SN Computer Science (2023) 4:141
https://doi.org/10.1007/s42979-022-01545-8

ORIGINAL RESEARCH

Quantitative and Qualitative Analysis of 18 Deep Convolutional Neural Network (CNN) Models with Transfer Learning to Diagnose COVID-19 on Chest X-Ray (CXR) Images

Li Sze Chow[1] · Goon Sheng Tang[1] · Mahmud Iwan Solihin[2] · Nadia Muhammad Gowdh[3] · Norlisah Ramli[3] · Kartini Rahmat[3]

Received: 22 March 2022 / Accepted: 3 December 2022 / Published online: 5 January 2023

Abstract

Coronavirus disease 2019 (COVID-19) is a disease caused by a novel strain of coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), severely affecting the lungs. Our study aims to combine both quantitative and qualitative analysis of the convolutional neural network (CNN) model to diagnose COVID-19 on chest X-ray (CXR) images. We investigated 18 state-of-the-art CNN models with transfer learning, which include AlexNet, DarkNet-19, DarkNet-53, DenseNet-201, GoogLeNet, Inception-ResNet-v2, Inception-v3, MobileNet-v2, NasNet-Large, NasNet-Mobile, ResNet-18, ResNet-50, ResNet-101, ShuffleNet, SqueezeNet, VGG-16, VGG-19, and Xception. Their performances were evaluated quantitatively using six assessment metrics: specificity, sensitivity, precision, negative predictive value (NPV), accuracy, and F1-score. The top four models with accuracy higher than 90% are VGG-16, ResNet-101, VGG-19, and SqueezeNet. The accuracy of these top four models is between 90.7% and 94.3%; the F1-score is between 90.8% and 94.3%. The VGG-16 scored the highest accuracy of 94.3% and F1-score of 94.3%. The majority voting with all the 18 CNN models and top 4 models produced an accuracy of 93.0% and 94.0%, respectively. The top four and bottom three models were chosen for the qualitative analysis. A gradient-weighted class activation mapping (Grad-CAM) was used to visualize the significant region of activation for the decision-making of image classification. Two certified radiologists performed blinded subjective voting on the Grad-CAM images in comparison with their diagnosis. The qualitative analysis showed that SqueezeNet is the closest model to the diagnosis of two certified radiologists. It demonstrated a competitively good accuracy of 90.7% and F1-score of 90.8% with 111 times fewer parameters and 7.7 times faster than VGG-16. Therefore, this study recommends both VGG-16 and SqueezeNet as additional tools for the diagnosis of COVID-19.

Keywords COVID-19 · Convolutional neural networks (CNN) · Transfer learning · Gradient-weighted class activation mapping (Grad-CAM)

Introduction

Coronavirus disease 2019 (COVID-19) is an illness caused by a novel coronavirus, which is now called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first outbreak was on December 21, 2019, in Wuhan City, China [1]. The World Health Organization (WHO) declared COVID-19 a global pandemic on March 11, 2020. It has escalated to 180 million cases with 3.9 million deaths and 165 million recovered as recorded on June 24, 2021 [2]. Among the worst-hit nations are the USA, India, and Brazil.

Effective screening is essential to triage the patients and treat them accordingly. COVID-19 is diagnosed by the real-time reverse transcription-polymerase chain reaction

✉ Li Sze Chow
chowls@ucsiuniversity.edu.my

1 Department of Electrical and Electronic Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, 1, Jalan Puncak Menara Gading, Taman Connaught, Cheras, 56000 Kuala Lumpur, Malaysia

2 Department of Mechanical and Mechatronics Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, 1, Jalan Puncak Menara Gading, Taman Connaught, Cheras, 56000 Kuala Lumpur, Malaysia

3 Department of Biomedical Imaging, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia

(RT-PCR) of nasopharyngeal swabs [3]. Chest radiography imaging and computed tomography (CT) are essential supplementary diagnostic tools for investigating patients suspected of having COVID-19. It is also a vital tool for patient follow-up. However, it requires an experienced and certified radiologist to triage the patients accurately. CXR findings are often non-specific and thus challenging to categorize due to COVID-19 or not. Therefore, a computer-aided diagnosis with automatic classification of lung abnormalities would be beneficial to assist radiologists to confirm their diagnosis and speed up the process.

Recently, many researchers used a convolutional neural network (CNN), a deep learning algorithm to assist in the diagnosis of COVID-19. Deep learning uses automatic feature extraction and pattern recognition to classify an image. CNN is based on the shared-weight architecture of the convolution kernels or filters, which slide along input features and produce the feature maps. CNN uses fully connected networks where each neuron in one layer is connected to all neurons in the next layer. In each layer, the data are transformed into a higher and more abstract level. The deeper the network, the more complex is the information learned. CNN is commonly used for image classification and segmentation.

Wang et al. proposed the COVID-Net model, which combined human-driven principled network design prototyping with machine-driven design exploration to detect COVID-19 cases from CXR images [4]. They used residual architecture design principles in the first stage of human-driven principled network design. Then they used generative synthesis to identify the optimal macro-architecture and micro-architecture designs for the COVID-Net model. They reported an accuracy of 92.6% on the test dataset, a sensitivity of 87.1% for COVID-19 cases, and a high positive predictive value (PPV) of 96.4% for COVID-19 cases. Mangal et al. used a pre-trained CheXNet [5], which consists of a 121-layer dense convolutional network (DenseNet) [6] backbone [7]. The final classifier was replaced with three classes (normal, pneumonia, and COVID-19) and four classes (normal, bacterial pneumonia, viral pneumonia, and COVID-19) of the classification layer. They showed an accuracy of 90.5% and 87.2% for three classes and four classes classification, respectively. Their study demonstrated a better performance than COVID-Net by > 0.1 area under the receiver operating characteristic (AUROC) curve in detecting pneumonia and COVID-19. Kumar et al. used deep features and a support vector machine (SVM) to detect COVID-19, pneumonia, and those with no infection [8]. They applied 13 pre-trained CNN models, which include AlexNet, VGG-16, VGG-19, GoogleNet, ResNet-18, ResNet-50, ResNet-101, Inception-v3, Inception-ResNet-v2, DenseNet-201, XceptionNet, MobileNet-v2, and ShuffleNet. Their study found that the best model for detecting COVID-19 is ResNet-50 with an accuracy of 95.33%, sensitivity of 95.33%, false-positive rate (FPR) of 2.33%, and F1-score of 95.34%. Chaudhary et al. proposed the Fourier–Bessel series expansion (FBSE)-based dyadic decomposition (FBD) method to diagnose COVID-19 [9]. The CXR images are decomposed into subband images, which are fed to the ResNet-50 pre-trained CNN model. Then, the deep features from each CNN are ensembled before being fed to the softmax classifier. They reported an overall accuracy of 98.6%. They further expanded their work by using Fourier–Bessel series expansion-based decomposition (FBSED) for the image decomposition on X-ray and CT images and applied it to five different pre-trained CNN models [10]. They demonstrated an accuracy of 100% on CXR images and 97.6% on CT images.

Recently, Loey et al. proposed a novel Bayesian optimization-based CNN model, which consists of two main components [11]. The first component deploys CNN to extract and learn deep features related to COVID-19. The second component used a Bayesian-based optimizer to tune the CNN hyperparameters according to an objective function. They achieved an accuracy of 96%. Gour et al. developed an uncertainty-aware convolutional neural network (UA-ConvNet), which utilizes the EfficientNet-B3 model and Monte Carlo (MC) dropout [12]. The MC dropout was applied for M forward passes to obtain the posterior predictive distribution, which was then used to get the mean prediction and model uncertainty. Their results demonstrated a G-mean of 98.02% and sensitivity of 98.15% for the multiclass classification task; and, a G-mean of 99.16% and sensitivity of 99.30% for the binary classification.

The majority of the above deep learning methods required large datasets of images (> 1000) for training and validation. In the case of limited datasets, several research groups used transfer learning on the state-of-the-art CNN models in detecting COVID-19 cases. Transfer learning exploits the capabilities of the pre-trained CNN to new data with a smaller population instead of training a CNN from scratch. The pre-trained model retains both its initial architecture and all the learned weights. Then, the CNN model can be used as a feature extractor via transfer learning [13]. Apostolopoulos et al. applied transfer learning on five CNN models including VGG-19, MobileNet-v2, Inception, Xception, and Inception-ResNet-v2 for automatic detection of COVID-19 from CXR images [14]. They found that MobileNet-v2 is the most effective model with an accuracy of 96.78%, sensitivity of 98.66%, and specificity of 96.46%. Minaee et al. applied transfer learning on four CNN models including ResNet-18, ResNet-50, SqueezeNet, and DenseNet-121 [15]. They reported a sensitivity of 98% for all four CNN models, whereas the SqueezeNet showed the highest specificity of 92.9%. Soares et al. used three CNN models including Xception, ResNet, and VGG-16 with transfer learning [16]. They showed that the VGG-16 model had the highest accuracy of 97.3%. Majeed et al. performed a more extensive

study with 12 CNN models and their proposed shallow CNN architecture called CNN-X for multiclass classification of COVID-19, bacteria, viral, and normal class [17]. The 12 CNN models are AlexNet, GoogleNet, VGG-16, VGG-19, ResNet-18, ResNet-50, ResNet-101, Inception-v3, Inception-ResNet-v2, SqueezeNet, DenseNet-201, and Xception. Their study showed that Inception-v3 had the highest sensitivity of 97.26%. VGG-19, DenseNet-201, and Xception equally demonstrated the highest specificity of 100%. Nayak et al. also applied similar comprehensive study on eight pre-trained CNN models, including VGG-16, Inception-v3, ResNet-34, MobileNet-v2, AlexNet, GoogleNet, ResNet-50, and SqueezeNet [18]. They have investigated several hyperparameters to improve performance. Their results showed that ResNet-34 achieved the highest performance with an accuracy of 98.33%, sensitivity of 100%, and specificity of 96.67%.

The summary of the studies that used various CNN models with transfer learning to detect COVID-19 cases is recorded in Table 1. Each of these studies reported a different best CNN model for the detection of COVID-19 cases. Therefore, it is inconclusive which CNN model with transfer learning is the best model for detecting COVID-19. Furthermore, none of these studies examines the subjective accuracy compared to the diagnosis of a certified radiologist. Therefore, there was no subjective qualitative analysis performed in the previous studies. We do not know whether the detection of COVID-19 is genuinely based on the essential characteristic feature of the COVID-19 abnormality in the lung.

Table 1 Summary of the studies using CNN models with transfer learning for the detection of COVID-19 cases

Authors	Number of CNN models	Dataset ratio	Accuracy	Sensitivity	Specificity
Apostolopoulos et al. [14]	5	Train: test (not reported)	(2-class on Dataset_1) VGG-19: 98.75% MobileNet-v2: 97.40% Inception: 86.13% Xception: 85.57% Inception-ResNet-v2: 84.38% (2-class on Dataset_2) MobileNet: 96.78%	VGG-19: 92.85% MobileNet-v2: 99.10% Inception: 12.94% Xception: 0.08% Inception-ResNet-v2: 0.01% (2-class on Dataset_2) MobileNet: 98.66%	VGG-19: 98.75% MobileNet-v2: 97.09% (2-class on Dataset_2) MobileNet: 96.46%
Minaee et al. [15]	4	Train: test 2084:3100	–	ResNet-18: 98% ResNet-50: 98% SqueezeNet: 98% DenseNet-121: 98%	ResNet-18: 90.7% ResNet-50: 89.6% SqueezeNet: 92.9% DenseNet-121: 75.1%
Soares et al. [16]	3	Train: valid: test 80:10:10	VGG-16: 97.3% Xception: 95.9% ResNet: 94.6%	–	–
Majeed et al. [17]	12 and CNN-X	Train: test 80:20	–	AlexNet: 90.41% GoogleNet: 93.15% VGG-16: 84.93% VGG-19: 0% ResNet-18: 95.89% ResNet-50: 95.89% ResNet-101: 91.78% Inception-v3: 97.26% Inception-ResNet-v2: 95.89% SqueezeNet: 93.15% DenseNet-201: 90.41% Xception: 93.15% CNN-X: 93.15%	AlexNet: 88.03% GoogleNet: 96.15% VGG-16: 97.86% VGG-19: 100% ResNet-18: 98.72% ResNet-50: 97.01% ResNet-101: 97.86% Inception-v3: 92.74% Inception-ResNet-v2: 99.57% SqueezeNet: 99.57% DenseNet-201: 100% Xception: 100% CNN-X: 97.86%
Nayak et al. [18]	8	Train: test 286:120	ResNet-34: 98.33% ResNet-50: 97.50% GoogleNet: 96.67% VGG-16: 95.83% AlexNet: 97.50% MobileNet-v2: 95.83% Inception-v3: 92.50% SqueezeNet: 96.67%	ResNet-34: 100% ResNet-50: 100% GoogleNet: 96.67% VGG-16: 96.67% AlexNet: 98.33% MobileNet-v2: 93.33% Inception-v3: 88.33% SqueezeNet: 95%	ResNet-34: 96.67% ResNet-50: 95% GoogleNet: 96.67% VGG-16: 95% AlexNet: 96.67% MobileNet-v2: 98.33% Inception-v3: 96.67% SqueezeNet: 98.33%

Therefore, our study aims to combine quantitative and qualitative analysis of the CNN models with transfer learning. This is an interdisciplinary study between computer scientists and radiologists. We investigated 18 CNN models with transfer learning for the detection of COVID-19 cases. It includes AlexNet, DarkNet-19, DarkNet-53, DenseNet-201, GoogLeNet, Inception-ResNet-v2, Inception-v3, MobileNet-v2, NasNet-Large, NasNet-Mobile, ResNet-18, ResNet-50, ResNet-101, ShuffleNet, SqueezeNet, VGG-16, VGG-19, and Xception. Their performances were evaluated using six assessment metrics including specificity, sensitivity, precision, negative-predictive value (NPV), accuracy, and F1-score. From these assessment metrics, the top four models with accuracy higher than 90% and the bottom three models were chosen for qualitative analysis. A gradient-weighted class activation mapping (Grad-CAM) was used to visualize the significant region of activation for the decision-making in image classification. Two certified radiologists performed blinded subjective voting on the Grad-CAM heatmaps in comparison with their diagnosis. This is a new contribution to the existing study, where we have adopted the qualitative assessment of the Grad-CAM heatmaps to enhance the evaluation of the CNN models. This study investigated the best CNN model with transfer learning to diagnose COVID-19 by combining both quantitative and qualitative analysis.

The main contributions of this work are:

- Quantitative analysis using six assessment metrics on 18 CNN models with transfer learning for diagnosing COVID-19 on CXR images. This is an objective assessment by computer.
- Identification of COVID-19 pneumonia-related lung changes on CXR identified visually by two certified radiologists on 50 CXR images. This is the ground truth of the diagnosis.
- Qualitative analysis of the top four and bottom three CNN models using Grad-CAM heatmaps, performed by two certified radiologists in comparison with the ground truth. This is a subjective assessment by radiologists.

Material and Methods

Overview of 18 CNN Architectures

VGG uses up to 19 weight layers, which is a very deep convolutional network during its era for large-scale image classification. They explored the conventional Convolutional Networks (ConvNets) and increased the depth of architecture with very small (3×3) convolution filters [19]. Our study used two versions of VGG, which are VGG-16 and VGG-19, where the number represents the number of layers. ResNet explicitly reformulates the layers as learning residual functions with reference to the layer inputs. Their baselines were inspired by the VGG nets except that this model has fewer filters and lower complexity. [20]. Our study used three versions of ResNet, which are ResNet-18, ResNet-50, and ResNet-101, where the number represents the number of layers. AlexNet comprises 5 convolution layers and 3 fully connected layers with a final 1000-way softmax layer. They used the "dropout" regularization method to reduce overfitting and non-saturating neurons to make training faster [21]. SqueezeNet is a small CNN architecture with equivalent accuracy to AlexNet although it is 50 times fewer parameters and 510 times smaller than AlexNet. It replaced the 3×3 filters with 1×1 filters, decreased the number of input channels to 3×3 filters, and downsample late in the network so that the convolution layers have a large activation map [22].

Inception-v3 scales up the networks by factorizing convolutions and aggressive dimension reductions inside the neural network. They demonstrated the training of high-quality networks on relatively modest size training sets using the combination of lower parameter count and additional regularization with batch-normalized auxiliary classifiers and label smoothing. They showed high-quality results for low receptive field resolution of 79×79, which could help detect relatively small objects [23]. GoogLeNet applies the Inception network, and its architecture is based on the Hebbian principle and the intuition of multi-scale processing. The main benefit is that it allows the increase of the depth and width of the network without a huge computational complexity [24]. Inception-ResNet-v2 combined the ideas of residual connections and the Inception architecture. It shows the benefit of accelerating the Inception networks' training speed and improving the recognition performance significantly [25]. Xception architecture was inspired by the Inception module, but it is entirely based on depth-wise separable convolutions with linear residual connections. It uses the same number of parameters as Inception-v3 but in a more efficient use of these parameters [26].

DarkNet-19 uses 3×3 filters and doubles the number of channels after every pooling step. It uses global average pooling to make predictions and a 1×1 filter to compress the feature representation between 3×3 convolutions [27]. DarkNet-53 is a variant of DarkNet-19, where it has 53 convolutional layers [28]. DenseNet-201 uses a feed-forward to link each layer to every other layer. In each layer, the feature maps of all the preceding layers are used as inputs. Its feature maps are then used as inputs into all the following layers. It solves the vanishing gradient problem, improves feature propagation, encourages feature reuse, and reduces the number of parameters [6].

MobileNet-v2 is a mobile architecture based on an inverted residual structure and linear bottleneck. The shortcut connections are between the thin bottleneck layers. The

intermediate expansion layer used lightweight depth-wise convolutions to filter the features. Its architecture consists of an initial fully convolution layer with 32 filters and 19 residual bottleneck layers [29]. ShuffleNet utilizes pointwise group convolution and channel shuffle. It reduces the computation cost while maintaining accuracy. Its computation is 13 times faster than AlexNet for comparable classification accuracy. It was designed for mobile devices [30]. NasNet designs a new search space to search for an architectural building block on a small dataset and then transfer the block to a larger dataset. They used the neural architecture search (NAS) as the primary search method. The model used a new regularization technique called "Scheduled Drop Path" that improves generalization [31]. Our study used two versions of NasNet, which are NasNet-Large and NasNet-Mobile.

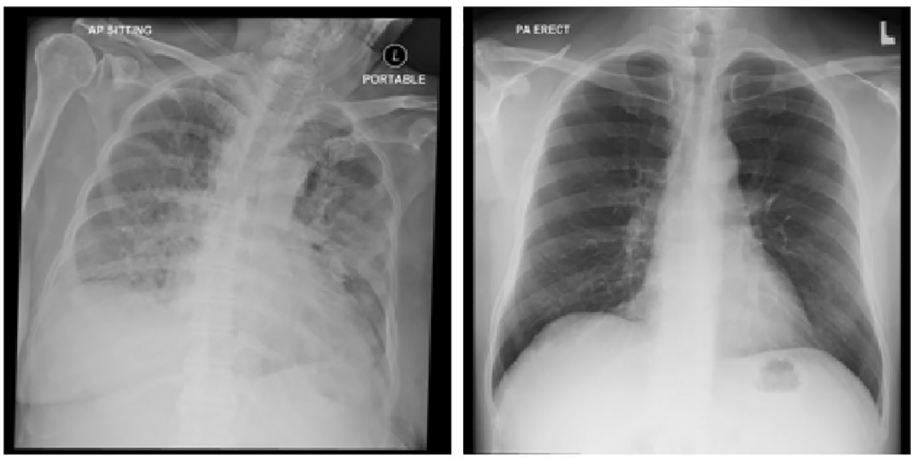
Dataset Preparation

The CXR images in our study were obtained from the public and private domains. The dataset from the public domain is called COVIDx [32], which consists of CXR images from five sources: Actualmed COVID-19 Chest X-Ray Dataset Initiative (Actmed) [33], COVID-19 Image Data Collection: Prospective Predictions Are the Future (COHEN) [34], Fig. 1 COVID-19 Chest X-Ray Dataset Initiative (Fig1) [35], (COVID-19 Radiography Database (SIRM) [36], and RSNA Pneumonia Detection Challenge (RSNA) [37]. The dataset from the public domain are available in the websites listed in the references. The dataset from the private domain was provided by the Department of Biomedical Imaging, Faculty of Medicine, University of Malaya (UM), Malaysia. The dataset from the private domain is not available to the public following the ethnic agreement which is specified for this study only. We obtained both CXR images of normal and COVID-19 subjects from both public and private domains. We chose the CXR images in the posteroanterior (PA) and anteroposterior (AP) views of the lung for this study. The number of images from each domain and source is recorded in Table 2. The size of the normal images range from 1024×1024 (smallest) to 2520×3032 (largest); the COVID-19 images range from 220×206 (smallest) to 4280×3520 (largest). There are no specific gray levels in the public domain images since they were taken from various databases. The private domain DICOM images were 12-pixel depth indicating 4096 gray levels in each CXR image. Figure 1 shows a COVID-19 CXR image and a normal lung CXR image provided by UM.

Table 2 The number of CXR images obtained from the public and private domain

Class	Dataset		
	Public (COVIDx)	Private (UM)	Total
Normal	RSNA (200)	UM (150)	350
COVID-19	Actmed (58) COHEN (65) Fig1 (35) SIRM (42) (Total 200)	UM (150)	350

The 18 CNN models were trained with a combined dataset consisting of 200 normal CXR images (100 from COVIDx and 100 from UM) and 200 COVID-19 CXR images (100 from COVIDx and 100 from UM). These images were split to a ratio of 7:3 for training and validation. For each class (normal and COVID-19), 140 images were used for training, and 60 images were used for validation. The remaining images were used for testing the CNN models to evaluate their performances. The testing dataset consists of 150 normal CXR images (100 from COVIDx



(a) (b)

Fig. 1 CXR images for **a** a patient diagnosed with COVID-19 and **b** a normal lung

Table 3 The implementation details of the dataset split for training, validation, and testing

Class	Training	Validation	Testing	Total
Normal	140	60	150	350
COVID-19	140	60	150	350

and 50 from UM) and 150 COVID-19 CXR images (100 from COVIDx and 50 from UM). The dataset split for training, validation, and testing is recorded in Table 3.

Hardware and Software

The training, validation, and testing of the CNN models were performed using an Intel(R) Core (TM) i5-10,500 CPU @ 3.10 GHz with 8 GB RAM. The YAKAMI DICOM Tool [38] was used to convert the DICOM images to JPEG file format. Then, the Deep Network Designer Toolbox in MATLAB R2020b (The Mathworks, Inc.) was used for training and testing the 18 CNN models. The MATLAB Grad-CAM Library [39] was used to run the Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the classification decision.

Transfer Learning

Our study applied transfer learning to the 18 CNN models available in MATLAB's Deep Network Designer. The 18 CNN models were previously trained using the ImageNet images [40]. Since we do not have a large dataset of CXR images to train a deep learning model from scratch, transfer learning was applied to the pre-trained CNN models. In this approach, the CNN models are used as a feature extractor while keeping their initial architecture. Referring to Fig. 2, the lower layers for the feature extractor portion are frozen. The original fully connected, softmax and classification output layers are removed and replaced with a new set with an output size of 2 to indicate the binary classification of COVID-19 or normal classes. We did not attempt to optimize the CNN models or adjust their weights in the feature learning portions. The transfer learning approach is a more efficient and common way for the considerably small size of data; therefore, we do not need to train the CNN models from scratch.

This study used the recommended default hyperparameter settings provided by MathWorks' Deep Learning Guide. Figure 3 is the training setting used for all the CNN models in this study. No tuning approach was done since it is not the main focus of this study.

Assessment Metric

There are four possible outcomes in a confusion matrix for binary classification: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). True positive (TP) refers to the number of cases correctly classified as positive where the disease is present. True negative (TN) refers to the number of cases correctly classified as negative where the disease is absent. False negative (FN) refers to the number of cases wrongly classified as negative where the disease is present. False positive (FP) refers to the number

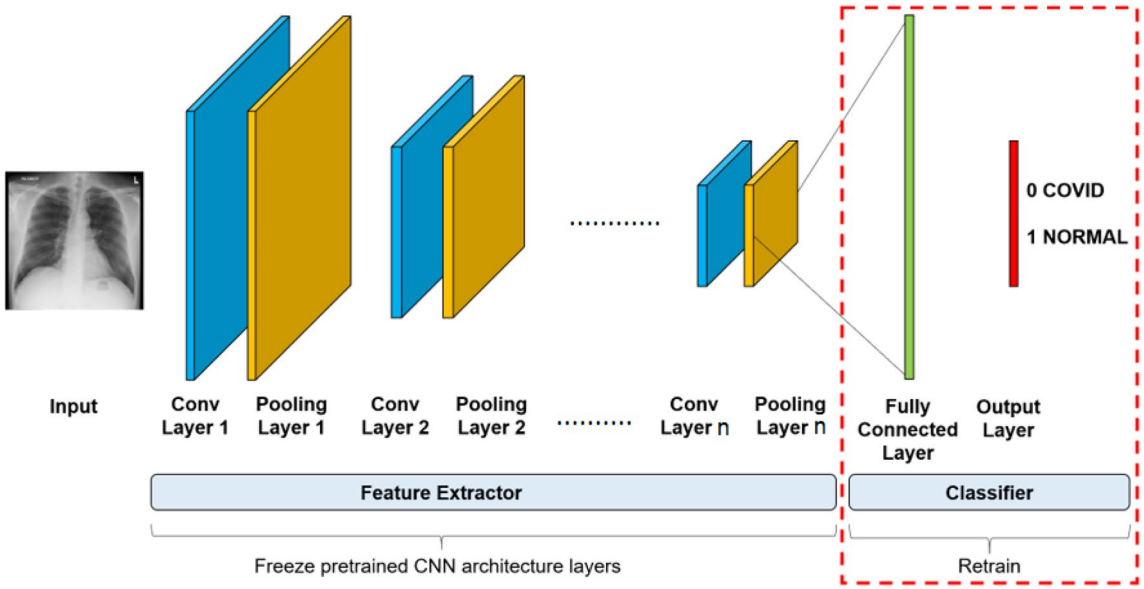


Fig. 2 A pre-trained CNN architecture is adapted with transfer learning to perform a binary classification (COVID-19 or normal)

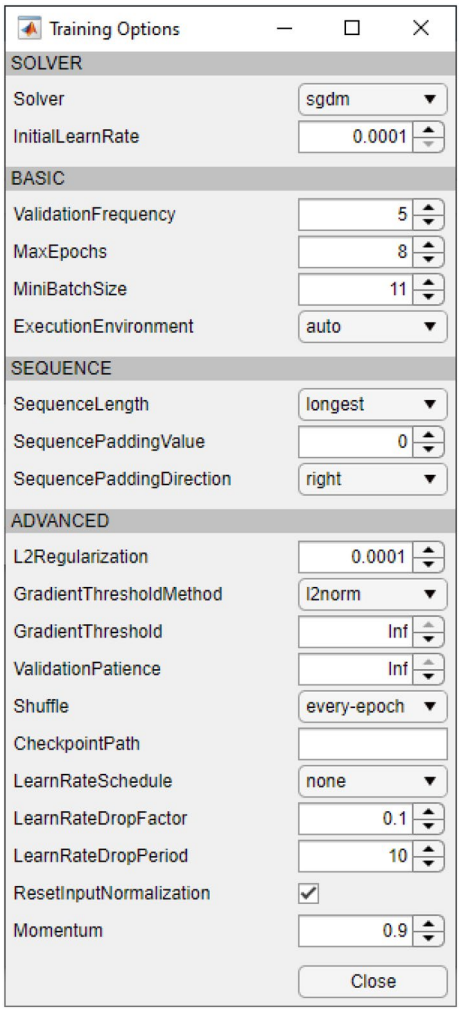


Fig. 3 Training setting of the hyperparameters for all the CNN models

of cases wrongly classified as positive where the disease is absent.

The TP, TN, FN, and FP are used to calculate the assessment metrics including specificity, sensitivity, precision, NPV, accuracy, and F1-score. These metrics are used to evaluate the performance of the 18 CNN models in this study. The formulas for the specificity, sensitivity (or recall), precision, NPV, accuracy, and F1-score are given in Eq. (1) to Eq. (6), respectively:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \tag{1}$$

$$\text{Sensitivity/Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{3}$$

$$\text{Negative Predictive Value (NPV)} = \frac{\text{TN}}{\text{TN}+\text{FN}}, \tag{4}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{5}$$

$$\text{F1 score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\times\text{TP}}{(2\times\text{TP})+\text{FN}+\text{FP}}. \tag{6}$$

Majority Voting

Majority voting has been adopted with deep learning to improve the COVID-19 detection accuracy [41, 42]. Our study used the hard approach of majority voting, which gives a label of the class for each image according to the highest number of labels (votes) among all the CNN models. It is applied for the 18 CNN models, then repeated for the top 4 CNN models with an accuracy higher than 90%.

Qualitative Analysis with Grad-CAM

The prediction made by the CNN models can be evaluated quantitatively using the assessment metrics described earlier. However, we do not know which part of the images was used as the features in the decision-making for the prediction. Therefore, it is equally important to display some sort of "visual explanation" for the decision made by the CNN models. We used Grad-CAM for this purpose [39]. It uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the significant regions in the image for the prediction. It is a useful tool to interpret the model's decision. For the 18 CNN models, the feature map layer was specified for each model to produce the Grad-CAM heatmap as shown in Table 4.

From the quantitative analysis, we chose the top four and bottom three CNN models for further qualitative analysis. We produced the Grad-CAM heatmaps of the testing dataset with COVID-19 (50 CXR images from UM). Two certified radiologists with more than 5 and 10 years of CXR interpretation experience independently evaluated these CXR images by drawing a contour over the infected region within the lung using the ITK-SNAP software [43]. For each CXR image, the radiologists were given seven Grad-CAM heatmaps (top four and bottom three) to vote for the closest heatmap with their diagnosis indicated by the contour of the infected region. If there were more than one heatmap with the correct region identified by the CNN model, they were all given one vote. If all the heatmaps showed the wrong region, no vote was given for that image. This process was

Table 4 The selected feature map layer used to produce Grad-CAM heatmap in each CNN models

CNN models	Feature map layer
AlexNet	relu5
DarkNet-19	conv19
DarkNet-53	res23
DenseNet-201	conv5_block32_concat
GoogLeNet	inception_5b-output
Inception-ResNet-v2	conv_7b_ac
Inception-v3	mixed10
MobileNet-v2	out_relu
NasNet-Large	activation_520
NasNet-Mobile	activation_188
ResNet-101	res5c_relu
ResNet-18	res5b_relu
ResNet-50	activation_49_relu
ShuffleNet	node_199
SqueezeNet	relu_conv10
VGG-16	relu5_3
VGG-19	relu5_4
Xception	block14_sepconv2_act

repeated for 50 CXR images with seven heatmaps each. The bottom three CNN models were included in this vote to ensure that they were the least accurate model compared to the top four models. The radiologists performed blind analysis during the voting without knowing the name of the CNN models. We aim to find the most suitable CNN models for COVID-19 detection by combining both quantitative and qualitative analysis.

Results

Quantitative Analysis

Table 5 records the depth of layers, total layers (convolution, dense, pooling, etc.), and the number of parameters (in million) for the 18 CNN models, arranged from the highest to the lowest number of parameters. All the models used the same input image size of $224 \times 224 \times 3$. Figure 4 shows the training time (left bars), and the validation and testing accuracy (right bars) for each model arranged from the highest to the lowest number of parameters. Generally, a model with a larger number of parameters requires a longer training time.

SqueezeNet used the lowest number of parameters (1.24 million) and the shortest training time (514 s = 8 min 34 s), yet a relatively good validation accuracy of 92.5% and testing accuracy of 90.67%. VGG-16 has the highest validation accuracy of 96.67% and testing accuracy of 94.33%, but a relatively long training time (3942 s = 1 h 5 min 42 s).

Table 5 The depth of layers, total layers, number of parameters, and image input size for the 18 CNN models are arranged from the highest to the lowest number of parameters

CNN model	Depth of layer	Total layers	Parameters (million)
VGG-19	19	47	144
VGG-16	16	41	138
NasNet-Large	*	1243	88.9
AlexNet	8	25	61
Inception-ResNet-v2	164	824	55.9
ResNet-101	101	347	44.6
DarkNet-53	53	184	41.6
ResNet-50	50	177	25.6
Inception-v3	48	315	23.9
Xception	71	170	22.9
DarkNet-19	19	64	20.8
DenseNet201	201	708	20
ResNet-18	18	71	11.7
GoogLeNet	22	144	7
NasNet-Mobile	*	913	5.3
MobileNet-v2	53	154	3.5
ShuffleNet	50	172	1.4
SqueezeNet	18	68	1.24

*The NASNet-Mobile and NASNet-Large networks do not consist of a linear sequence of modules

In general, there is a tradeoff in achieving higher accuracy. Nevertheless, NasNet-Large used the longest training time yet it has the lowest validation and testing accuracy. Therefore, the relation between the training time with the validation and testing accuracy is inconclusive among these 18 CNN models.

Table 6 records the classification results (TP, FP, FN, and TN) for the 18 CNN models, arranged from the highest to the lowest number of parameters, and for the majority voting with 18 models and the top 4 models. These values were used to calculate the assessment metric specificity, sensitivity, precision, NPV, accuracy, and F1-score as recorded in Table 7. The 18 CNN models were arranged from the highest to the lowest accuracy (%) in Table 7. It was found that VGG-16 has the highest accuracy of 94.3%, highest specificity of 93.5%, highest precision of 93.3%, and highest F1-score of 94.3%. VGG-19 demonstrated the highest sensitivity value of 95.6% and the highest NPV value of 96.0%. DarkNet-19 and GoogLeNet also demonstrated the highest NPV value of 96.0%. The top 4 models were identified based on an accuracy higher than 90%, which are VGG-16, ResNet-101, VGG-19, and SqueezeNet. The majority voting with 18 models produced an accuracy of 93.0%, which is lower than the majority voting with the top 4 models with an accuracy of 94.0%.

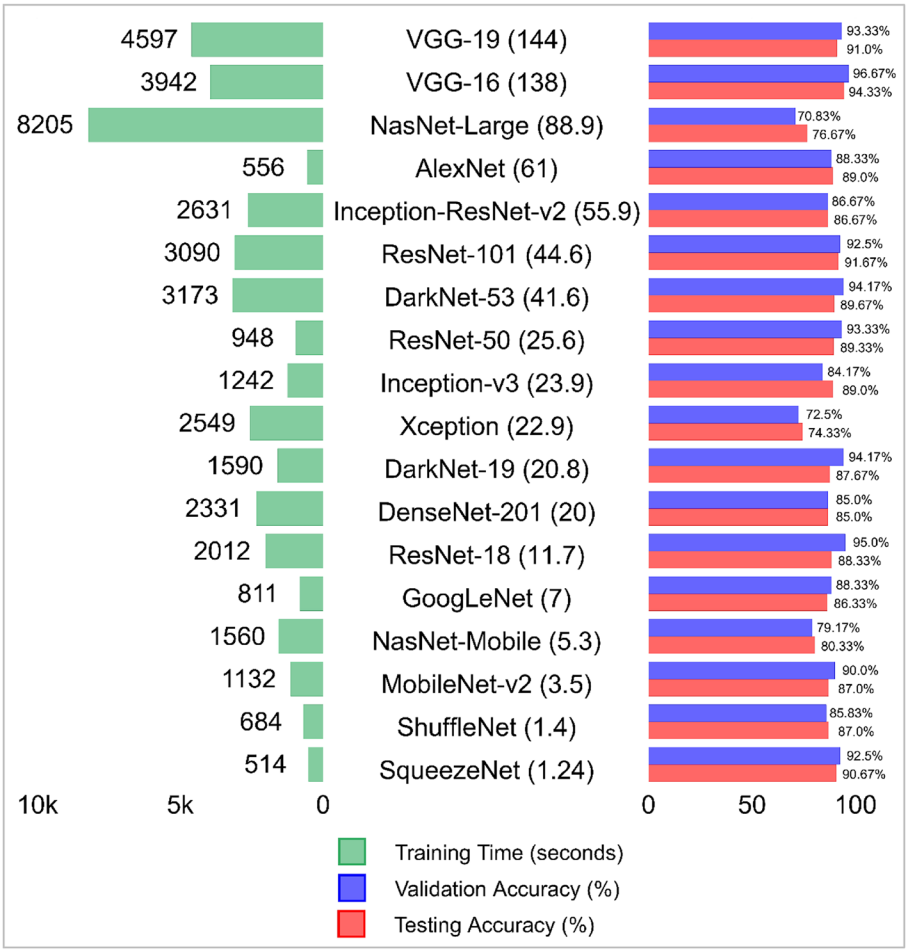


Fig. 4 Training time (left bars), validation, and testing accuracy (right bars) for 18 CNN models, with descending order of the number of parameters (written in bracket, in the unit of million)

The assessment metric results in Table 7 are plotted in Fig. 5 from the highest to the lowest number of parameters as the plot moves from the left to the right side. DarkNet-53 demonstrated the most consistent values among the six assessment metrics, while Xception has the largest variation of values. It is observed that there is no specified trend of performance with the number of parameters used in each CNN model. Our study focuses on the performance of different types of CNN models, instead of the number of parameters, to diagnose COVID-19. The majority voting with either 18 or the top 4 models produced consistently higher values of all the assessment metrics. The confusion matrices of the top 4 models, the majority voting with 18 models, and the majority voting with the top 4 models are plotted in Fig. 6.

Qualitative Analysis

From the above quantitative results in Table 7 and Fig. 5, it is inconclusive which CNN model is the best model for identifying COVID-19 from the normal lung CXR images. Therefore, it is necessary to perform qualitative analysis to investigate the most suitable CNN model for diagnosing COVID-19. Figure 7(a) and b show the Grad-CAM heatmaps of the 18 CNN models for the correctly classified COVID-19 and normal CXR images, respectively. The red region is the most significant region where the CNN models extracted the "features" during the prediction process. The blue region is the least significant region for decision-making. It is observed that some of the red regions for decision-making are not within the thoracic cavity. Therefore, the prediction performed by some CNN models was based on the features of a wrong region although it produced the true positive (TP) or true negative (TN) results. The ground truth of the infected lung area is shown in the bottom right corner of Fig. 7(a) by two radiologists. The majority voting method does not have a Grad-CAM heatmap because it is a different approach that produces the label of the images based on the majority votes of the prediction from each CNN model.

To identify which CNN model interpreted the correct region within the lung during the classification process, the qualitative analysis of these heatmaps is necessary with the assistance of the radiologist. Only the top four models (VGG-16, ResNet-101, VGG-19, and SqueezeNet) and bottom three models (NasNet-Mobile, NasNet-Large, and

Table 6 Classification results (TP, FP, FN, TN) for the 18 CNN models, arranged in the descending order of the number of parameters; and for the majority voting with 18 models and the top 4 models

CNN model	COVID (TP)	COVID (FP)	Normal (FN)	Normal (TN)
VGG-19	129	21	6	144
VGG-16	140	10	7	143
NasNet-Large	101	49	21	129
AlexNet	140	10	23	127
Inception-ResNet-v2	119	31	9	141
ResNet-101	137	13	12	138
DarkNet-53	134	16	15	135
ResNet-50	139	11	21	129
Inception-v3	127	23	10	140
Xception	90	60	17	133
DarkNet-19	119	31	6	144
DenseNet-201	118	32	13	137
ResNet-18	138	12	23	127
GoogLeNet	115	35	6	144
NasNet-Mobile	115	35	24	126
MobileNet-v2	136	14	25	125
ShuffleNet	138	12	27	123
SqueezeNet	139	11	17	133
Majority voting (18 models)	138	12	9	141
Majority voting (top 4 models)	142	8	10	140

Table 7 Assessment metric values for the 18 CNN models (arranged from the highest to the lowest accuracy) and for the majority voting with 18 and the top 4 models

CNN model	Specificity (%)	Sensitivity (%)	Precision (%)	NPV (%)	F1-Score (%)	Accuracy (%)
Majority voting (top 4 models)	94.6	93.4	94.7	93.3	94.0	94.0
Majority voting (18 models)	92.2	93.9	92.0	94.0	92.9	93.0
VGG-16	**93.5**	95.2	**93.3**	95.3	**94.3**	**94.3**
ResNet-101	91.4	91.9	91.3	92.0	91.6	91.7
VGG-19	87.3	**95.6**	86.0	**96.0**	90.5	91.0
SqueezeNet	92.4	89.1	92.7	88.7	90.8	90.7
DarkNet-53	89.4	89.9	89.3	90.0	89.6	89.7
ResNet50	92.1	86.9	92.7	86.0	89.7	89.3
AlexNet	92.7	85.9	**93.3**	84.7	89.5	89.0
Inception-v3	85.9	92.7	84.7	93.3	88.5	89.0
ResNet-18	91.4	85.7	92.0	84.7	88.7	88.3
DarkNet-19	82.3	95.2	79.3	**96.0**	86.5	87.7
MobileNet-v2	89.9	84.5	90.7	83.3	87.5	87.0
ShuffleNet	91.1	83.6	92.0	82.0	87.6	87.0
Inception-ResNet-v2	82.0	93.0	79.3	94.0	85.6	86.7
GoogLeNet	80.4	95.0	76.7	**96.0**	84.9	86.3
DenseNet-201	78.7	90.1	81.1	91.3	85.0	84.0
NasNet-Mobile	78.3	82.7	76.7	84.0	79.6	80.3
NasNet-Large	72.5	82.8	67.3	86.0	74.3	76.7
Xception	68.9	84.1	60.0	88.7	70.0	74.3

The highest value for each type of assessment metric is highlighted in bold font

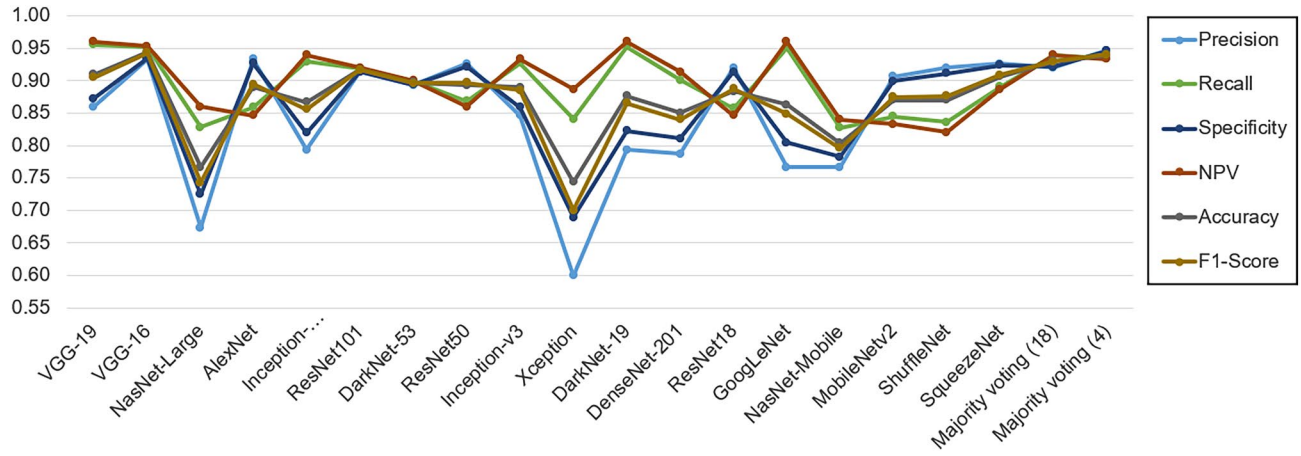


Fig. 5 Assessment metrics for the 18 CNN models, arranged from the highest to the lowest number of parameters as the plot moves from the left to right, and for the majority voting with 18 models and the top 4 models

Xception) from Table 7 were chosen to produce the Grad-CAM heatmaps for 50 CXR images (from UM datasets) for the qualitative analysis. Figure 8 shows another COVID-19 CXR image with the ground truth drawn by two radiologists and seven Grad-CAM heatmaps of the top four and bottom three models. The radiologists voted the best heatmap by comparing them with the contour of the infected region drawn by themselves. The result of their voting is recorded in Table 8. The total number of voting is unequal between the two radiologists because in any case without a correct heatmap, no score was given. Referring to Table 8, SqueezeNet has the highest score (printed in bold) on its Grad-CAM heatmaps to the radiologist's diagnosis. The bottom three models have the least score among both radiologists. This result confirms that the poorly performed CNN models in terms of quantitative analysis agreed with the qualitative analysis by the radiologists.

Discussion

This study has demonstrated both quantitative and qualitative analysis of 18 CNN models with transfer learning to diagnose COVID-19 on CXR images. The state-of-the-art CNN models can classify COVID-19 from normal lung CXR images with accuracy between 74.3% and 94.3% in our study as recorded in Table 7. Six assessment metrics were calculated including specificity, sensitivity, precision, NPV, accuracy, and F1-score. Yet, it is difficult to conclude which is the most suitable model from the quantitative analysis result. Most of the CNN models produced competitively good results of assessment metric values. Referring to Table 7, the top four CNN models with accuracy higher than 90% are VGG-16, ResNet101, VGG-19, and SqueezeNet. The majority voting with the hard approach produced aN accuracy of 94.0% when combining the top 4 models and 93.0% when combining all the 18 models. The slightly lower accuracy in combining 18 models is due to the averaging effect from the poorer models.

To date, the majority of the CNN studies for the detection of COVID-19 excluded qualitative analysis by radiologists. The new contribution from our study is the subjective qualitative analysis of the CNN models by certified radiologists alongside the quantitative analysis. Our study has combined both objective assessment (quantitative analysis by computer) and subjective assessment (qualitative analysis by radiologists) to enhance the evaluation of the CNN models. It gives us better confidence in our investigation of the best CNN model for diagnosing COVID-19 on CXR images.

Mangal et al. used RISE [44] to generate saliency maps to visualize their model's predictions [7]. The saliency map specifies parts of the input image that contribute to the activity of a specific layer or the decision of the neural network. It is a local gradient-based backpropagation interpretation method. However, a study showed that saliency maps are not totally reliable, where the data preprocessing such as input invariance and normalization could produce an undesirable effect on the saliency maps [45]. Another study found that the saliency map is vulnerable to adversarial attacks [46]. Therefore, our study only considers the Grad-CAM for the visualization of the model's predictions.

The Grad-CAM heatmap is a good visualization tool to illustrate the significant region used by each model for feature extraction and decision-making. Our study reveals that the Grad-CAM heatmaps of SqueezeNet are the closest to two independent radiologists' subjective diagnoses of COVID-19. SqueezeNet demonstrated an accuracy of 90.7% and an F1-score of 90.8%. Nevertheless, VGG-16 demonstrated the highest accuracy of 94.3% and F1-score of 94.3%

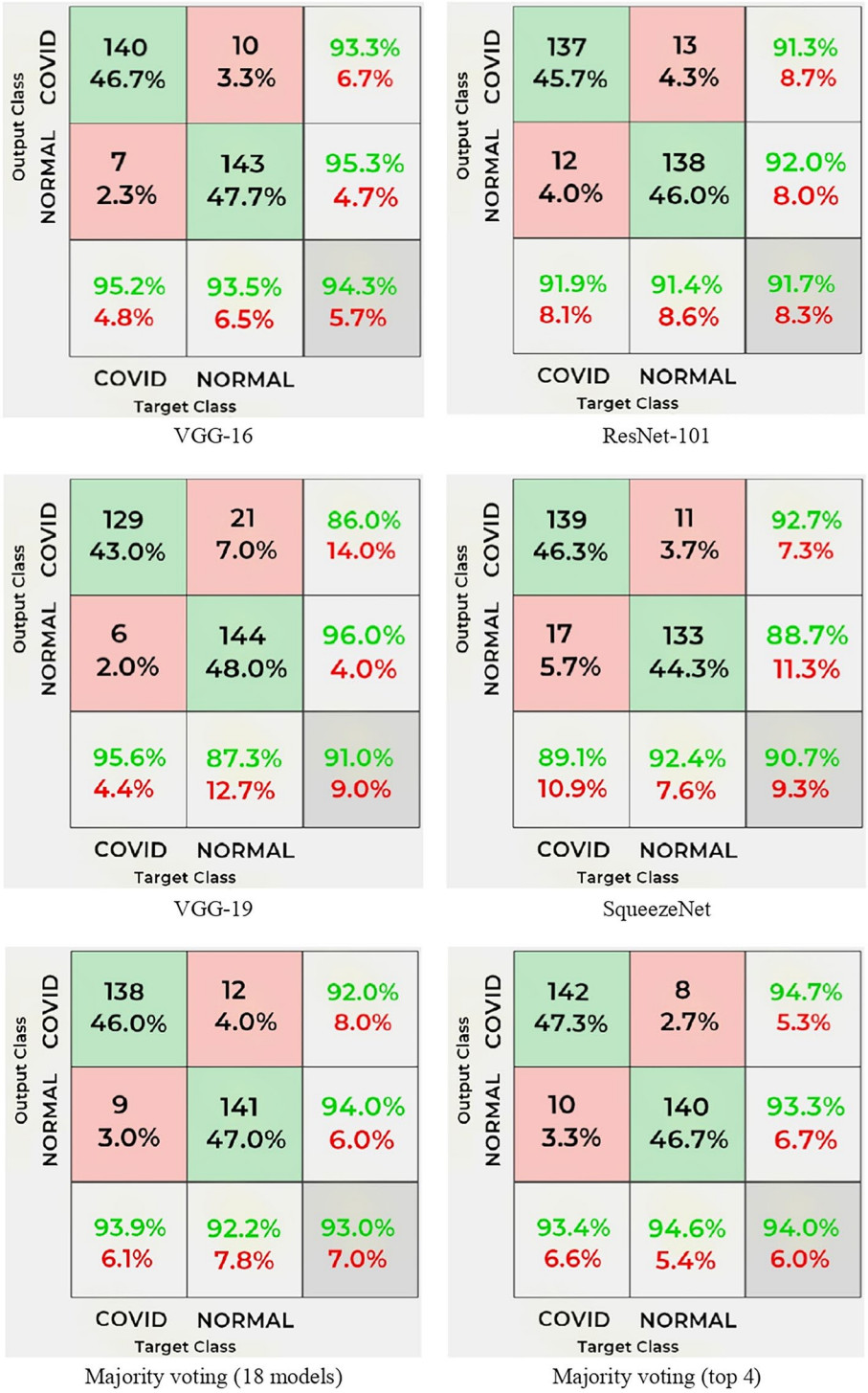


Fig. 6 Confusion matrices for the top 4 models (VGG-16, ResNet-101, VGG-19, SqueezeNet), majority voting with 18 models and the top 4 models

among all the CNN models. Although VGG-16 was not the highest voted model by radiologists, it does not rule out its credential for diagnosing COVID-19. Visual assessment is subject to human error and inter-observer discrepancies.

VGG is one of the most commonly used CNN models given its high accuracy in large-scale image recognition [19]. It uses very deep convolutional networks, a depth of 16 or 19 weight layers, which is beneficial for classification accuracy on a wide range of datasets and tasks. SqueezeNet maintains a competitive accuracy but with the least number of parameters (1.24 million) as shown in Table 5 and the shortest training time (8 min 34 s) as shown in Fig. 4. Its design strategies were to replace 3×3 filters with 1×1 filters, decrease the number of input

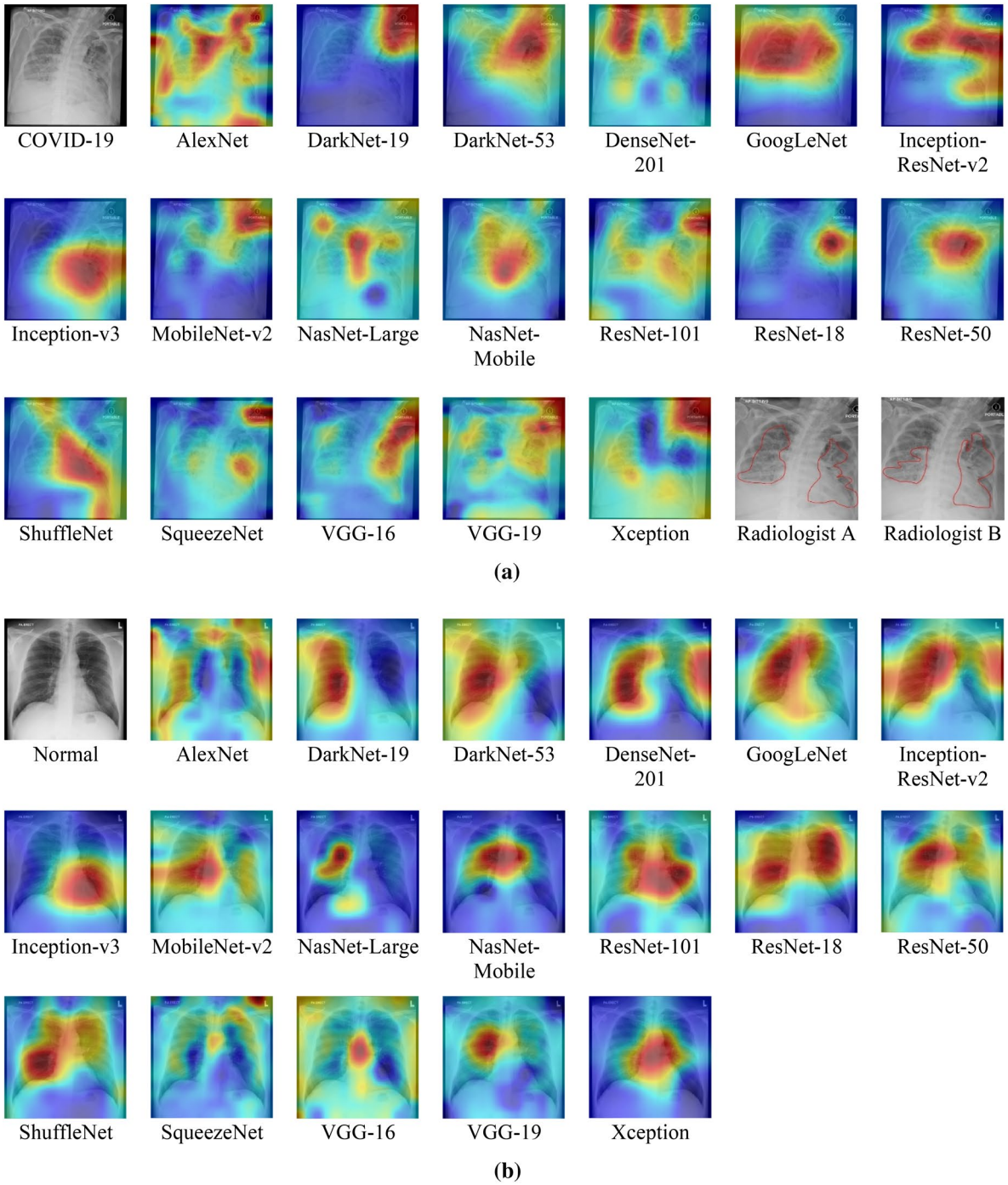


Fig. 7 The Grad-CAM heatmaps of 18 CNN models for a correctly classified **a** COVID-19 CXR (116_1.Ser2.Img1.jpg) where the ground truth identified by two radiologists are in grayscale with red contour indicating the affected area and **b** normal CXR (102.Ser1.Img1_anon.jpg)

channels, and downsample late in the network to achieve large activation in the convolution layers. VGG-16 is 1.04 times more accurate than SqueezeNet, but SqueezeNet used 111 times fewer parameters than VGG-16 and its training time was 7.7 times faster than VGG-16. The CNN model with fewer parameters has the advantage of more efficient distributed training and less overhead when exporting new models to clients [22]. Both VGG-16 and SqueezeNet could be recommended as supplementary tools to aid CXR interpretation to determine COVID-19 or normal findings.

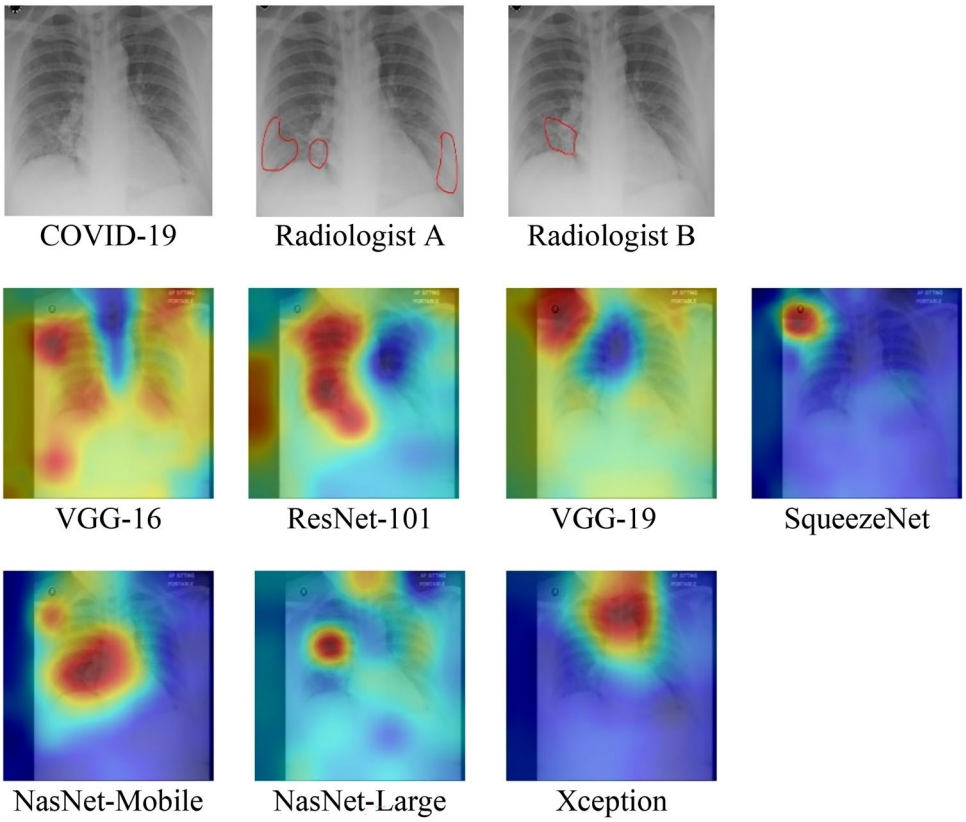


Fig. 8 COVID-19 CXR image (115_1.Ser1.Img1) with ground truth identified by two radiologists; Grad-Cam heatmaps of the top four and bottom three models

Table 8 The voting results by two blinded radiologists on 50 Grad-CAM heatmaps of top 4 and bottom 3 CNN models

CNN architecture	Radiologist A	Radiologist B
VGG-16	7	14
ResNet101	8	3
VGG-19	4	16
SqueezeNet	**11**	**20**
NasNet-Mobile	3	1
NasNet-Large	0	0
Xception	4	0

Although the CXR images could be pre-processed to exclude the artifacts of text, medical traces, or other anatomical regions, they would become tailor-made lung images in this investigation. In any natural image classification, a "complex" or "crowded" image should not be avoided during a fair evaluation process of a CNN model. Therefore, we did not selectively choose "perfect" lung images for this study. On the other hand, we adopted the qualitative assessment to enhance the evaluation of the CNN models in diagnosing COVID-19.

Our study used CXR as opposed to chest CT because CXR is the most used imaging tool for the assessment of COVID-19 patients in many parts of the world. It is used for baseline assessments, to assess disease severity and follow-up. It is done more commonly than CT as it can be done faster and easily at the patient's bedside. It uses low resources and does not incur long machine downtime or extensive room and machine cleaning as CT does. Therefore, CXR is widely used as a supplementary imaging tool to aid the diagnosis of COVID-19 pneumonia. RT-PCR is accurate and indeed the gold standard, however, RT-PCR results take a long time, often 24 h and up to 4 to 5 days during the peak of the pandemic. CXR is fast, cheap, and readily available thus CXR features that are suggestive of COVID-19 can be used to triage patients with a high index of suspicion of COVID-19. These patients can be isolated and treated as highly suspicious of COVID-19 pneumonia while awaiting the formal RT-PCR results. The RT-PCR test only gives a result as positive or negative of COVID-19 infection without any information on its severity. On the other hand, CXR provides information about the status of infection and the disease severity. Furthermore, CXR imaging is an efficient and cost-effective procedure with relatively cheap and portable equipment which could be performed rapidly in isolated rooms for COVID-19 patients. In addition, CXR images can be used as an input to the automatic diagnosis with the CNN model with high accuracy. A reliable CNN tool in the diagnosis of COVID-19 pneumonia with this inexpensive and easily available imaging tool will be highly beneficial.

There are a few shortcomings in this study yet to be improved in future studies. Our study used a considerably small dataset for the training, validation, and testing. However, it is possible given the advantage of the transfer learning on the CNN model. We have adopted diverse CXR images from the west (public domain, COVIDx) and east (private domain, UM). Future work could explore a larger number of datasets to verify the findings. The training and validation datasets could be expanded through data augmentation, i.e., rotations, translations, image scaling, reflections, shearing, and cropping transformation. It may overcome the overfitting of the training dataset encountered during the training process. Future work may explore training with a higher number of epochs to improve accuracy. The hyperparameters such as learning rates and different solvers can be experimented with to determine the best learning rate and solvers for the models to be trained on. The binary classification of either COVID-19 or normal lung is not directly applicable in actual clinical diagnosis. There could be other types of pneumonia infection that might be wrongly classified as COVID-19. Therefore, a multiclass classification model is necessary for a practical clinical diagnosis. Nevertheless, the automated diagnosis with the CNN model should only be used as a secondary tool to assess clinical suspicion of COVID-19. RT-PCR test remains the confirmatory test for this disease. Future work could develop a CNN model to monitor or classify stages of severity of lung deformation after the diagnosis of COVID-19.

Conclusion

The main contribution of this study is the combination of both objective quantitative and subjective qualitative analysis in evaluating the performance of CNN models with transfer learning to diagnose COVID-19. In this study, the quantitative analysis of 18 CNN models with transfer learning revealed that the top four models for diagnosing COVID-19 on CXR images are VGG-16, ResNet-101, VGG-19, and SqueezeNet. The VGG-16 scored the highest accuracy of 94.3% and the highest F1-score of 94.3%. The majority voting with all the 18 CNN models and top 4 models produced an accuracy of 93.0% and 94.0% respectively. The qualitative analysis using Grad-CAM heatmaps of the top four and bottom three models revealed that SqueezeNet is the closest model to the subjective diagnosis of two certified radiologists. SqueezeNet demonstrated a competitively good accuracy of 90.7% and F1-score of 90.8% with the shortest training time of 8 min 34 s. It used 111 times fewer parameters than VGG-16 and its training time was 7.7 times faster than VGG-16. Therefore, our study recommends both VGG-16 and SqueezeNet as additional tools for the diagnosis of COVID-19.

Funding This study was supported in part by the University Malaya Research Grant (Grant No: CSRG002-2020ST). We like to express our gratitude to the Department of Biomedical Imaging, the University of Malaya for providing the CXR images for this study.

Data availability The dataset from the public domain are available in the websites listed in the references. The dataset from the private domain is not available to the public following the ethnic agreement which is specified for this study only.

Declarations

Conflict of Interest The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Ethical Approval This research has been approved on ethical grounds by the Medical Research Ethics Committee, University Malaya Medical Centre Ethics Board on 19 June 2020 (2020417-8530).

References

1. Tan W, et al. A novel coronavirus genome identified in a cluster of pneumonia cases—Wuhan, China 2019–2020. China CDC Wkly. 2020;2(4):61–2. https://doi.org/10.46234/ccdcw2020.017.
2. "COVID Live Update: 167,653,596 Cases and 3,480,642 Deaths from the Coronavirus - Worldometer." https://www.worldometers.info/coronavirus/ (accessed 24 May 2021).
3. Wang W, et al. "Detection of SARS-CoV-2 in different types of clinical Specimens," JAMA - Journal of the American Medical Association, vol. 323, no. 18. American Medical Association, pp. 1843–1844, May 12, 2020, doi: https://doi.org/10.1001/jama.2020.3786.
4. Wang L, Wong A. "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images," 2020.
5. Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv. 2017. https://doi.org/10.4855/arXiv.1711.05225
6. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. "Densely connected convolutional networks," In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, vol. 2017-Janua, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.
7. Mangal KS, Arora. CovidAID: COVID-19 detection using chest X-ray. arXiv. https://doi.org/10.48550/arXiv.2004.09803 2020
8. Kumar Sethy P, Kumari Behera S, Kumar Ratha P, Biswas P. "Detection of coronavirus disease (COVID-19) based on Deep Features and Support Vector Machine," Preprints, Apr. 2020. Accessed: 24 May 2021. [Online]. Available: www.preprints.org.
9. Chaudhary PK, Pachori RB. "Automatic diagnosis of COVID-19 and pneumonia using FBD method," Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020, pp. 2257–2263, Dec. 2020. https://doi.org/10.1109/BIBM49941.2020.9313252.
10. Chaudhary PK, Pachori RB. FBSED based automatic diagnosis of COVID-19 using X-ray and CT images. Comput Biol Med. 2021. https://doi.org/10.1016/J.COMPBIOMED.2021.104454.

11. Loey E-SS, Mirjalili S. Bayesian-based optimized deep learning model to detect COVID-19 patients using chest X-ray image data. Comput Biol Med. 2022;142: 105213.
12. Gour M, Jain S. Uncertainty-aware convolutional neural network for COVID-19 X-ray images classification. Comput Biol Med. 2022;140: 105047.
13. Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data. 2016;3(1):1–40. https://doi.org/10.1186/s40537-016-0043-6.
14. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med. 2020;43(2):635–40. https://doi.org/10.1007/s13246-020-00865-4.
15. Minaee S, Kafieh R, Sonka M, Yazdani S, Jamalipour Soufi G. "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," Med. Image Anal., vol. 65, Oct. 2020, doi: https://doi.org/10.1016/j.media.2020.101794.
16. Soares LP, Soares CP. "Automatic detection of COVID-19 cases on X-ray images Using Convolutional Neural Networks," 2020. [Online]. Available: http://arxiv.org/abs/2007.05494.
17. Majeed T, Rashid R, Ali D, Asaad A. Issues associated with deploying CNN transfer learning to detect COVID-19 from chest X-rays. Phys Eng Sci Med. 2020;43(4):1289–303. https://doi.org/10.1007/s13246-020-00934-8.
18. Nayak SR, Nayak DR, Sinha U, Arora V, Pachori RB. Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: a comprehensive study. Biomed Signal Process Control. 2021;64: 102365. https://doi.org/10.1016/J.BSPC.2020.102365.
19. Simonyan K, Zisserman A. "Very deep convolutional networks for large-scale image recognition," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., vol. 75, no. 6, pp. 398–406, 2015.
20. He K, Zhang X, Ren S, Sun J. "Deep residual learning for image recognition," In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2016, vol. 2016-Decem, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
21. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. https://doi.org/10.1145/3065386.
22. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," pp. 1–13, 2016.
23. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. "Rethinking the Inception Architecture for Computer Vision," In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2016, vol. 2016-December, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
24. Szegedy C, et al. "Going deeper with convolutions," In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol. 07–12-June, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.
25. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. "Inception-v4, inception-ResNet and the impact of residual connections on learning," In: 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 2017, pp. 4278–4284.
26. Chollet F. "Xception: deep learning with depthwise separable convolutions," In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, vol. 2017-Janua, pp. 1800–1807. https://doi.org/10.1109/CVPR.2017.195.
27. Redmon J, Farhadi A. "YOLO9000: Better, Faster, Stronger," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 6517–6525, Dec. 2016, Accessed: 24 May 2021. [Online]. Available: http://arxiv.org/abs/1612.08242.
28. Redmon J, Farhadi A. "YOLOv3: An incremental improvement," *arXiv*. 2018.
29. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. "MobileNetV2: Inverted Residuals and Linear Bottlenecks," In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520, doi: https://doi.org/10.1109/CVPR.2018.00474.
30. Zhang X, Zhou X, Lin M, Sun J. "ShuffleNet: an extremely efficient convolutional neural network for mobile devices," In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856, doi: https://doi.org/10.1109/CVPR.2018.00716.
31. Zoph B, Vasudevan V, Shlens J, Le QV. "Learning transferable architectures for scalable image recognition," In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697–8710, doi: https://doi.org/10.1109/CVPR.2018.00907.
32. "COVID-Net/COVIDx.md at master · lindawangg/COVID-Net · GitHub." https://github.com/lindawangg/COVID-Net/blob/master/docs/COVIDx.md (accessed 02 Apr 2021).
33. "GitHub - agchung/Actualmed-COVID-chestxray-dataset: Actualmed COVID-19 Chest X-ray Dataset Initiative." https://github.com/agchung/Actualmed-COVID-chestxray-dataset (accessed 02 Apr 2021).
34. Cohen JP, Morrison P, Dao L, Roth K, Duong TQ, Ghassemi M. "COVID-19 image data collection: prospective predictions are the future," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.11988.
35. "GitHub - agchung/Figure1-COVID-chestxray-dataset: Figure 1 COVID-19 Chest X-ray Dataset Initiative." https://github.com/agchung/Figure1-COVID-chestxray-dataset (accessed 02 Apr 2021).
36. T. Rahman, M. Chowdhury, and A. Khandakar, "COVID-19 Radiography Database," *Kaggle*, 2020. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database/data# (accessed 29 Jul 2020).
37. "RSNA Pneumonia Detection Challenge | Kaggle." https://www.kaggle.com/c/rsna-pneumonia-detection-challenge (accessed 29 Jul 2020).
38. "YAKAMI DICOM Tools (Free DICOM Viewer/Converter/etc.)." https://www.kuhp.kyoto-u.ac.jp/~diag_rad/intro/tech/dicom_tools.html#INSTALL (accessed 02 Apr 2021).
39. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128(2):336–59. https://doi.org/10.1007/s11263-019-01228-7.
40. "ImageNet." https://www.image-net.org/ (accessed 24 May 2021).
41. Chandra TB, Verma K, Singh BK, Jain D, Netam SS. Coronavirus disease (COVID-19) detection in chest X-ray images using majority voting based classifier ensemble. Expert Syst Appl. 2021;165: 113909. https://doi.org/10.1016/j.eswa.2020.113909.
42. Jabra MB, Koubaa A, Benjdira B, Ammar A, Hamam H. COVID-19 diagnosis in chest X-rays using deep learning and majority voting. Appl Sci. 2021;11:2884. https://doi.org/10.3390/app11062884.
43. Yushkevich PA, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage. 2006;31(3):1116–28. https://doi.org/10.1016/j.neuroimage.2006.01.015.
44. Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models. Proceedings of the British Machine Vision Conference (BMVC). 2018
45. Kindermans P, Hooker S, Adebayo J. The unreliability of saliency methods. arXiv:1711.00867. 2017
46. Ghorbani A, Abid A, Zou J. Interpretation of neural networks in fragile. arXiv:1710.10547. 2018