# COVID-19/Pneumonia Classification Based on Guided Attention

**Viacheslav Danilov**
Quantori
**Alex Karpovsky**
Kanda Software
**Alexander Kirpich**
Georgia State University
**Diana Litmanovich**
Beth Israel Deaconess Medical Center
**Dato Nefaridze**
Quantori
**Oleg Talalov**
Quantori
**Semyon Semyonov**
Quantori
**Alexander Proutski**
Quantori
**Vladimir Koniukhovskii**
EPAM Systems
**Vladimir Shvartc**
EPAM Systems
**Yuriy Gankin** ( ✉ yuriy.gankin@quantori.com )
Quantori

**Research Article**

**Keywords:** Covid-19, pneumonia, classification, guided attention, deep learning

**Posted Date:** February 1st, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-149472/v1

# Abstract

With the novel coronavirus 19 (COVID-19) continually having a devastating effect around the globe, many scientists and clinicians are actively seeking to develop new techniques to assist with the tackling of this disease. Modern machine learning methods have shown promise in their adoption to assist the health care industry through their data and analytics-driven decision making, inspiring researchers to develop new angles to fight the virus. In this paper, we aim to develop a robust method for the detection of COVID-19 by utilizing patients' chest X-ray images. Despite recent progress, scarcity of data has thus far limited the development of a robust solution. We extend upon existing work by combing publicly available data across 5 different sources and carefully annotating the comprising images into three categories: normal, pneumonia, and COVID-19. To achieve a high classification accuracy, we propose a training pipeline based on the directed guidance of traditional classification networks, where the guidance is directed by an external segmentation network. Through this network, we observed that the widely used, standard networks can achieve an accuracy comparable to tailor-made models specifically for COVID-19, furthermore one network, VGG-16, outperformed the best of the tailor-made models.

# Research Highlights

1. Both direct and indirect supervision allow for networks to focus more on the desired object.
2. Basing network training on the guided attention mechanism results in accuracies comparable to tailor-made networks made for distinguishing between COVID-19 and pneumonia.
3. Direct supervision based on U-net influences network performance more than indirect supervision based on Grad-CAM.
4. VGG-16 trained using guided attention has demonstrated the most accurate classification at the level of 88% and 84% on the validation and testing subsets respectively.

# 1. Introduction

Since its introduction into the human population in late 2019, COVID-19 continues to have a devastating effect on the global populace with the number of infected individuals steadily rising [1]. With widely available treatments still outstanding and the continued strain placed on many healthcare systems across the world, efficient screening of suspected COVID-19 patients and their subsequent isolation is of paramount importance to mitigate the further spread of the virus. Presently, the accepted gold standard for patient screening is reverse transcriptase-polymerase chain reaction (RT-PCR) where the presence of COVID-19 is inferred from analysis of respiratory samples [2]. Despite its success, RT-PCR is a highly involved manual process with slow turnaround times, and with results becoming available up to several days after the test is performed. Furthermore, its variable sensitivity, lack of standardized reporting, and a widely ranging total positive rate [3–5] calls for alternative screening methods.

Chest radiography imaging (such as X-ray or computed tomography (CT) imaging) has gained traction as a powerful alternative, where the diagnosis is administered by expert radiologists who analyze the resulting images and infer the presence of COVID-19 through subtle visual cues [6–10]. Of the two imaging methods studied, X-ray imaging has distinct advantages with regards to accessibility, availability, and rate of testing [11]. Furthermore, the existence of portable X-ray imaging systems does not require patient transportation or physical contact between healthcare professionals and suspected infected individuals, thus allowing for efficient virus isolation and a safer testing methodology. Despite its obvious promise, the main challenge facing radiography examination is the scarcity of trained experts that could conduct the analysis at a time when the number of possible patients continues to rise. As such, a computer system that could accurately analyze and interpret chest X-ray images could significantly alleviate the burden placed on expert radiologists and further streamline patient care. Image identification techniques are readily adopted in Artificial Intelligence (AI) and could prove to be a powerful solution to the problem at hand.

Despite recent progress in the development of AI algorithms [12–15], one of the fundamental issues facing the development of a robust solution is the scarcity of publicly available data. We extend upon existing works by combining various publicly available data sources and carefully annotating the images across three classes: normal, pneumonia, and COVID-19. The data is then divided into training, validation, and testing subsets with an 8:1:1 split respectively with a strict class balance maintained across all sets.

Deep learning models, such as convolutional neural networks (CNNs), have gained traction in the field of medical imaging [16] and here we train 10 promising CNNs for the purpose of COVID-19 classification in chest X-ray images. To assist the models, we utilize a purpose-built extraction mask as part of a three-stage procedure. The mask accurately extracts the lung areas from the CXRs, with the subsequent images fed into one of the CNNs. To better quantify the performance of our proposed framework we benchmark our results against recently developed COVID-Net models [12]. To ensure consistency we utilize our dataset to output predictions across an array of different COVID-Net models.

The structure of the rest of this paper is as follows. Section 2 summarizes the data collected based on 5 most relevant datasets. Section 3 describes a proposed three-stage workflow using a guided attention mechanism. Results obtained during the all stages, Further improvements of the proposed workflow, its advantages over other models and possible implementation are shown in Sect. 4. Section 5 represents a synthesis of key points of the developed model based on the guided attention mechanism.

## 2. Data

To train a high-precision classifier, we collected data from different publicly available sources. At the time of publication, we identified the following five datasets; Covid Chest X-Ray Dataset (CCXRD) [17,18], Actualmed COVID-19 Chest X-Ray Dataset (ACCXRD) [19], Fig. 1 COVID-19 Chest X-Ray Dataset (FCCXRD) [20], COVID-19 Radiography Database (CRD) [21,22], and RSNA Pneumonia Detection Dataset (RSNAPDD) [23]. Since the datasets include different labels for their findings, we conducted the following mapping. We assigned viral and bacterial pneumonias to the “Pneumonia” label; SARS, MERS-CoV, COVID-19, and COVID-19 (ARDS) to the “COVID-19” label; “no findings” and “normal” diagnosis to the “Normal” label. Table 1 summarizes statistical information of the study data set.

Table 1
– Statistical information on the dataset used in the study

| **Dataset** | **Diagnosis** | | | **Images in dataset** |
|---|---|---|---|---|
| | **Normal** | **Pneumonia** | **COVID-19** | |
| CCXRD | 18 | 162 | 503 | 683 |
| ACCXRD | 127 | – | 58 | 185 |
| FCCXRD | 3 | 2 | 35 | 40 |
| CRD | – | – | 219 | 219 |
| RSNAPDD | 800 | 700 | – | 1500 |
| **Total** | **952 (34.1%)** | **918 (32.9%)** | **920 (33.0%)** | **2790 (100%)** |

It should be noted that the RSNAPDD dataset includes only normal and pneumonia cases. Originally, this dataset consisted of 20672 normal cases and 9555 cases of pneumonia. In order to keep the class balance in our datasets, we added only 800 normal and 700 pneumonia cases. It is worth noticing that normal and pneumonia cases from the CRD dataset were excluded because they duplicated images from the CCXRD dataset.

The final dataset only includes images acquired from the anterior-posterior (AP) and posteroanterior (PA) directions. Lateral CXR has no clinical applicability to distinguish COVID-19 patients [24].

# 3. Methods

The proposed workflow in this study is divided into three stages. During the first stage 10 industry-standard networks including MobileNet V2, DenseNet-121, EfficientNet B0, EfficientNet B1, EfficientNet B3, EfficientNet B5, VGG-16, ResNet-50 V2, Inception V3, and Inception ResNet V2 were trained on the prepared dataset. All those networks are the de-facto industry standard in the field of deep learning. During the second stage, we chose the 4 most accurate networks which were then fine-tuned. In the process of fine-tuning, both a feature extractor and a classifier were trained. In the third stage, the networks were trained with a guided attention mechanism. This mechanism is based on the usage of the U-net segmentation network, where the output is used to focus the classifier on the lung area of an image. Besides direct guidance by U-net, the network is additionally trained based on indirect supervision through the application of Grad-CAM. Indirect supervision is used in the training process since Grad-CAM’s attention heatmaps reflect the areas of an input image supporting the network’s prediction. In this regard, the prediction is based on the areas which we expect the network to focus on, while indirect supervision forces networks to focus on the desired object in the image rather than its other parts. The training workflow of the model is shown in Fig. 1 below. All three stages are described in Sect. 3.1 and 3.2 in more detail.

It should be noted that different COVID-Net models [12] are also considered in this study. To date, COVID-Net models are state-of-the-art models used for distinguishing COVID-19 and pneumonia cases. All COVID-Net models are abbreviated CXR further in the paper.

## 3.1. Stage I and Stage II

As we mentioned above, we chose 10 deep learning networks in order to find out which network architectures are most effective in recognizing COVID-19 and pneumonia. All the networks vary by the number of weights, architecture topology, the way of data processing, etc. Additionally, CXR models are used for comparison purposes. Table 2 summarizes information about the networks we used in the first stage.

Table 2

– Description of the models used during the first stage

| Model | Size of an input image | Size of an output feature vector | Parameters, millions | Size, Mb | Reference |
|---|---|---|---|---|---|
| MobileNet V2 | 224x224 | 7x7x1280 | 2.6 | 14 | 25 |
| DenseNet-121 | 224x224 | 7x7x1024 | 7.2 | 33 | 26 |
| EfficientNet B0 | 224x224 | 7x7x1280 | 4.2 | 29 | 27 |
| EfficientNet B1 | 240x240 | 8x8x1280 | 6.7 | 31 | 27 |
| EfficientNet B3 | 300x300 | 10x10x1536 | 11.0 | 48 | 27 |
| EfficientNet B5 | 456x456 | 15x15x2048 | 28.8 | 75 | 27 |
| VGG-16 | 224x224 | 7x7x512 | 14.9 | 528 | 28 |
| ResNet-50 V2 | 224x224 | 7x7x2048 | 25.6 | 98 | 29 |
| InceptionV3 | 299x299 | 8x8x2048 | 22.1 | 92 | 30 |
| Inception ResNet V2 | 299x299 | 5x5x1536 | 54.5 | 215 | 31 |
| CXR Small | 224x224 | 7x7x2048 | 117.4 | 1448 | 32 |
| CXR Large | 224x224 | 7x7x2048 | 127.4 | 1486 | 32 |
| CXR-3A | 480x480 | 13x13x1536 | 40.2 | 617 | 32 |
| CXR-3B | 480x480 | 15x15x2048 | 11.7 | 293 | 32 |
| CXR-3C | 480x480 | 15x15x2048 | 9.2 | 210 | 32 |
| CXR-4A | 480x480 | 13x13x1536 | 40.2 | 617 | 32 |
| CXR-4B | 480x480 | 15x15x2048 | 11.7 | 293 | 32 |
| CXR-4C | 480x480 | 15x15x2048 | 9.2 | 210 | 32 |

To train the abovementioned networks, we used bodies of these networks with ImageNet weights frozen. Using Amazon SageMaker [33], we tuned a given model and found its best version through a series of training jobs run on the collected dataset. Having performed hyperparameter tuning based on Bayesian optimization strategy, a set of hyperparameter values for the best performing model was found, as measured by a validation accuracy. The optimal architecture of the network head consists of the following layers:

- Global Average Pooling layer;
- Densely-connected layer with 128 neurons and ELU activation;
- Dropout layer with dropout rate equal to 0.10;
- Densely-connected layer with 64 neurons and ELU activation;
- Dropout layer with dropout rate equal to 0.05;
- Densely-connected layer with 3 neurons;
- Softmax activation layer.

It is important to note that for the first stage that only the classification heads were trained with body weights frozen. According to the results of the hyperparameter tuning procedure, gradient descent optimizer SGD with a learning rate equal to $10^{-4}$ proved to be optimal. Having trained several state-of-the-art networks, we found that most of them diverged. In this connection, L2-regularization with λ of 0.001 was applied to all training networks. All networks were trained with a batch size equal to 32. In order to avoid overfitting during network training, we applied Early Stopping regularization monitoring validation loss with patience equal to 10 epochs. For training networks in on both first and second stages we used the cross-entropy, calculated as follows:

$$Loss = -\sum_{i=1}^{C} \bar{y}_i * log(p_i + \varepsilon) \tag{1}$$

where $C$ is the number of classes (3 in our study), $p_i$ is the predicted probability, $y_i$ is the ground-truth label (ternary indicator), $\varepsilon$ is a small positive constant.

For the training and testing networks during the first stage, the dataset was split in an 8:1:1 ratio i.e. the training subset includes 2122 images (80.7%), the validation subset – 242 images (9.2%), and the testing subset – 267 images (10.1%). The split of data within training, validations, and testing phases was performed according to the distribution shown in Table 3.

Table 3
– Description of the data distribution within training, validation, and testing subsets

| Dataset | Diagnosis | Training | Validation | Testing |
|---|---|---|---|---|
| CCXRD | Normal | 14 | 2 | 2 |
| | Pneumonia | 133 | 15 | 17 |
| | COVID-19 | 407 | 46 | 51 |
| ACCXRD | Normal | 102 | 12 | 13 |
| | Pneumonia | 0 | 0 | 0 |
| | COVID-19 | 46 | 6 | 6 |
| FCCXRD | Normal | 1 | 1 | 1 |
| | Pneumonia | 0 | 1 | 1 |
| | COVID-19 | 27 | 4 | 4 |
| CRD | Normal | 0 | 0 | 0 |
| | Pneumonia | 0 | 0 | 0 |
| | COVID-19 | 177 | 20 | 22 |
| RSNAPDD | Normal | 567 | 63 | 70 |
| | Pneumonia | 648 | 72 | 80 |
| | COVID-19 | 0 | 0 | 0 |
| **Total** | | **2122 (80.7%)** | **242 (9.2%)** | **267 (10.1%)** |

## 3.2. Stage III

Once the performance and accuracy metrics of all networks were estimated, 4 networks that showed the best results on the first stage were chosen for fine-tuning. Besides training both bodies and heads of the networks, we introduced a guided attention mechanism for the considered networks. We were inspired by [34], where the authors proposed a framework that provides guidance on the attention maps generated by a weakly supervised deep learning neural network. The attention block in our pipeline is based on the usage of U-net [35]. As shown in Fig. 1, the proposed algorithm applies segmentation masks to the features of the network body (feature extractor) using multiplication. Applying an attention block to the output feature vector of the network's backbone allows networks to put more weight on the features that will be more relevant in the distinction of the different classes. Additionally during this stage, we applied attention maps obtained with help of the Grad-CAM technique [36]. Furthermore, the loss differs from the one on Stage I and Stage II and it is calculated as follows:

$$Loss = L_{clas} + \alpha L_{attn} \quad (2)$$

where $L_{clas}$ is the cross-entropy loss, $L_{attn}$ is the attention loss, $\alpha$ is the coefficient used to scale the total loss and the attention component. $L_{attn}$ is calculated according to Eq. (5) in [34].

To correctly apply U-net in the guided attention mechanism, we trained this network on the lung segmentation task. The data used for the training of this network is taken from the V7 Labs repository [37]. The segmentation dataset contains 6500 images of AP/PA chest X-ray images with pixel-level polygonal lung segmentations. Some examples of COVID-19 affected patients with segmented areas of lungs are shown in Fig. 2.

## 3.3. Visual model validation

While modern neural networks enable superior performance, their lack of decomposability into intuitive and understandable components makes them hard to interpret. In this regard, an achievement of the model transparency is useful to explain their predictions. Nowadays, one of the techniques used for model interpretation is known as Class Activation Map (CAM) [38]. Though CAM is a good technique to demystify the working of CNNs, it suffers from some limitations. One of the drawbacks of CAM is that it requires feature maps to directly precede the softmax layers, so it applies to a particular kind of network architecture that performs global average pooling over convolutional maps immediately before prediction. Such architectures may achieve inferior accuracies compared to general networks on some tasks or simply be inapplicable to new tasks. De facto deeper representations of a CNN capture the best high-level constructs. Furthermore, CNN’s naturally retrain spatial information which is lost in fully connected layers, so we expect the last convolutional layer to have the best tradeoff between high-level semantics and detailed spatial information. In this connection, we decided to use another popular technique known as Grad-Cam. This model interpretation technique, published in [36], aims to improve the shortcomings of CAM and claims to be compatible with any kind of architecture. The technique does not require any modifications to the existing model architecture, and this allows its application to any CNN based architecture. Unlike CAM, Grad-Cam uses the gradient information flowing into the last convolutional layer of a CNN to understand each neuron for a decision of interest. Grad-Cam improves on its predecessor, provides better localization and clear class discriminative saliency maps.

# 4. Results

## 4.1. Stage I

Having trained 10 neural networks, we found that 2 tend to diverge more than others. This is likely connected with the normalization layers. Networks such as MobileNet V2 and VGG-16 do not have Batch/Instance/Layer/Group Normalization layers in their architecture. In this regard, these networks start diverging (MobileNet V2) or hit a validation loss/accuracy plateau (VGG-16) after approximately 100 epochs. Popular regularization techniques such as Lasso Regression (L1 Regularization), Ridge Regression (L2 regularization), ElasticNet (L1-L2 regularization), Dropout and Early Stopping may help to avoid this problem. In this regard we applied Ridge Regression, Dropout layers and Early Stopping in our training pipeline. As for the remaining networks, they did not suffer from overfitting; however, they could not reach better validation loss/accuracy values. When a given model reached its best validation loss, we saved the associated model weights using saving callback. Figure 3 demonstrates how the networks were trained during the first stage. Blue asterisks reflect the best value of the accuracy on the validation subsets.

Since the loss value is poorly interpreted, we compared commonly used network metrics such as accuracy and F1-score. Table 4 and Table 5 summarize these metrics estimated in the first stage. As seen, MobileNet V2, EfficientNet B1, EfficientNet B3, and VGG-16 achieved better results than other networks.

Table 4
– Performance metrics within different subsets obtained after the first stage

| **Model** | **Accuracy** | | | **F1-score** | | |
|---|---|---|---|---|---|---|
| | **Training** | **Validation** | **Testing** | **Training** | **Validation** | **Testing** |
| MobileNet V2 | 0.95 | 0.79 | 0.77 | 0.95 | 0.80 | 0.78 |
| DenseNet-121 | 0.76 | 0.72 | 0.74 | 0.76 | 0.72 | 0.75 |
| EfficientNet B0 | 0.95 | 0.79 | 0.70 | 0.95 | 0.80 | 0.70 |
| EfficientNet B1 | 0.79 | 0.76 | 0.74 | 0.79 | 0.76 | 0.75 |
| EfficientNet B3 | 0.77 | 0.75 | 0.71 | 0.78 | 0.75 | 0.72 |
| EfficientNet B5 | 0.77 | 0.74 | 0.70 | 0.77 | 0.74 | 0.70 |
| VGG-16 | 0.90 | 0.79 | 0.78 | 0.90 | 0.80 | 0.79 |
| ResNet-50 V2 | 0.80 | 0.71 | 0.69 | 0.80 | 0.71 | 0.70 |
| Inception V3 | 0.77 | 0.71 | 0.73 | 0.77 | 0.71 | 0.74 |
| Inception ResNet V2 | 0.71 | 0.68 | 0.70 | 0.71 | 0.67 | 0.70 |

Table 5
– Performance metrics within different classes obtained after the first stage

| Model | Accuracy | | | F1-score | | |
|---|---|---|---|---|---|---|
| | Normal | Pneumonia | Covid-19 | Normal | Pneumonia | Covid-19 |
| MobileNet V2 | 0.70 | 0.78 | 0.83 | 0.74 | 0.75 | 0.83 |
| DenseNet-121 | 0.75 | 0.82 | 0.63 | 0.76 | 0.73 | 0.73 |
| EfficientNet B0 | 0.74 | 0.69 | 0.66 | 0.71 | 0.66 | 0.72 |
| EfficientNet B1 | 0.73 | 0.73 | 0.75 | 0.74 | 0.69 | 0.79 |
| EfficientNet B3 | 0.70 | 0.72 | 0.72 | 0.70 | 0.69 | 0.75 |
| EfficientNet B5 | 0.66 | 0.75 | 0.67 | 0.68 | 0.68 | 0.73 |
| VGG-16 | 0.80 | 0.76 | 0.78 | 0.77 | 0.75 | 0.82 |
| ResNet-50 V2 | 0.68 | 0.70 | 0.68 | 0.69 | 0.65 | 0.74 |
| Inception V3 | 0.74 | 0.77 | 0.68 | 0.75 | 0.71 | 0.75 |
| Inception ResNet V2 | 0.70 | 0.76 | 0.61 | 0.70 | 0.68 | 0.70 |

## 4.2. Stage II

Based on the results of the first stage, MobileNet V2, EfficientNet B1, EfficientNet B3, and VGG-16 demonstrated their ability to distinct COVID-19 and pneumonia on X-ray images much better than other networks. In this regard, these networks are chosen for fine-tuning. Additionally, we compared how fine-tuned networks differ from the best networks of the first stage. The results of the models' performance are shown in Fig. 4, where blue asterisks reflect the best value of the accuracy on the validation subsets.

Having compared accuracy and F1-score values on the first (Table 4 and Table 5) and second stage (Table 6 and Table 7), we can state that MobileNet V2 and VGG-16 have a larger boost in accuracy than EfficientNet models. Once the fine-tuning was performed, MobileNet V2 and VGG-16 got a + 6% and + 9% accuracy change on the validation subset and a + 1% and + 4% accuracy change on the testing subset. On the other hand, EfficientNet B1 and EfficientNet B3 a + 2% and + 3% accuracy change on the validation subset and a -1% and + 6% accuracy change on the testing subset. It should also be noted, that the largest boost in classification of COVID-19 was achieved by VGG-16. This network had an + 11% boost, while MobileNet V2, EfficientNet B1, and EfficientNet B3 could reach the level of + 2%, 0%, and + 6%, respectively.

Table 6
– Performance metrics within different subsets obtained after the second stage

| Model | Accuracy | | | F1-score | | |
|---|---|---|---|---|---|---|
| | Training | Validation | Testing | Training | Validation | Testing |
| MobileNet V2 | 1.00 | 0.85 | 0.78 | 1.00 | 0.85 | 0.79 |
| EfficientNet B1 | 0.83 | 0.78 | 0.73 | 0.83 | 0.78 | 0.74 |
| EfficientNet B3 | 0.83 | 0.78 | 0.77 | 0.83 | 0.78 | 0.77 |
| VGG-16 | 1.00 | 0.87 | 0.82 | 1.00 | 0.87 | 0.83 |

Table 7
– Performance metrics within different classes obtained after the second stage

| Model | Accuracy | | | F1-score | | |
|---|---|---|---|---|---|---|
| | Normal | Pneumonia | Covid-19 | Normal | Pneumonia | Covid-19 |
| MobileNet V2 | 0.74 | 0.75 | 0.85 | 0.75 | 0.74 | 0.85 |
| EfficientNet B1 | 0.70 | 0.74 | 0.75 | 0.72 | 0.71 | 0.78 |
| EfficientNet B3 | 0.77 | 0.75 | 0.78 | 0.76 | 0.74 | 0.81 |
| VGG-16 | 0.81 | 0.78 | 0.89 | 0.80 | 0.79 | 0.89 |

## 4.3. Stage III

Having trained the chosen networks according to the pipeline described in Sect. 3 and 3.2, we compared them on the validation and testing subsets (Fig. 5 and Fig. 6). Based on the obtained results we established that the proposed pipeline allows for boosting of the model accuracy. VGG-16 and MobileNet V2 showed the best accuracy on the validation and testing subsets. It is worth noticing that the VGG-16 network outperformed the best CXR model (CXR-4A) on these subsets. The performance of other CXR models is shown in Appendix A. It is observed that the VGG-16 (S3) network trained based on the proposed pipeline has a + 9% and + 1% of accuracy boost on the validation subset compared to VGG-16 (S1) and VGG-16 (S2) respectively. Similar positive dynamics of using our pipeline is observed for other models as well. It should be noted that the CXR-4A and lightweight MobileNet V2 have almost the same accuracy, while the complexity of the latter is 15.5 times lower. The MobileNet V2 network includes 2.6 mln. weights, while CXR-4A – 40.2 mln. weights.

## 4.4. Model validation using Grad-CAM

As we mentioned in Sect. 3.3, despite deep learning models having facilitated unprecedented accuracy in image classification, one of their biggest problems is model interpretability representing a core component in understanding and debugging of a model. We used the Grad-CAM technique to validate the models and their correct/incorrect ability for making predictions, and to verify which series of neurons activated in the forward-pass during the prediction. For the sake of visualization, we choose 3 patients with different findings: normal, pneumonia, and COVID-19. Source images of these findings with their ground truth (GT) heatmaps are shown in Fig. 7 and Fig. 8.

Using Grad-CAM, we validated where our 4 best networks (MobileNet V2, EfficientNet B1, EfficientNet B3, VGG-16) are looking, verifying that they are properly looking at the correct patterns in the image and activating around those patterns. The Grad-CAM technique uses the gradients, flowing into the final convolutional layer to produce a coarse localization heatmap highlighting the important regions in the image for predicting the target concept i.e. COVID-19 or pneumonia areas. However, the localization heatmaps may differ from the traditional localization techniques such as segmentation masks or bounding boxes. In this regard, these heatmaps are used for the sake of approximate localization.

In order to interpret the models, Fig. 7 and Fig. 8 reflect the visualization of gradient class activation maps. Additional cases of the networks’ heatmaps are shown in Appendix B and Appendix C. Based on the obtained results, we may state that the training of the models using masks (Stage III) has a positive effect on the search for the correct patterns by the models. Networks such as MobileNet V2 (Fig. 7c and Fig. 8c) and VGG-16 (Fig. 7f and Fig. 8f) identify affected areas correctly, despite the inaccuracies in the location of the heatmaps. On the other hand, interpretation of the EfficientNet networks showed that they are not activating around the proper patterns in the image. This allows us to assume that EfficientNet B1 and EfficientNet B3 have not properly learned the underlying patterns in our dataset and/or we may need to collect additional data.

# 5. Conclusion

In this study, we demonstrated a training pipeline based on directed guidance for neural networks. This guidance forces the neural networks to pay attention to the areas obtained by the external network. Having trained a set of deep learning models, we found that the proposed pipeline allows for increased classification accuracy. This pipeline was used for the detection of COVID-19 and distinguishing its presence from that of pneumonia. Of the obtained results, MobileNet V2 performed comparably to the tailor-made CXR model CXR-4A, despite being 15 times less complex. According to the performed experiments, the networks trained based on the proposed pipeline perform comparably to practicing radiologists when it comes to the classification of multiple thoracic pathologies in chest X-ray radiographs. Our pipeline may have the potential to improve healthcare delivery and increase access to chest radiograph expertise for the detection of a variety of acute diseases.

# Declarations

## Acknowledgements

The study was supported in part by the Ministry of Science and Higher Education, project No. FFSWW-2020-0014 “Development of the technology for robotic multiparametric tomography based on big data processing and machine learning methods for studying promising composite materials”.

## Author Contributions

Y.G., S.S., and O.T. conceived the idea of the study. V.D., Y.G., and O.T. developed the plan of execution. V.D. and O.T. collected and annotated the data. V.D. and D.N. developed, trained, and analyzed the performance of deep learning networks on the collected data. A.P. tested the performance of purpose-built deep learning networks (COVID-Nets) on the collected data. V.D. and A.P. wrote the manuscript with input from all the co-authors. A.K., A.K., V.K., V.S., and D.L. assisted in study direction and data quality discussions. Y.G. supervised the project.

## Competing interests

The authors have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this paper

## Additional information

**Correspondence** and requests for materials should be addressed to V.D. and Y.G.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

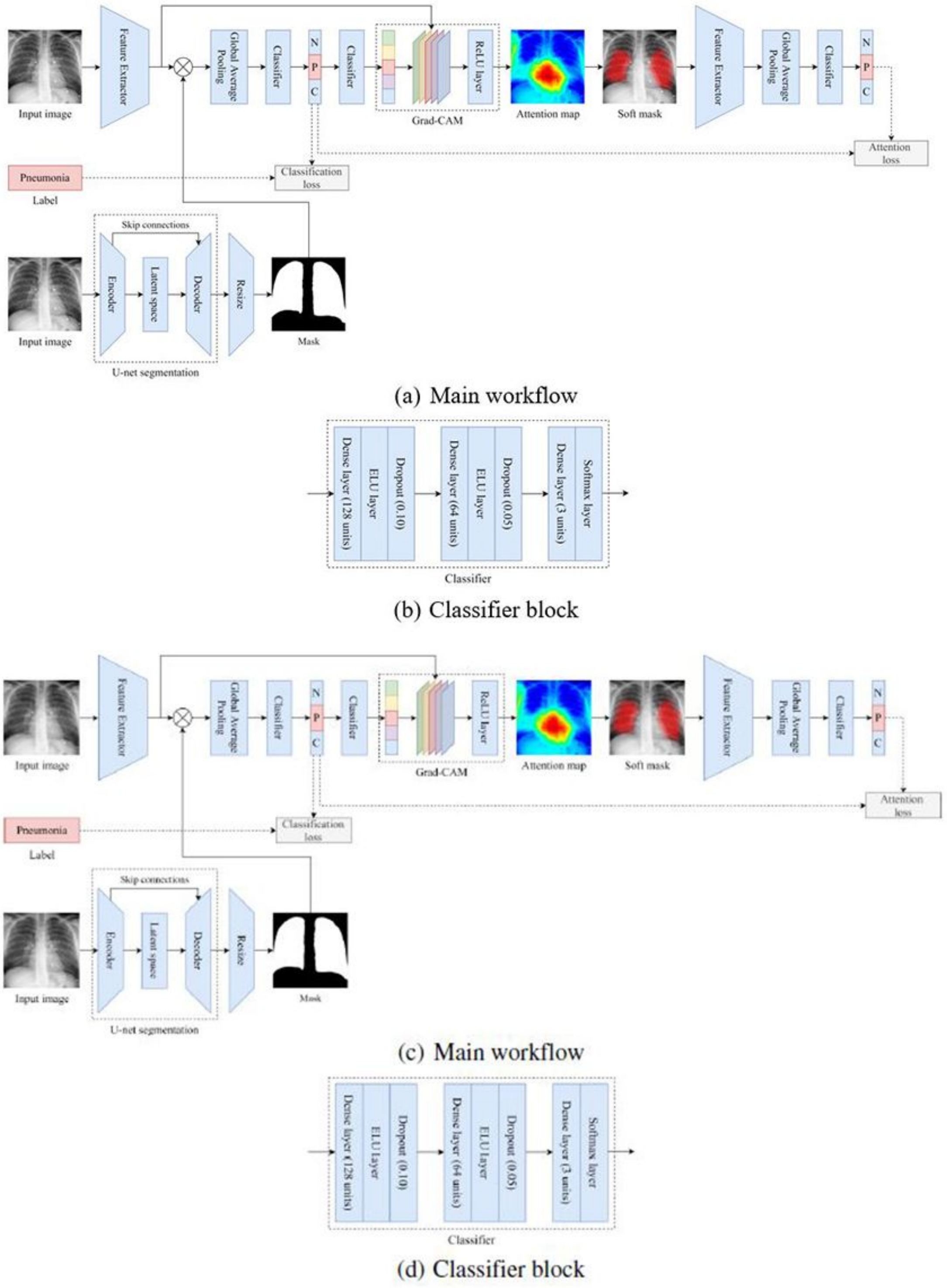
# References

1. COVID-19 Virus Pandemic - Worldometer. https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1.
2. Wang, W. *et al.* Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA - Journal of the American Medical Association.* **323**, 1843–1844 (2020).
3. Wikramaratna, P., Paton, R., Ghafari, M. & Lourenco, J. Estimating false-negative detection rate of SARS-CoV-2 by RT-PCR. *medRxiv* 2020.04.05.20053355 (2020) doi:10.1101/2020.04.05.20053355.
4. Yang, Y. *et al.* Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections. *medRxiv* 2020.02.11.20021493 (2020) doi:10.1101/2020.02.11.20021493.
5. Fang, Y. *et al.* Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. *Radiology.* **296**, E115–E117 (2020).
6. Guan, W. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).
7. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet.* **395**, 497–506 (2020).

8. Ng, M. Y. *et al.* Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review. *Radiol. Cardiothorac. Imaging.* **2**, e200034 (2020).
9. Kanne, J. P., Little, B. P., Chung, J. H., Elicker, B. M. & Ketai, L. H. Essentials for radiologists on COVID-19: An update-radiology scientific expert panel. Radiology vol. 296 E113–E114(2020).
10. Ai, T. *et al.* Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology.* **296**, E32–E40 (2020).
11. Rubin, G. D. *et al.* The Role of Chest Imaging in Patient Management During the COVID-19 Pandemic: A Multinational Consensus Statement From the Fleischner Society. *Chest.* **158**, 106–116 (2020).
12. Wang, L., Lin, Z. Q. & Wong, A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **10**, 19549 (2020).
13. Mahmud, T., Rahman, M. A., Fattah, S. A. & 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**, 103869 (2020).
14. Farooq, M., Hafeez, A. & COVID-ResNet: A Deep Learning Framework for Screening of COVID19 from Radiographs. arXiv(2020).
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.* **65**, 101794 (2020).
16. Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T. & Saalbach, A. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *Sci. Rep.* **9**, 6381 (2019).
17. Cohen, J. P. *et al.* COVID-19 Image Data Collection: Prospective Predictions Are the Future. arXiv(2020).
18. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 Image Data Collection. arXiv(2020).
19. Wang, L. *et al.* Actualmed COVID-19 Chest X-ray Dataset Initiative. https://github.com/agchung/Actualmed-COVID-chestxray-dataset (2020).
20. Wang, L. *et al. Figure 1* COVID-19 Chest X-ray Dataset Initiative. https://github.com/agchung/Figure1-COVID-chestxray-dataset (2020).
21. COVID-19 Radiography Database | Kaggle. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.
22. Chowdhury, M. E. H. *et al.* Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access.* **8**, 132665–132676 (2020).
23. RSNA Pneumonia Detection Challenge | Kaggle. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge.
24. Litmanovich, D. E., Chung, M., Kirkbride, R. R., Kicska, G. & Kanne, J. P. Review of Chest Radiograph Findings of COVID-19 Pneumonia and Suggested Reporting Language. *J. Thorac. Imaging.* **35**, 354–360 (2020).
25. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 4510–4520(2018).

26. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. in *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR* 2017 vols 2017-Janua 2261–2269 (Institute of Electrical and Electronics Engineers Inc., 2017).
27. Tan, M., Le, Q. V. & EfficientNet Rethinking Model Scaling for Convolutional Neural Networks. 36th Int. Conf. Mach. Learn. ICML 2019 2019-June, 10691–10700(2019).
28. Liu, S. & Deng, W. Very deep convolutional neural network based image classification using small training sample size. in Proceedings – 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015 730–734 (Institute of Electrical and Electronics Engineers Inc., 2016). doi:10.1109/ACPR.2015.7486599.
29. He, K., Zhang, X., Ren, S. & Sun, J. Identity Mappings in Deep Residual Networks. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics).* **9908 LNCS**, 630–645 (2016).
30. 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 vols 2016-Decem 2818–2826 (IEEE Computer Society, 2016).
31. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-ResNet and the impact of residual connections on learning. in *31st AAAI Conference on Artificial Intelligence, AAAI 2017* 4278–4284 (AAAI press, 2017).
32. Wang, L. *et al.* COVID-Net: COVID-Net Open Source Initiative. https://github.com/lindawangg/COVID-Net.
33. Amazon SageMaker – Machine Learning – Amazon Web Services. https://aws.amazon.com/sagemaker/.
34. Li, K., Wu, Z., Peng, K. C., Ernst, J. & Fu, Y. Tell Me Where to Look: Guided Attention Inference Network. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 9215–9223(IEEE Computer Society, 2018). doi:10.1109/CVPR.2018.00960.
35. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol. 9351 234–241(Springer Verlag, 2015).
36. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. in *Proceedings of the IEEE International Conference on Computer Vision* vols 2017-Octob 618–626 (Institute of Electrical and Electronics Engineers Inc., 2017).
37. COVID-19 X-ray dataset. https://github.com/v7labs/covid-19-xray-dataset.
38. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning Deep Features for Discriminative Localization. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition vols 2016-Decem 2921–2929 (IEEE Computer Society, 2016).

# Figures



(a) Main workflow

(b) Classifier block

(c) Main workflow

(d) Classifier block

## Figure 1

Proposed workflow for the classification with the guided attention mechanism

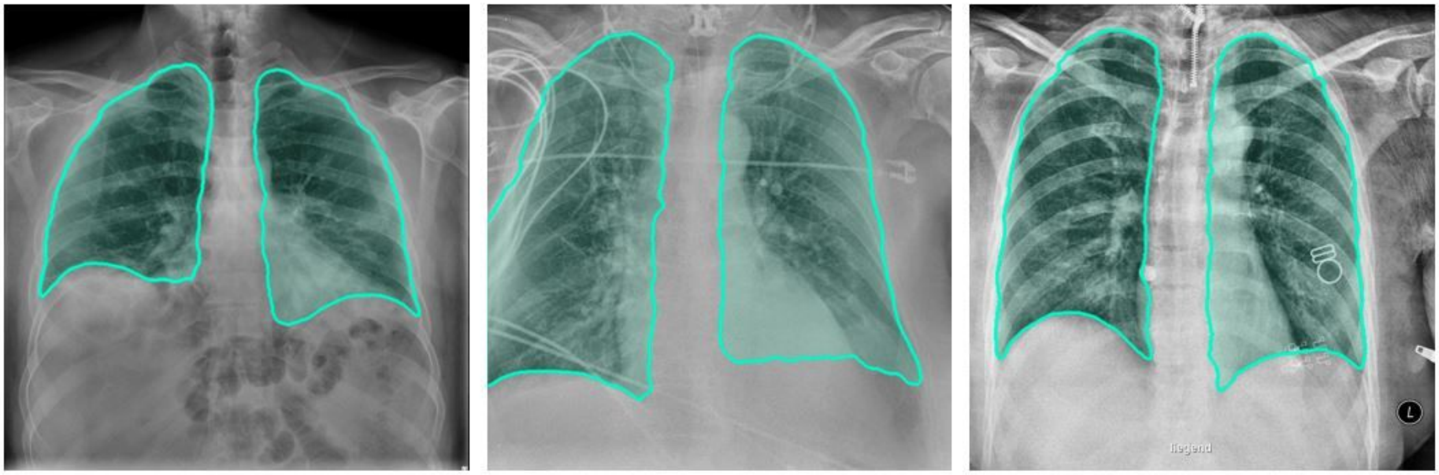
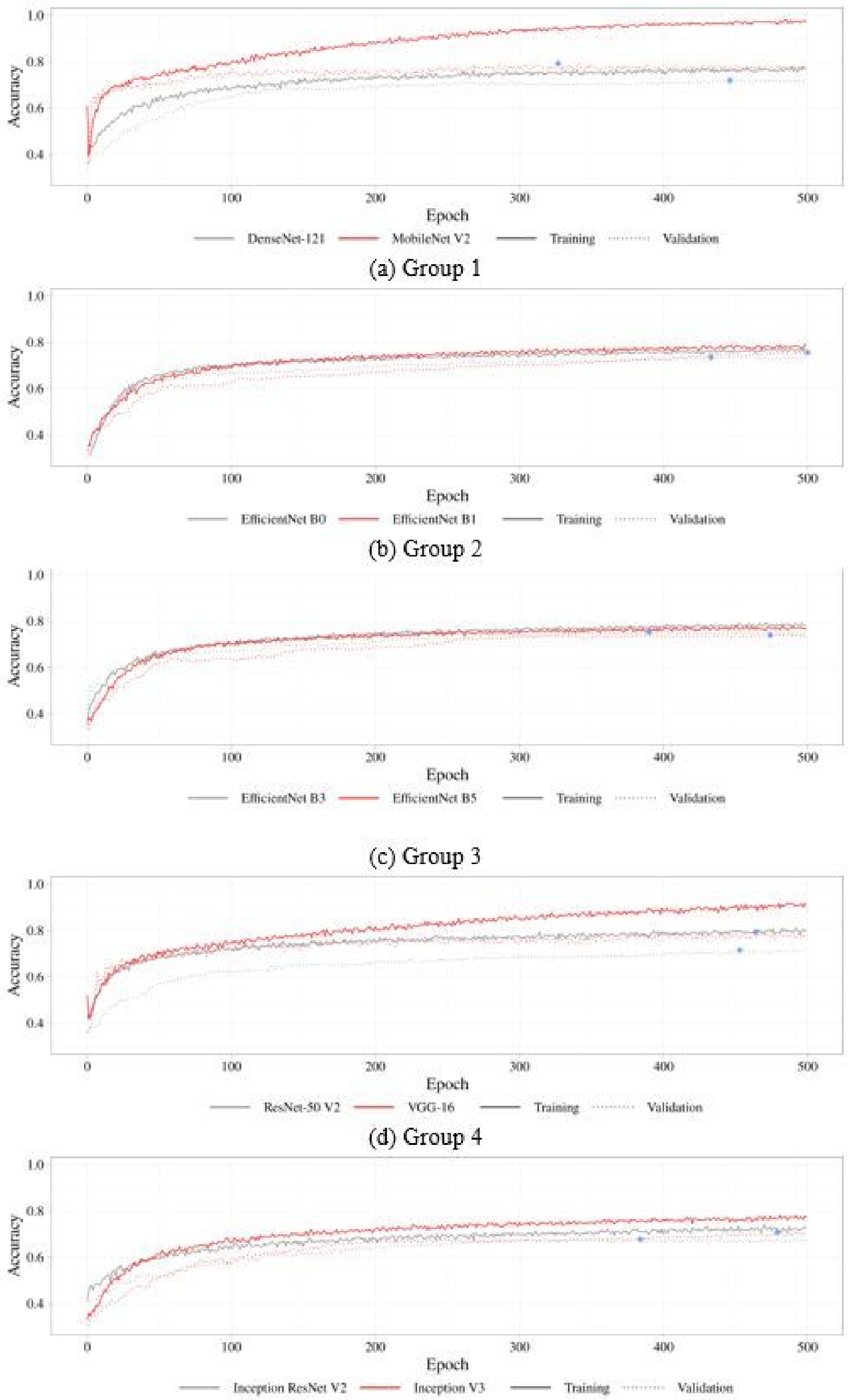


Figure 2

Examples of COVID-19 patients with segmented lungs



(a) Group 1

(b) Group 2

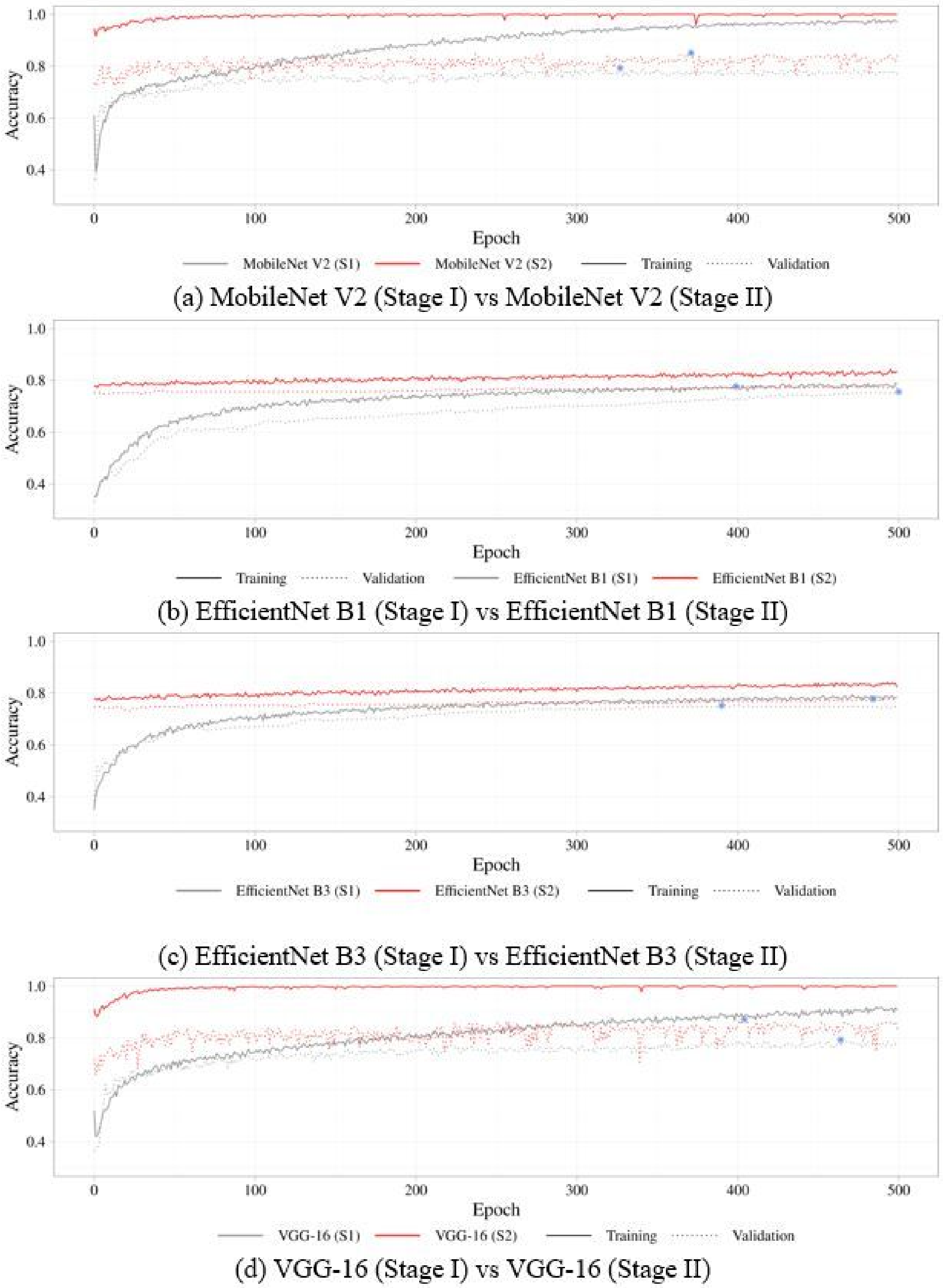
(c) Group 3

(d) Group 4

(e) Group 5

## Figure 3

Accuracy dynamics over training during the first stage



(a) MobileNet V2 (Stage I) vs MobileNet V2 (Stage II)

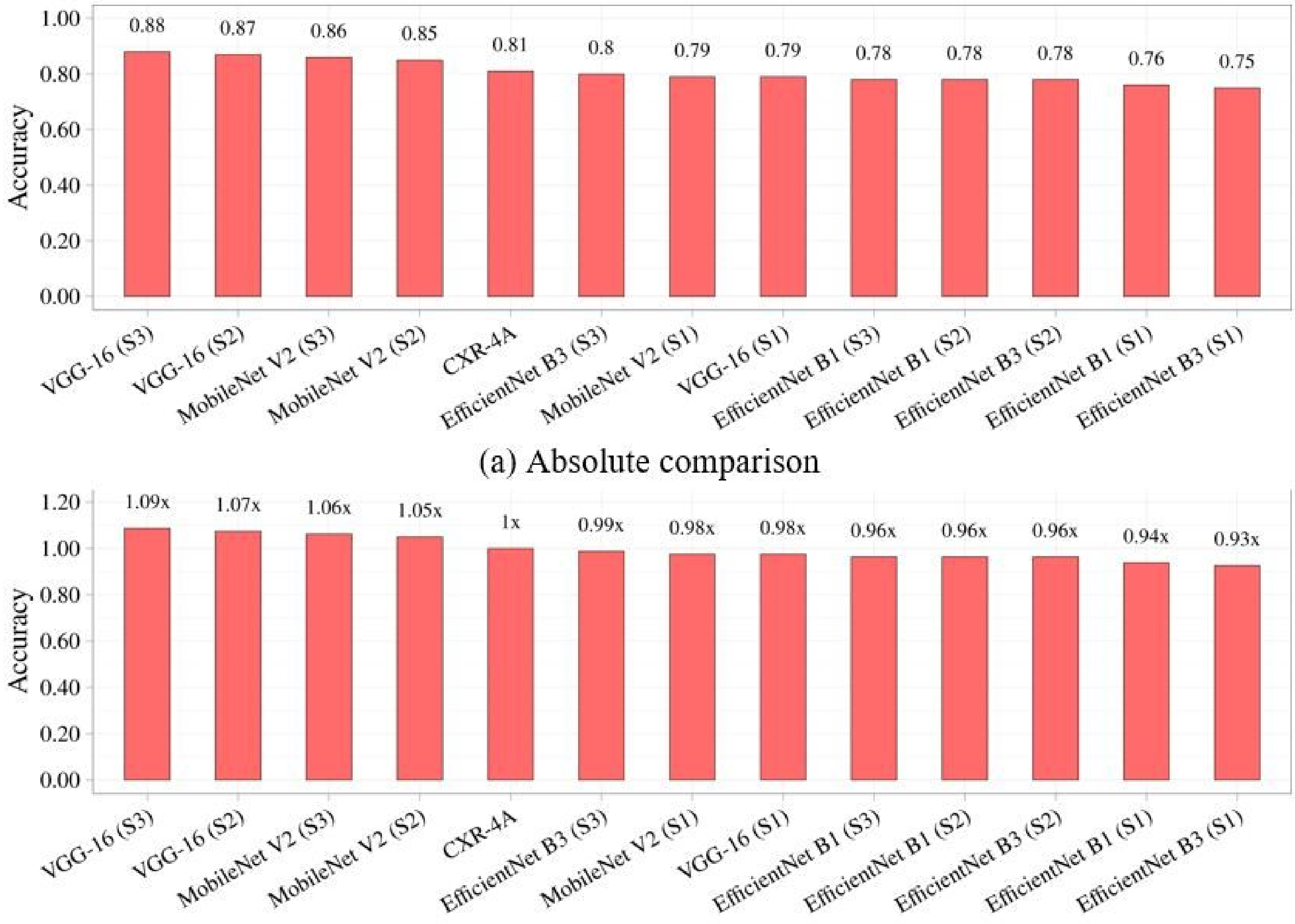
(b) EfficientNet B1 (Stage I) vs EfficientNet B1 (Stage II)

(c) EfficientNet B3 (Stage I) vs EfficientNet B3 (Stage II)

(d) VGG-16 (Stage I) vs VGG-16 (Stage II)

## Figure 4

Accuracy dynamics over training during the second stage

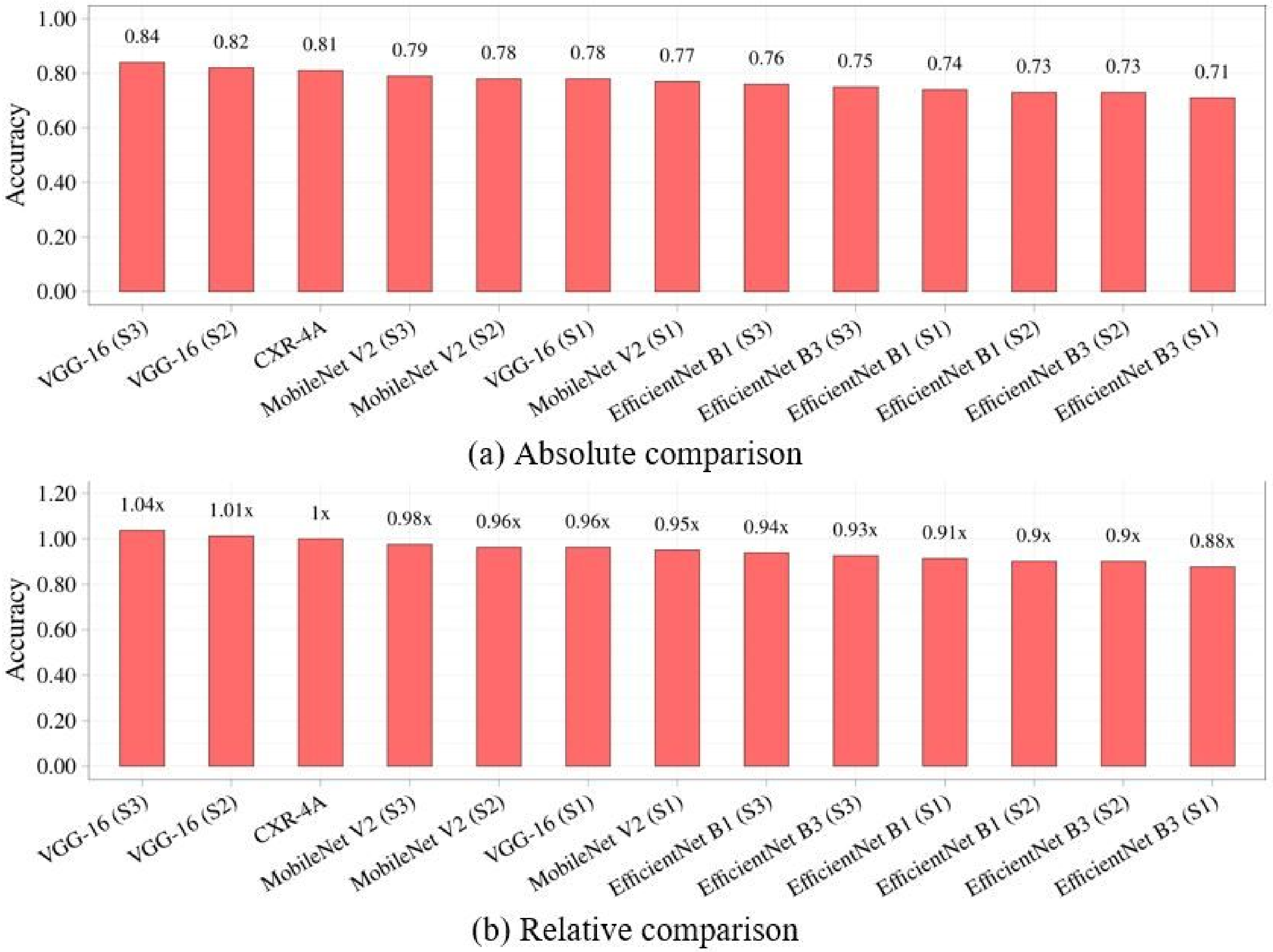


(a) Absolute comparison

(b) Relative comparison

## Figure 5

Comparison of accuracy based on the validation subset

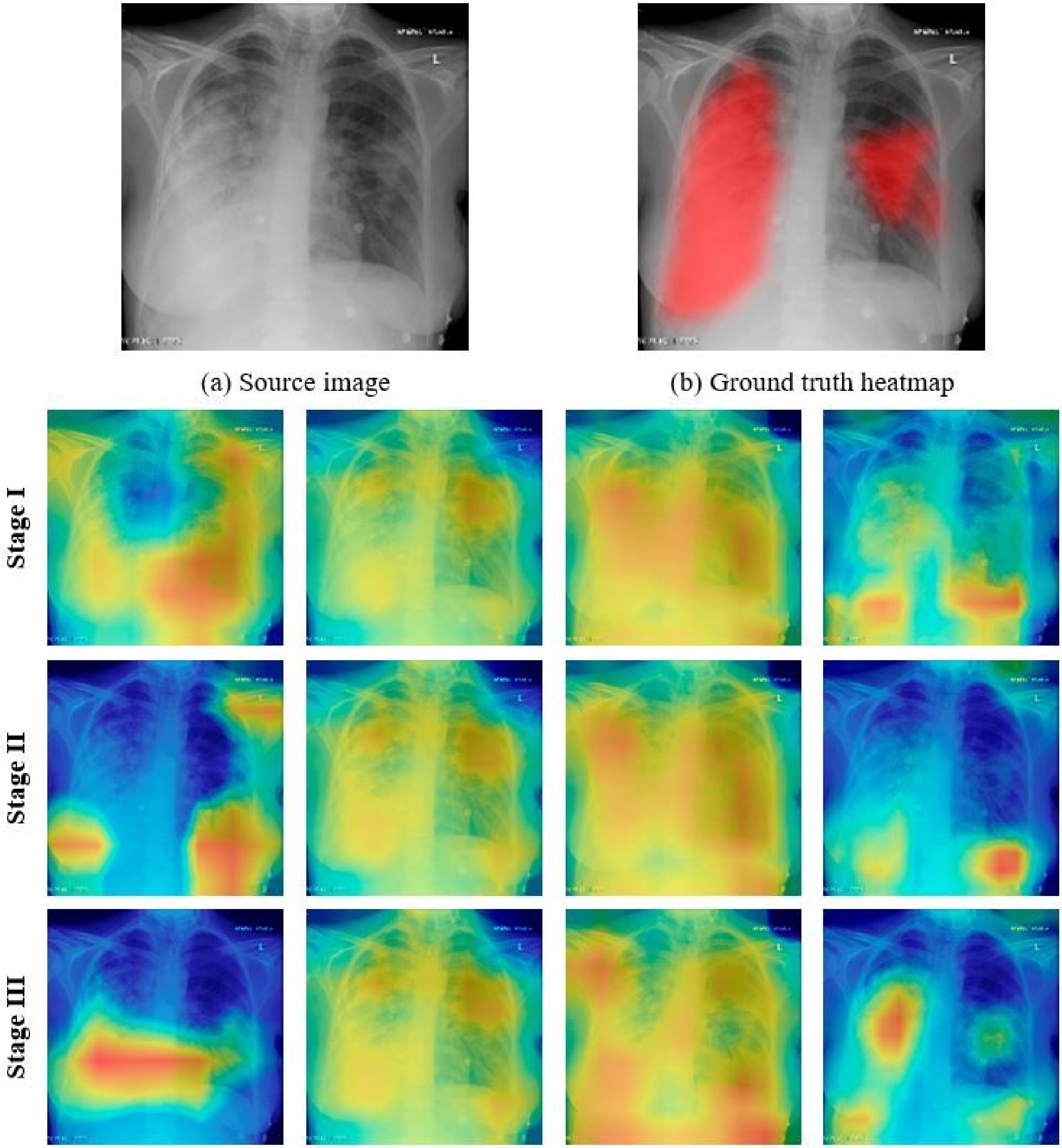


(a) Absolute comparison

(b) Relative comparison

## Figure 6

Comparison of accuracy based on the testing subset



(a) Source image

(b) Ground truth heatmap

(c) MobileNet V2

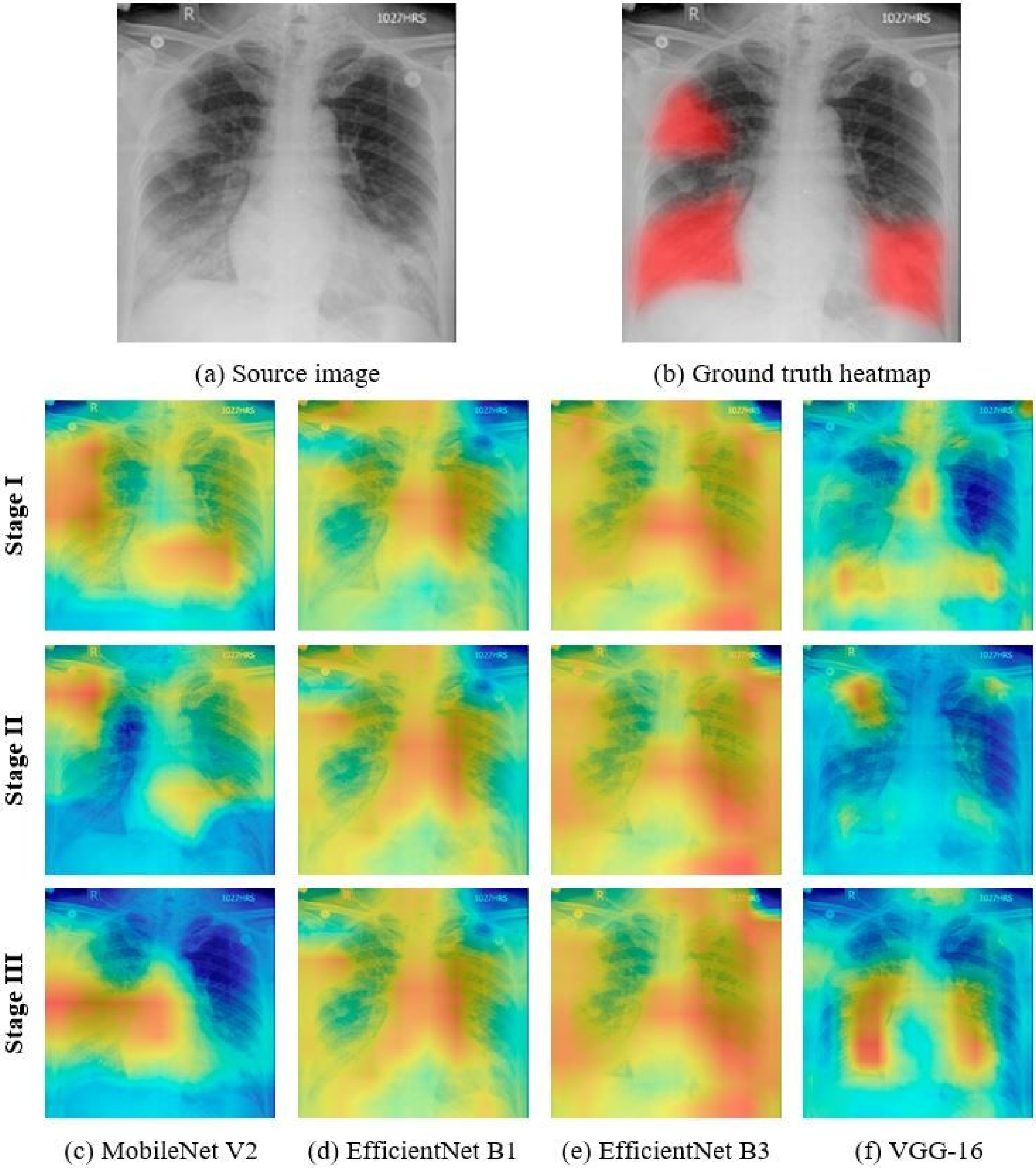
(d) EfficientNet B1

(e) EfficientNet B3

(f) VGG-16

**Figure 7**

Visualization of network heatmaps for a pneumonia finding



(a) Source image

(b) Ground truth heatmap

(c) MobileNet V2

(d) EfficientNet B1

(e) EfficientNet B3

(f) VGG-16

### Figure 8

Visualization of network heatmaps for a COVID-19 finding

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx