Heliyon 9 (2023) e15137

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

An XAI approach for COVID-19 detection using transfer learning with X-ray images

Salih Sarp [a], Ferhat Ozgur Catak [b], Murat Kuzlu [c], Umit Cali [d,*], Huseyin Kusetogullari [e], Yanxiao Zhao [a], Gungor Ates [f], Ozgur Guler [g]

[a] Electrical & Computer Engineering, Virginia Commonwealth University, Richmond, VA, USA
[b] Department of Electrical Engineering & Computer Science, University of Stavanger, Rogaland, Norway
[c] Old Dominion University, Batten College of Engineering & Technology, Norfolk, VA, USA
[d] Department of Electric Power Engineering, Norwegian University of Science and Technology, Trondheim, Norway
[e] Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden
[f] Department of Pulmonary Medicine, Private Genesis Hospital, Diyarbakir, Turkey
[g] eKare, Inc Fairfax, VA, USA

ARTICLE INFO

Dataset link: https://github.com/ieee8023/covid-chestxray-dataset

Dataset link: https://github.com/v7labs/covid-19-xray-dataset

Keywords:
COVID-19
Explainable artificial intelligence
Transfer learning

ABSTRACT

The coronavirus disease (COVID-19) has continued to cause severe challenges during this unprecedented time, affecting every part of daily life in terms of health, economics, and social development. There is an increasing demand for chest X-ray (CXR) scans, as pneumonia is the primary and vital complication of COVID-19. CXR is widely used as a screening tool for lung-related diseases due to its simple and relatively inexpensive application. However, these scans require expert radiologists to interpret the results for clinical decisions, i.e., diagnosis, treatment, and prognosis. The digitalization of various sectors, including healthcare, has accelerated during the pandemic, with the use and importance of Artificial Intelligence (AI) dramatically increasing. This paper proposes a model using an Explainable Artificial Intelligence (XAI) technique to detect and interpret COVID-19 positive CXR images. We further analyze the impact of COVID-19 positive CXR images using heatmaps. The proposed model leverages transfer learning and data augmentation techniques for faster and more adequate model training. Lung segmentation is applied to enhance the model performance further. We conducted a pre-trained network comparison with the highest classification performance (F1-Score: 98%) using the ResNet model.

1. Introduction

The COVID-19 outbreak has been the most significant pandemic of the 21st century [1], with hundreds of millions of reported cases and over five million deaths worldwide as of 2021 [2]. Though reverse transcription-polymerase chain reaction (RT-PCR) is the reference standard method to identify patients with a COVID-19 infection, Chest X-ray (CXR) and Computed Tomography (CT) have been extensively used in diagnosis, monitoring, and treatment decisions regarding COVID-19 cases [3–5]. Pneumonia is the most common radiological manifestation of COVID-19, which can be detected using CXR images [6,7]. Many thoracic imaging societies

* Corresponding author.
E-mail addresses: sarps@vcu.edu (S. Sarp), f.ozgur.catak@uis.no (F.O. Catak), mkuzlu@odu.edu (M. Kuzlu), umit.cali@ntnu.no (U. Cali), huseyin.kusetogullari@bth.se (H. Kusetogullari), yzhao7@vcu.edu (Y. Zhao), gungorates@gmail.com (G. Ates), oguler@ekare.ai (O. Guler).

https://doi.org/10.1016/j.heliyon.2023.e15137
Received 8 March 2022; Received in revised form 19 June 2022; Accepted 27 March 2023

Available online 7 April 2023

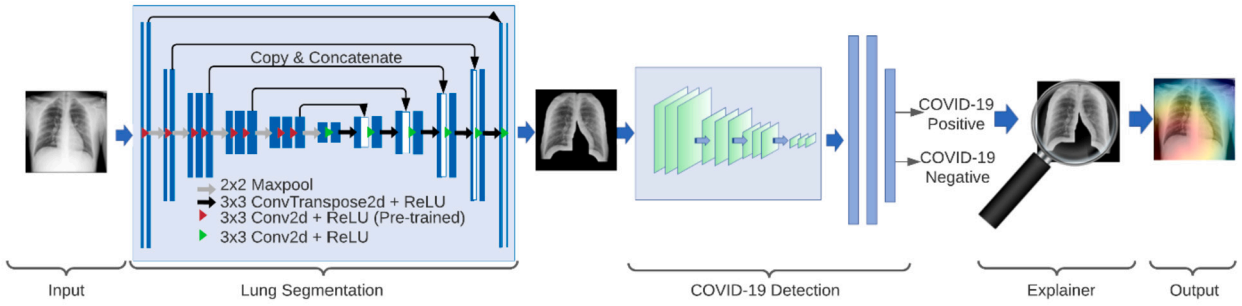


Fig. 1. The proposed model framework includes input, segmentation, detection, explainer, and output.

like the Radiological Society of North America state that routine CT for the identification of COVID-19 pneumonia is currently not recommended in the diagnosis of COVID-19 unless the patient is seriously ill [8–10]. Moreover, X-ray images are preferable for COVID-19 case detection because they are captured faster at low cost and are more readily available than CT images [10–12]. Manual diagnosing pneumonia in X-ray images is a challenging, time-consuming process and has poor diagnostic performance [13], [14]. Recently, a variety of Machine Learning (ML) based COVID-19 detection methods using X-rays have been developed and implemented [15]. Pneumonia with X-rays could be detected as the first stage of COVID-19 disease [16]. Some models use AI and computer vision associated with the CXR imagery of patients to identify if the patients are diagnosed as COVID-19 positive [17]. The second stage of the analysis is designed to detect if the pneumonia is caused by COVID-19. Antagonistically, modern technology such as AI with imagery inputs like heatmaps and similar data can be used by physicians as decision support tools to minimize human errors and increase diagnosis efficiency [18,19]. For several decades, AI has been used by academia and industry; it is inspired by human mental learning by mimicking the brain's cognitive features to learn and make decisions like a human artificially.

Traditional AI models function as black-box models for most researchers and professionals using them for various tasks, including medical diagnostic purposes. Such traditional AI methods lack details and explanations to help physicians make better decisions and interpretations. Explainable AI (XAI) provides this opportunity, which transfers the AI-based black-box models to more explainable and transparent gray-box models. The major limitations of all the mentioned methods are that they cannot: 1) analyze the level of COVID-19 cases and 2) provide sufficient insights regarding model details. This study proposes a model for COVID-19 case detection and its interpretation using XAI, depicted in Fig. 1. The contributions of the proposed framework are:

(i) Detection and classification of COVID-19 cases from affordable CXR images,
(ii) Automatic interpretation of COVID-19 cases using a LIME-based heatmap implementation with XAI from X-ray images to assist clinicians and radiologists.

The proposed model utilized lung segmentation, transfer learning, and data augmentation technique for faster and adequate model training. A pre-trained network comparison was performed where the ResNet model achieved the highest classification performance (F1-Score: 98%).

The remaining sections of this paper are organized as follows. Related works are introduced in Section 2. Section 3 presents data collection and processing steps and the validation methods. Section 4 provides details about the methodology and the implementation of the proposed model. The results and related discussions are examined in Section 5. New trends and future work are provided in Section 6. Conclusions are drawn in Section 7.

2. Related works

Recently, there have been many studies that utilize ML techniques to combat COVID-19 pandemic. For instance, the authors in [20] proposed multi-level thresholding with a Support Vector Machine (SVM) classifier for the early detection of COVID-19 cases. Firstly, features were extracted using a multi-level thresholding technique. After that, an SVM classifier was applied to the extracted features of 40 contrast-enhanced CXR images, and classification accuracy was obtained at 97%. In another study [21], the authors applied an improved SVM classifier to detect COVID-19 cases. They collected an image dataset from 235 patients, of which 43% were confirmed COVID-19 cases. Five ML algorithms, i.e., logistic regression, random forests, gradient boosting trees, neural networks, and SVM, were trained with 70% of the dataset and evaluated their performances with 30%. The results showed that the SVM classifier performs the best in detecting COVID-19 cases compared with other conventional ML methods with an accuracy of 85%. In [22], Random Forest and XGBoost algorithms were applied to the X-ray images to detect COVID-19 cases. The results showed that XGBoost, with an accuracy of 97.7%, provides similar performance to the Random Forest method, with an accuracy of 97.3%. Advanced learning methods based on Convolutional Neural Networks (CNN) have also been proposed and employed to detect COVID-19 cases using X-ray images to overcome the limitation of the conventional ML approaches. Ozturk et al. [23] proposed a Deep Learning (DL) model for the early detection of COVID-19 cases using X-ray images. The proposed model consists of 17 convolutional layers and five pooling layers using Maxpool. Moreover, these layers have different filter numbers, sizes, and stride values. The model was employed on 1125 X-ray images, including 125 for the COVID-19 class, 500 for the pneumonia class, and 500 for the

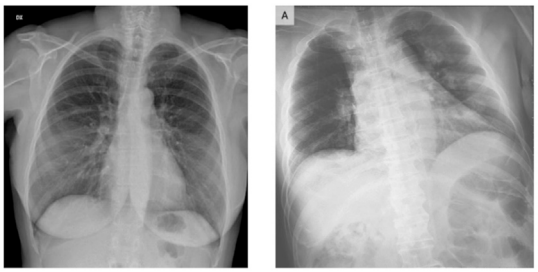
normal class. The model provided a classification accuracy of 98.08% for binary classes and 87.02% for multiclass classification. Toraman et al. [24] proposed Convolutional Capsule Network architecture (CapsNet) to detect COVID-19 cases using CXR images. The method was applied to a dataset containing X-ray images containing COVID-19 [25], No-Findings, and pneumonia [26]. The results showed that the CapsNet approach provides highly accurate diagnostics for COVID-19 with 97.24% and 84.22% for binary and multiclass classification, respectively. In [27], a CNN model has been designed and developed using EfficientNet architecture to automatically diagnose COVID-19 cases with X-ray images. The proposed model uses EfficientNet with 10-fold stratified cross-validation, which was applied to classify binary multiclass cases using X-ray images containing COVID-19, pneumonia, and normal patients. The proposed method achieved an average recall result of 99.63% and 96.69% for binary and multiclass classification, respectively. A DL-based ML method has been developed by Apostolopoulos et al. to detect COVID-19 cases [28]. The method was applied for both binary and multiclass analysis, and they used a dataset composed of 224 COVID-19 X-rays, 700 bacterial pneumonia, and 500 no-findings images. The proposed model finds high accuracy results, which are 98.78% and 93.48% for binary (COVID-19 vs. No-findings) and multiclass (COVID-19 vs. No-findings vs. pneumonia), respectively. Moreover, Hemdan et al. [29] designed a COVID-19 case detection method based on DL using X-ray images, and the proposed method was compared with seven other DL-based COVID-19 case detection methods. The method was performed for only binary class classification, and an accuracy rate of 74.29% was estimated. Three different automated COVID-19 case detection methods have been developed based on three different DL models, which are ResNet50, InceptionV3, and InceptionResNetv2 in [30]. The developed methods were applied for binary class classification only, and the highest accuracy rate was achieved by ResNet50 with an average of 98%. Islam et al. [31] proposed a combination of two different methods, which are CNN and long short-term memory (LSTM), for detecting COVID-19 cases using X-ray images. In the proposed approach, CNN was first applied to the X-ray images to extract the features. After the obtained features were used by LSTM to classify COVID-19 cases. The method was performed on a collection of 4,575 X-ray images, including 1525 images of COVID-19. The experimental results indicated that the CNN-LSTM performs better than the state-of-the-art methods with an accuracy of 99.4%. Loey et al. [32] used a Generative Adversarial Network (GAN) with deep transfer learning to diagnose COVID-19 from X-ray images. The proposed approach used three different transfer learning pre-trained models, i.e., AlexNet, GoogleNet, and RestNet18. The method was performed on a collection of datasets consisting of 69 COVID-19, 79 pneumonia bacterial, 79 pneumonia viruses, and 79 normal cases. The experimental results showed that using GAN with pre-trained GoogleNet provides the highest accuracy rate with 99.9% for binary class classification problems. Bandyopadhyay et al. [33] developed a hybrid model based on two different ML methods, LSTM and Gated Recurrent Unit (GRU), to detect COVID-19 cases automatically. The proposed method obtained 87% accuracy for the confirmed COVID-19 cases. A DL method was presented in [34] to automatically classify COVID-19 cases from CXR. The proposed model achieved an accuracy of 89.5%, a precision of 97%, and a recall of 100% for COVID-19 cases. In [35], a multi-dilation DL approach (CovXNet) for automatic COVID-19 and other pneumonia case detection from CXR images was proposed. Experiments were performed on two different datasets to evaluate the performance of the CovXNet. The first dataset consisted of 5,856 X-ray images, and another dataset contained 305 X-ray images of different COVID-19 patients. The results showed that the CovXNet method achieved an accuracy of 97.4% for COVID/Normal detection and an accuracy of 96.9% for binary class, and 90.2% for multiclass classification. Other DL methods have been designed and developed based on different pre-trained models such as VGG16, VGG19, ResNet50, DenseNet121, Xception, and capsule networks [36–42]. Generally, existing approaches attempt to resolve binary and multiclass COVID-19 cases classification problems.

3. Data collection, preprocessing, and validation

The collected images for COVID-19 cases were preprocessed using various methods. This section explains them, including the data collection, preprocessing, validation, and test/computational environment.

3.1. Data collection

The publicly accessible GitHub dataset of CXR and CT images for lung disease patients suspected of having COVID-19 or other viral and bacterial conditions such as MERS, SARS, and ARDS is available [25]. This dataset is gathered from both public sources and indirectly from hospitals and physicians [43]. In this research, the X-ray scan images have been used to create an XAI-based COVID-19 detection model, while the CT images have been disregarded. Also, low-quality photos and pictures with foreign objects (metals, cables, etc.) were omitted. First, the selected X-ray scan images were rescaled to 512×512. Second, various image enhancement techniques were applied to produce enhanced input images, including flipping (right/left and up/down), rotation and translation with five random angles. Our previous work [44] had only 50 positive and 50 negative X-ray scan images for the training and 20 positive and 20 negative samples for testing. In this study, we have benefitted from the existing data repositories to extend and improve the classifier's performance and added an explainer. Another issue with the dataset is class distribution. This dataset contains X-ray scan images from those infected by COVID-19 and other diseases. There were only three records in the dataset with COVID-19 negative samples. Therefore, X-ray pictures with ARDS and Streptococcus results were labeled as COVID-19 negative samples. In this study, 6,000 images are collected from the GitHub repositories mentioned above to increase the number of training and testing samples in the dataset. Thus, an analysis using different neural network structures is conducted, similar to the previous study. This study created a binary classification model by marking labels other than COVID-19 as the "0" class. Of these 6,000 samples, 5,500 are COVID-19 negative, and the rest are COVID-19 positive X-ray images. Furthermore, 1,200 of them were used for testing, and the remaining 4,800 for training.



(a) (b)

Fig. 2. X-ray scan image example from the dataset, with (a) a label of COVID-19 negative and (b) COVID-19 positive.

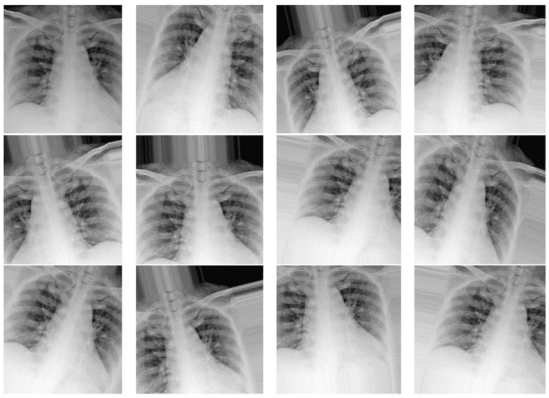


Fig. 3. Example of data augmentation. The first image is the original, and the remaining images are augmented versions.

Table 1
Confusion matrix prediction.

		Prediction	
		$y' = 0$	$y' = 1$
True Label	$y = 0$	True Negative	False Positive
	$y = 1$	False Negative	True Positive

3.2. Data preprocessing

Collected raw data were processed with various techniques to increase the model's classification performance. Samples of the dataset are depicted in Fig. 2.

To improve the classification model's performance and increase the number of samples in the dataset, various image augmentation techniques were employed. The parameters used were a rotation range of 20 degrees, zoom range of 15, width shift range of 0.2, height shift range of 0.2, shear range of 0.15, and horizontal flipping. An example of the image augmentation techniques is illustrated in Fig. 3.

3.3. Validation

To assess the performance of the COVID-19 detection part of the framework, we utilized the following performance metrics, F1-Score, recall, precision, and accuracy extracted from the confusion matrix shown in Table 1.

Performance measures are given in Eqs. (1)–(4) below.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{True Negative} + \text{True Positive}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \tag{4}$$

Evaluation of the explanation part of the framework was performed by an MD specialized in COVID-19 by reviewing the test images. Generated heatmaps were reviewed and evaluated one by one by the medical professional.

3.4. Test computational environment

The proposed COVID-19 XAI framework in this study was implemented using the Keras DL framework built with Python 3.6. The workstation – an Intel® Core™ i7-8700 processor @3.20 GHz – was used to run AI-based models, which has 32 GB memory and GTX 1080 GPU (NVIDIA GeForce). The classification part of the model is trained for 60 epochs, whereas the lung segmentation part is trained for 50 epochs since the accuracy and loss values do not improve notably after these epochs. The complete training of the framework took around 3 hours. We used a constant learning rate of 0.001 and the "RMSprop" optimizer to train classification parts. The segmentation is realized using an "adam" optimizer with a 0.005 learning rate. XAI module training takes around 1 minute on average.

4. Methodology and implementation

COVID-19 detection has been explored in many studies [45]. In contrast to existing approaches, lung segmentation is utilized in the proposed pipeline to enhance the COVID-19 detection and explanation tasks. It forces the system to better learn the features inside the lung region with improved training. Further, it provides greater emphasis on the lungs during the explanation phase of the framework. After this step, transfer learning is adopted to speed up and ease the feature extraction since the number of X-ray images is limited. Following feature extraction, X-rays are classified, and the classification performance of the model is measured. The framework's third stage focuses on explaining the COVID-19 cases. The COVID-19 positive cases are then fed into the LIME-based heatmap explanation part of the pipeline to spotlight the areas with COVID-19 pneumonia to help physicians during the diagnosis in a non-invasive manner. The following sections discuss lung segmentation, transfer learning models for classification, and XAI tools used in this study.

4.1. Lung segmentation

Anomalies in the lung provide information about many diseases. Our study examines the CXR to determine whether the patient has COVID-19. An additional lung segmentation part is added to the proposed pipeline to increase the performance of the detection and explanation part of the proposed model. Manual segmentation is time-consuming and not available for many biomedical applications. Also, human annotations are prone to cause inconsistencies as well as to make mistakes. With lung segmentation, COVID-19 detection and its explanation networks are fed with masked lung images, which force these parts to detect and explain only in the lung section of the CXR. The output of the explanation part is shown in the whole CXR for a better interpretation.

A reference hybrid U-Net [46] architecture that uses a pre-trained VGG11 feature extractor is utilized in the encoder part of U-Net to obtain lung segmentation. Pre-trained networks on a large dataset, i.e., ImageNet, outperform the networks trained from scratch. The U-Net architecture consists of an encoder and decoder structure with skip connections to carry low-level feature maps to the decoder. Concatenating feature maps from encoder to decoder improves the performance and convergence of the network. To train this model, publicly available lung segmentation images are used [47], and various augmentation techniques such as horizontal and vertical shift, minor zoom, and padding are applied. The lung segmentation model has a Jaccard index of 92% and a dice score of 96%. Additionally, morphological operations, i.e., dilation, are implemented to ensure the correct segmentation of the lungs with a kernel size of 90×30 and three iterations using the OpenCV function, cv2.dilate (bc, kernel, iterations = 3). Our aim is to segment and interpret all parts of the lung in X-ray images using the proposed approach. Including the perimeter outside the lung does not have an important effect, but the entirety of the lung has a great impact on classification and especially on explanation tasks. Therefore, we have applied dilation operations. Lung segmentation is used as a pre-processing method to increase the performance of COVID-19 classification and explanation tasks. The comparison of the model results with/without lung segmentation is summarized in Table 2.

4.2. Transfer learning

Transfer learning is a widely used technique in machine learning that simplifies the process of building models. It involves utilizing knowledge gained from a previously trained model on a related task and applying it to a new, but related problem. This approach is based on extracting features from the input data obtained from a related initial task and transferring these features to the new task to improve accuracy and reduce training time. Specifically, pre-trained DL network models have already been trained on large datasets and have achieved high accuracy, and these pre-trained models can be used as a starting point for the new task. The transfer learning process starts with the previously learned patterns while solving a different problem. Further, it decreases the time-consuming training process and enables the creation of a model with high classification performance. These pre-trained models are based on Deep Evolution Neural Networks (DENN). In deep learning, this method involves initially training a CNN for a classification problem using large-scale training datasets. Since a CNN model can learn to extract the image's discriminative features, the availability of initial training data is an essential part of successful training. The model performance evaluation depends on the model's fitness for transfer learning, which relies on CNN's capacity to select the most important image features.

The VGG-Net model was used in both the segmentation and detection part of this framework, which was developed with a tiny convolution in the neural network by Simonyan et al. [46]. Compared to previous models, the most significant difference is the widespread use of CNN models due to their deeper structure, which typically includes multiple layers of convolution and association. This model consists of nearly 138 million parameters. VGG is one of the popular networks, which is trained with more than a million images from the ImageNet dataset with 1,000 different classes. Therefore, the model can be applied as a helpful tool for the feature extractor of new images.

The ResNet (residual networks) won the ImageNet challenge in 2015 and was proposed by He et al. [48] with a paper titled "Deep Residual Learning for Image Recognition". The version that is used in this model has 50 neural network layers and was trained on the ImageNet dataset having 1,000 different classes. Increased layer quantity brings some challenges, such as model complexity and vanishing gradient. ResNet was inspired by the VGG networks, but it has fewer filters and less complexity. The vanishing gradient problem was mitigated by the skip connections, which allow gradients to flow through alternative paths. This method is the core concept in residual blocks of ResNet to alleviate the vanishing gradient problem. The ResNet50 model has more than 23 million parameters.

The Inception V3 model was developed by Szegedy et al. with a paper titled "Rethinking the Inception Architecture for Computer Vision" published in 2015 [49]. This iteration of the inception architecture is more computationally efficient than the previous models. Larger convolutions are changed within parallel smaller convolutions. Additionally, factorized convolutions and an auxiliary classifier are utilized to improve the model's performance. A new grid size reduction was proposed to combat bottlenecks of expensive computation.

Our study used VGG16, VGG19, ResNet, and Inception V3 neural network models to improve the performance of our X-ray image based on the COVID-19 detection model.

4.3. Explainable artificial intelligence (XAI)

AI has diverse applications and provides unprecedented advantages, such as higher efficiency and broader data analysis, for many daily tasks such as manufacturing, finance, and entertainment [50,51]. However, the use of AI is lagging in high-risk systems, especially in healthcare [52]. The inner workings of AI systems comprise complicated mathematical and statistical processes, which are not interpretable. Fortunately, a black-box AI model can be converted to a glass-box model by applying explainable AI tools.

AI models are converted to more understandable systems by making them interpretable or comprehensible. Shallow Learning (SL) methods such as decision trees and regression algorithms are more transparent as the mathematical backgrounds are well-defined and studied. These methods are interpreted by utilizing the underlying math. On the other hand, the inner workings of DL methods, such as CNNs and Recurrent Neural Networks (RNNs), are conceived by finding the relationship between the inputs and the outputs. DL methods consist of nodes and weights associated with the inputs and outputs. This relationship should be clarified to mitigate risks and build trust in AI models for enhanced adoption. A comprehensive program driven by the DARPA showed that XAI improves user trust significantly and increases user adoption through the provided explanation [53,54].

Grad-Cam, Tylor decomposition, and LIME are some of the XAI tools used to make the AI models more understandable. Our model provides a LIME-based heatmap explanation method to detect and find COVID-19 in CXR scans.

In this study, the LIME model-independent general XAI method is used, which finds the statistical connection between the inputs and the outputs of the models. Inputs are perturbed during the training of local surrogates to understand their effects on the output instead of globally training them. This process results in the instance of an interpretable representation and visualization. The mathematical definition of LIME is given in the equation as follows:

$$Explanation(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{5}$$

Where:

x: An instance of out of the data space for which we desire an explanation for its predicted target value

$L(f, g, \pi_x)$: Fidelity function that measures how unfaithful the g is while approximating to f in the locality defined by the π_x. It is the locality-aware loss.

$\Omega(g)$: Measures the model complexity of the explainer (g)

f: The black-box model to be explained

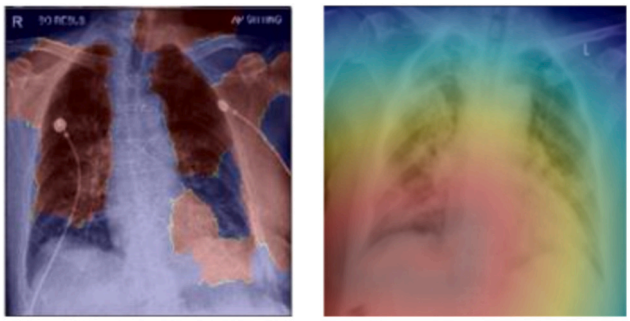
g: The explainer

G: The total set of interpretable models

π_x: Proximity measure

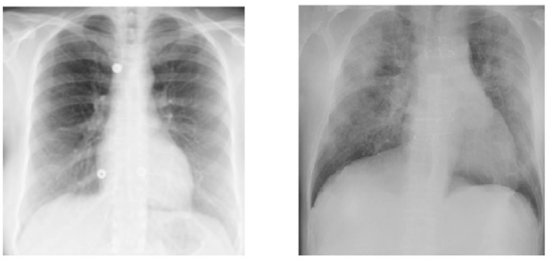
Locality aware loss ($L(f, g, \pi_x)$) is minimized for local faithfulness with low complexity of the second term ($\Omega(g)$) for interpretability.

After obtaining the explanations, the areas determined by the LIME are fed into the heatmap creation part of the explainer. These areas are highlighted with heatmaps to spotlight the areas with COVID-19 pneumonia to provide additional information to the physicians during the diagnosis of COVID-19. In comparison to the regular LIME output [55], our XAI part takes the LIME output one step further to better localize the COVID-19 affected areas as shown in Fig. 4.



a) b)

Fig. 4. a) Regular LIME [55] and b) our proposed LIME-based XAI model output on the right.



a) b)

Fig. 5. a) Healthy chest X-ray [56] and b) Chest X-ray of a patient with COVID-19 [36].

Table 2
Performance results for each pre-trained CNN model for the COVID-19 classification model (weighted average).

Model	Segm.	Accuracy	Precision	Recall	F1-Score
VGG16	Yes	0.96	0.96	0.96	0.96
VGG19	Yes	0.96	0.96	0.89	0.92
ResNet	Yes	0.98	0.98	0.98	0.98
InceptionV3	Yes	0.91	0.91	0.91	0.90
ResNet	No	0.96	0.96	0.89	0.92
COVID-Net [36]	No	0.93	0.99	0.91	0.95
CORODET [57]	No	0.99	0.98	0.95	0.97

5. Results and discussions

Under X-rays, dense structures such as bones and metal blocks are seen as white since they block the X-rays. Less dense areas appear in tones of gray, and the least dense regions, such as lungs, will be black. Healthy and COVID-19 CXR images are shown in Fig. 5.

CXR images of healthy lungs are shown as black (see Fig. 5a), whereas the areas with COVID-19 are shown in white. The complications in the lungs can be detected by examining the CXR. A CXR with COVID-19 has more white areas spread over the lungs, as shown in Fig. 5b.

Classification performances of seven DL models with/without transfer learning and with/without segmentation are given in Table 2. The models (VGG16, VGG19, ResNet, InceptionV3) are developed using transfer learning with getting lung-segmented CXR images as inputs. Performance results of another pre-trained ResNet model without lung segmentation and the model without transfer learning are also compared in Table 2. Weighted averages are used to take the sample size into consideration to provide a more accurate representation of the averages during the calculations.

In our previous studies [44,58], we investigated whether the transfer learning models could be used in detecting COVID-19 positives in the X-ray data to increase the model's classification performance without XAI, which we now focus on in the present study. The results showed we could detect COVID-19 from X-ray images in a similar manner to the current techniques for other imaging approaches, although this was more limited than explainable. In this study, we used the cognitive learning approach to confirm and measure the quality of the predictions based on those of medical doctors, which provides a generalizable model to evaluate the validity of the projections in the X-ray data and the use of the XAI techniques. Table 2 indicates that the pre-trained ResNet transfer model is better suited for detecting COVID-19 with 98% accuracy. The VGG16 and VGG19 models indicate a similar classification performance with an accuracy of 96%. The InceptionV3 model had the lowest classification accuracy at 91%. The

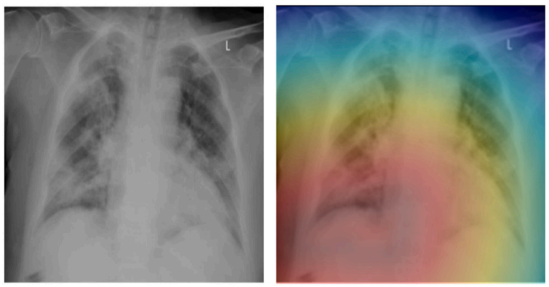


Fig. 6. The proposed method detects and highlights the affected regions of the COVID-19 CXR images. Lung areas with heavily infected areas with pneumonia are depicted with warmer colors.

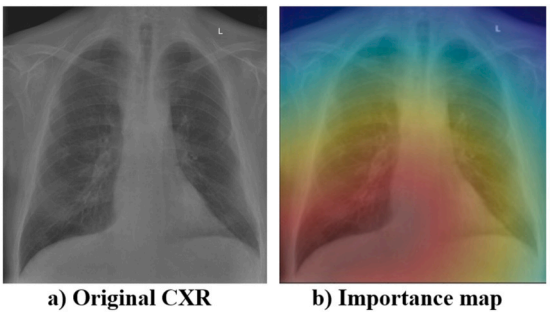


Fig. 7. The CXR image shows the regions with pneumonia in both lungs.

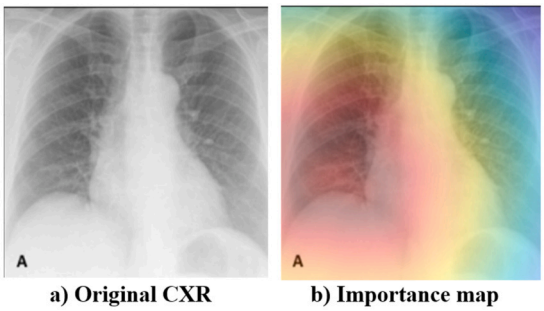


Fig. 8. The CXR image indicates the infected areas with pneumonia.

VGG16, VGG19, ResNet, and InceptionV3 models performed comparably in COVID-19 classification with F1 scores of 0.96, 0.92, 0.98, and 0.90, respectively. The ResNet model presents the highest performance in classification, i.e., 98% F1-score. It also suggests that the high F1 classification metric performance is more effective in predicting the best classification performance.

Additionally, Table 2 compares models with/without transfer learning and lung segmentation, i.e., the ResNet model without lung segmentation and other studies without transfer learning [36,57]. This comparison indicates that transfer learning provides better resource management and improved efficiency during training. Models without transfer learning require far more complicated models and high-performance computational sources, which also consume more time. However, models with transfer learning do not rely on high computational power as much as those without transfer learning. The previously trained networks are taken as a tool to solve the problem at hand efficiently and faster without expensive and extensive resources. Our proposed model outperforms both of the models without transfer learning [36,57]. In addition, the proposed model has better performance than ResNet without lung segmentation.

Fig. 6a shows a COVID-19-positive CXR image. The lung on the left has a pattern, with whiter areas showing more dense pneumonia regions. These areas are well detected and highlighted with a heatmap by the proposed method in Fig. 6b.

The COVID-19-positive patient CXR image is highlighted in Fig. 7. The model correctly identified the disease and stressed the infected regions by pneumonia for better understandability in a low-quality image.

The pneumonia infection areas are depicted and highlighted in Fig. 8a and Fig. 8b, respectively.

Another CXR image indicating and highlighting the areas with pneumonia is shown in Fig. 9 with cables and other medical devices. Our model is resilient to these external elements.

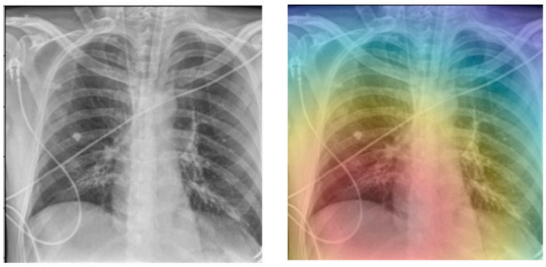


Fig. 9. The CXR image of a COVID-19 patient shows the infected areas with pneumonia.

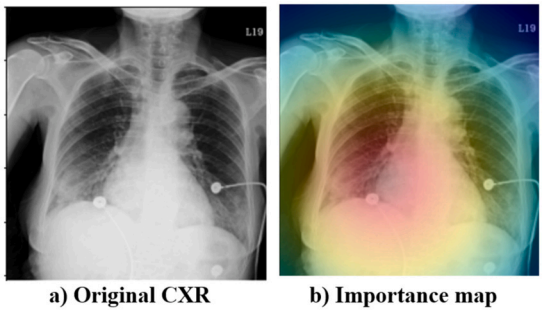


Fig. 10. Areas with pneumonia are highlighted by the proposed model.

The proposed model successfully identifies the COVID-19 positive case; however, the explanation of the CXR image is slightly off the correct place in Fig. 10. The heatmap indicates the infected areas with the perimeter of the lung on the left. The whiter areas on the lung perimeter are considered pneumonia.

The deficiencies due to the reflections were eliminated by introducing lung segmentation to the pipeline. The heatmap provided a better explanation for pneumonia areas where it is impossible to notice the color difference with the naked eye. Also, for most X-rays, it is evident that there is still room for improvement with the use of extensive data collection and labeling. The limited color change is observed on one side of the lung, which could be improved as well. In addition, it can be studied for different color scales that will create more contrast of the increased different heat zones specific to the limited area with the lesion in the heatmap. Many countries and regions in the world cannot access tomography. For these, direct radiographs can be the only diagnostic tool. Furthermore, computed tomography cannot be repeated one by one due to high radiation exposure. That's why a well-working heatmap can be very helpful. With the help of this study, even those who are not very experienced can easily see the area where the lesions are. The following observations are deducted from the overall results:

Observation 1. COVID-19 could be detected by AI using transfer learning with higher performance (F1-Score: 98%).

Observation 2. Application of an XAI tool, i.e., LIME, makes the COVID-19 detection AI model more understandable.

Observation 3. The XAI application on CXR images has a high potential for faster diagnosis and prognosis of COVID-19.

Observation 4. The treatment of COVID-19 can be tracked more easily by applying the proposed model to the CXR images. We would like to emphasize that many learning-based methods have been designed and proposed to classify COVID-19 cases, and these methods have been compared with those of radiologists. The results show that learning-based systems provide better results in terms of precision and time [59–61].

Observation 5. The application of XAI methods enables the adoption of AI applications in high-risk industries such as healthcare.

Observation 6. Segmenting lung images as a preprocessing step improves the COVID-19 detection performance and its explanation.

6. New trends and future work

Under normal conditions, a well-trained physician can detect a pneumonia case by looking at the CXR and providing diagnosis results relatively quickly without investigating thousands or millions of X-ray images. This study is motivated by the power of mental modeling, which is performed by the human brain to understand the concept of pneumonia and its causes by comprehending the

associated facts such as human anatomy, fundamentals of virology, how lungs and ribs function, and other information learned during their medical education at school or in a clinic. Researchers and engineers have developed various XAI tools to help professionals and academia improve their understanding and insights regarding AI-based implementations. The following XAI tools have significant potential to be implemented primarily in the field of medical sciences, where image processing and clustering play an important role:

- Local Interpretable Model-Agnostic Explanations (LIME)
- Class activation mapping (CAM)
- Deep SHapley Additive exPlanations (Deep SHAP)
- LRP (Layer-wise Relevance Propagation)
- DeepLIFT (Deep Learning Important FeaTures)

7. Conclusion

This paper presents an XAI approach for COVID-19 diagnosis using transfer learning with CXR images. The proposed model supports decision-making for COVID-19 cases, i.e., positive and negative. The framework accepts a CXR image as the input and predicts the COVID-19 classification and its explanation as the output. To improve the classification and explanation performances, a lung segmentation model is realized, and its output, segmented lung images, is fed to the framework. An XAI approach, i.e., LIME, is used to faithfully describe predictions of COVID-19 cases in an interpretable manner through heatmaps. The proposed model is also extended through the LIME and heatmap methods to offer better explainability. XAI tools help non-expert end-users understand the black box AI model by providing explainability and transparency. It provides feedback to the end-user and explains, i.e., by providing more information and tracing the insight right back to the inner workings of the black box AI model.

The internal working of AI models, especially the deep learning models, are black box concepts that cannot be explained why the AI model outputs a specific result. Our model requires lung segmentation before classification and explanation, extending the overall processing time. It is also trained on a limited number of CXR images. Additional data, CXR images, will increase the robustness and performance of classification. In addition to this, our model's XAI part has limitations while interpreting the CXR images. Our model first needs to classify the COVID-19 CXR images from the healthy ones; then, it will provide heatmaps indicating the areas with COVID-19 pneumonia. When a healthy CXR image is given to the XAI part of the model pipeline, the model tries to find COVID-19-affected areas and provides the highlighted results, which are the closest COVID-19-affected areas. Therefore the classification part is also critical in this project.

This study demonstrated how XAI techniques could be helpful for COVID-19 diagnosis in the healthcare domain when assessing trust and gaining insights into predictions. The proposed hybrid model provides two outputs: (1) COVID-19 diagnosis and (2) Model-decision explanation. Interpretation of the results obtained from the proposed XAI module offers adequate information about COVID-19 diagnosis. This study is expected to benefit researchers and physicians working on COVID-19 diagnosis or related studies by providing insight into XAI's potential.

CRediT authorship contribution statement

Salih Sarp, Ferhat Ozgur Catak, Murat Kuzlu: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Umit Cali, Huseyin Kusetoglu, Ozgur Guler, Yanxiao Zhao: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Gungor Ates: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data associated with this study has been deposited at https://github.com/ieee8023/covid-chestxray-dataset, and https://github.com/v7labs/covid-19-xray-dataset.

References

[1] H. Abid, J. Mohd, V. Raju, Effects of COVID 19 pandemic in daily life, Curr. Med. Res. Pract. (2020).
[2] WHO, WHO coronavirus (COVID-19) dashboard, https://covid19.who.int/.

[3] U.S. Food, Drug Administration, Accelerated Emergency Use Authorization (EUA) Summary SARS-CoV-2 Assay (Rutgers Clinical Genomics Laboratory), FDA, US, 2020, pp. 1–8.
[4] M.T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, arXiv preprint, arXiv:1606.05386, 2016.
[5] Z.Y. Zu, M.D. Jiang, P.P. Xu, W. Chen, Q.Q. Ni, G.M. Lu, L.J. Zhang, Coronavirus disease 2019 (COVID-19): a perspective from China, Radiology 296 (2) (2020) E15–E25, Radiological Society of North America.
[6] J. Ker, L. Wang, J. Rao, T. Lim, Deep learning applications in medical image analysis, IEEE Access 6 (2017) 9375–9389.
[7] S.S. Yadav, S.M. Jadhav, Deep convolutional neural network based medical image classification for disease diagnosis, J. Big Data 6 (1) (2019) 1–18.
[8] U.B. COVID, Guidance for the reporting radiologist, Br. Soc. Thorac. Imag. (2020).
[9] S. Simpson, F.U. Kay, S. Abbara, S. Bhalla, J.H. Chung, M. Chung, T.S. Henry, J.P. Kanne, S. Kligerman, J.P. Ko, Radiological society of north America expert consensus document on reporting chest CT findings related to COVID-19: endorsed by the society of thoracic Radiology, the American college of Radiology, and RSNA, Radiology 2 (2) (2020) e200152, Radiological Society of North America.
[10] G.D. Rubin, C.J. Ryerson, L.B. Haramati, N. Sverzellati, J.P. Kanne, S. Raoof, N.W. Schluger, A. Volpi, J.-J. Yim, I.B. Martin, The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society, Radiology 296 (1) (2020) 172–180, Radiological Society of North America.
[11] W.H. Self, D.M. Courtney, C.D. McNaughton, R.G. Wunderink, J.A. Kline, High discordance of chest x-ray and computed tomography for detection of pulmonary opacities in ED patients: implications for diagnosing pneumonia, Am. J. Emerg. Med. 31 (2) (2013) 401–405, Elsevier.
[12] R.R. Ayebare, R. Flick, S. Okware, B. Bodo, M. Lamorde, Adoption of COVID-19 triage strategies for low-income settings, Lancet Respir. Med. 8 (4) (2020) e22, Elsevier.
[13] R. Zhang, X. Tie, Z. Qi, N.B. Bevins, C. Zhang, D. Griner, T.K. Song, J.D. Nadig, M.L. Schiebler, J.W. Garrett, Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: value of artificial intelligence, Radiology 298 (2) (2021) E88–E97, Radiological Society of North America.
[14] S.H. Yoon, K.H. Lee, J.Y. Kim, Y.K. Lee, H. Ko, K.H. Kim, C.M. Park, Y.-H. Kim, Chest radiographic and CT findings of the 2019 novel coronavirus disease (COVID-19): analysis of nine patients treated in Korea, Korean J. Radiol. 21 (4) (2020) 494–500, The Korean Society of Radiology.
[15] Z. Sadiq, S. Rana, Z. Mahfoud, A. Raoof, Systematic review and meta-analysis of chest radiograph (CXR) findings in COVID-19, Clin. Imaging 80 (2021) 229–238.
[16] V. Perumal, V. Narayanan, S.J.S. Rajasekar, Detection of COVID-19 using CXR and CT images using transfer learning and haralick features, Appl. Intell. 51 (1) (2021) 341–358.
[17] M. Frid-Adar, R. Amer, O. Gozes, J. Nassar, H. Greenspan, Covid-19 in CXR: from detection and severity scoring to patient disease monitoring, IEEE J. Biomed. Health Inform. 25 (6) (2021) 1892–1903.
[18] M.M. Badža, M.Č. Barjaktarović, Classification of brain tumors from MRI images using a convolutional neural network, Appl. Sci. 10 (6) (2020) 1999.
[19] S. Sarp, Y. Zhao, M. Kuzlu, Artificial intelligence-powered chronic wound management system: towards human digital twins, 2022.
[20] L.N. Mahdy, K.A. Ezzat, H.H. Elmousalami, H.A. Ella, A.E. Hassanien, Automatic x-ray COVID-19 lung image classification system based on multi-level thresholding and support vector machine, MedRxiv Publisher: Cold Spring Harbor Laboratory Press, 2020.
[21] A.F. de Moraes Batista, J.L. Miraglia, T.H.R. Donato, A.D.P. Chiavegatto Filho, COVID-19 diagnosis prediction in emergency care patients: a machine learning approach, MedRxiv Publisher: Cold Spring Harbor Laboratory Press, 2020.
[22] R. Kumar, R. Arora, V. Bansal, V.J. Sahayasheela, H. Buckchash, J. Imran, N. Narayanan, G.N. Pandian, B. Raman, Accurate prediction of COVID-19 using chest X-ray images through deep feature learning model with SMOTE and machine learning classifiers, MedRxiv Publisher: Cold Spring Harbor Laboratory Press, 2020.
[23] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U.R. Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images, Comput. Biol. Med. 121 (2020) 103792, Elsevier.
[24] S. Toraman, T.B. Alakus, I. Turkoglu, Convolutional capsnet: a novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks, Chaos Solitons Fractals 140 (2020) 110122, Elsevier.
[25] J.P. Cohen, COVID-19 Image Data Collection, https://github.com/ieee8023/COVID-chestxray-dataset.
[26] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, 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.
[27] G. Marques, D. Agarwal, I. de la Torre Díez, Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network, Appl. Soft Comput. 96 (2020) 106691, Elsevier.
[28] I.D. Apostolopoulos, T.A. Mpesiana, Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks, Phys. Eng. Sci. Med. 43 (2) (2020) 635–640, Springer.
[29] E.E.-D. Hemdan, M.A. Shouman, M.E. Karar, Covidx-net: a framework of deep learning classifiers to diagnose COVID-19 in x-ray images, arXiv preprint, arXiv:2003.11055, 2020.
[30] A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks, Pattern Anal. Appl. (2021) 1–14, Springer.
[31] M.Z. Islam, M.M. Islam, A. Asraf, A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, Inform. Med. Unlocked 20 (2020) 100412, Elsevier.
[32] M. Loey, F. Smarandache, N.E.M. Khalifa, Within the lack of chest COVID-19 X-ray dataset: a novel detection model based on GAN and deep transfer learning, Symmetry 12 (4) (2020) 651, Multidisciplinary Digital Publishing Institute.
[33] S.K. Bandyopadhyay, S. Dutta, Machine learning approach for confirmation of COVID-19 cases: positive, negative, death and release, MedRxiv Publisher: Cold Spring Harbor Laboratory Press, 2020.
[34] A.I. Khan, J.L. Shah, M.M. Bhat, CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images, Comput. Methods Programs Biomed. 196 (2020) 105581, Elsevier.
[35] T. Mahmud, M.A. Rahman, S.A. Fattah, CovXNet: a multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization, Comput. Biol. Med. 122 (2020) 103869, Elsevier.
[36] L. Wang, Z.Q. Lin, A. Wong, COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images, Sci. Rep. 10 (1) (2020) 19549.
[37] P.K. Sethy, S.K. Behera, Detection of Coronavirus Disease (COVID-19) Based on Deep Features, MDPI AG, 2020.
[38] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K.N. Plataniotis, A. Mohammadi, Covid-caps: a capsule network-based framework for identification of COVID-19 cases from x-ray images, Pattern Recognit. Lett. 138 (2020) 638–643, Elsevier.
[39] M.J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, N. Shukla, X-Ray Image Based COVID-19 Detection Using Pre-trained Deep Learning Models, engrxiv, 2020.
[40] M. Singh, S. Bansal, S. Ahuja, R.K. Dubey, B.K. Panigrahi, N. Dey, Transfer learning–based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data, Med. Biol. Eng. Comput. 59 (4) (2021) 825–839, Springer.
[41] N.N. Das, N. Kumar, M. Kaur, V. Kumar, D. Singh, Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays, IRBM (2020), Elsevier.
[42] M. Heidari, S. Mirniaharikandehei, A.Z. Khuzani, G. Danala, Y. Qiu, B. Zheng, Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms, Int. J. Med. Inform. 144 (2020) 104284, Elsevier.

[43] v7labs, Covid-19 x-ray dataset, https://github.com/v7labs/covid-19-xray-dataset.
[44] K. Sahinbas, F.O. Catak, Transfer learning-based convolutional neural network for COVID-19 detection with x-ray images, in: Data Science for COVID-19, Elsevier, 2021, pp. 451–466.
[45] Y. Oh, S. Park, J.C. Ye, Deep learning COVID-19 features on CXR using limited training data sets, IEEE Trans. Med. Imaging 39 (8) (2020) 2688–2700.
[46] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015, p. 14.
[47] I. Ovcharenko, Lung segmentation, https://github.com/IlliaOvcharenko/lung-segmentation.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[50] M. Kuzlu, U. Cali, V. Sharma, Ö. Güler, Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools, IEEE Access 8 (2020) 187814–187823.
[51] S. Sarp, M. Knzlu, U. Cali, O. Elma, O. Guler, An interpretable solar photovoltaic power generation forecasting approach using an explainable artificial intelligence tool, in: 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference, ISGT, IEEE, 2021, pp. 1–5.
[52] S. Sarp, M. Kuzlu, E. Wilson, U. Cali, O. Guler, A Highly Transparent and Explainable Artificial Intelligence Tool for Chronic Wound Classification: XAI-CWC, Preprints, 2021.
[53] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (XAI) program, AI Mag. 40 (2) (2019) 44–58.
[54] D. Gunning, E. Vorm, J.Y. Wang, M. Turek, DARPA's explainable AI (XAI) program: a retrospective, Appl. AI Lett. (2021).
[55] L. Zou, H.L. Goh, C.J.Y. Liew, J.L. Quah, G.T. Gu, J.J. Chew, M.P. Kumar, C.G.L. Ang, A. Ta, Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections, IEEE Trans. Artif. Intell. (2022).
[56] Y. Bar, I. Diamant, L. Wolf, H. Greenspan, Deep learning with non-medical training used for chest pathology identification, in: Medical Imaging 2015: Computer-Aided Diagnosis, vol. 9414, International Society for Optics and Photonics, 2015, 94140V.
[57] E. Hussain, M. Hasan, M.A. Rahman, I. Lee, T. Tamanna, M.Z. Parvez, Corodet: a deep learning based classification for COVID-19 detection using chest x-ray images, Chaos Solitons Fractals 142 (2021) 110495.
[58] F.O. Catak, K. Şahinbaş, Human-in-the-loop enhanced COVID-19 detection in transfer learning-based CNN models, in: Computational Intelligence for COVID-19 and Future Pandemics, Springer, 2022, pp. 71–87.
[59] Y. Xu, et al., A collaborative online AI engine for CT-based COVID-19 diagnosis, medRxiv 51 (3) (2020) 1–20.
[60] N. Qianqian, et al., A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images, Eur. Radiol. 30 (12) (2020) 6517–6527.
[61] D. Qingli, et al., Chest CT images for COVID-19: radiologists and computer-based detection, Front. Mol. Biosci. 8 (2021) 1–5.