” *Kaggle*, Accessed: Jan. 20, 2023. [Online]. Available: <https://www.kaggle.com/datasets/preetviradiya/covid19-radiography-dataset>
- [23] E. Vantaggiato, “Covid-19 X-ray - Two proposed databases,” *Kaggle*, Accessed: Jan. 20, 2023. [Online]. Available: <https://www.kaggle.com/datasets/edoardovantaggiato/covid19-xray-two-proposed-databases>
- [24] Y. Kubota, “keisen/tf-keras-vis: Neural network visualization toolkit for tf.keras,” *GitHub*, Accessed: Jan. 20, 2023. [Online]. Available: <https://github.com/keisen/tf-keras-vis>
- [25] A. A. Odeh, “Explaining-transfer-learning-models-for-the-detection-of-COVID-19-on-X-ray-lung-images: This is the code for the under-progress paper with the title of: Explaining transfer learning models for the detection of COVID-19 on X-ray lung images,” *GitHub*, Accessed: Jan. 23, 2023. [Online]. Available: <https://github.com/AbdAlRahman-Odeh-99/Explaining-Transfer-Learning-Models-for-the-Detection-of-COVID-19-on-X-ray-Lung-Images>

BIOGRAPHIES OF AUTHORS



Abd Al-Rahman Odeh    received the BSc. degree in software engineering from Jordan University of Science and Technology, Jordan, in 2021 and the MSc. degree in data science from the same institution, in 2023. Currently, he is an instructor of computer science at the School of Engineering and Computing at the American International University in Kuwait. His research interests include artificial intelligence, machine learning, deep learning, computer vision, and natural language processing. He can be contacted at email: a.odeh@aiu.edu.kw or atodeh20@cit.just.edu.jo or abdalrahmanodeh1@gmail.com.



Ahmad Mustafa    received the Ph.D. degree in computer science from the University of Texas at Dallas in 2018. He is currently an assistant professor at the Jordan University of Science and Technology. His current research interests include continual learning, active learning, natural language processing, explainable artificial intelligence, and deep learning. He can be contacted at email: ammustafa@just.edu.jo.