Study on Transfer Learning Capabilities for Pneumonia Classification in Chest-X-Rays Images

Danilo Avola[a], Andrea Bacciu[a], Luigi Cinque[a], Alessio Fagioli[a,*], Marco Raoul Marini[a], Riccardo Taiello[a]

[a]*Department of Computer Science, Sapienza University, Via Salaria 113, 00185, Rome, Italy*

arXiv:2110.02780v1 [eess.IV] 6 Oct 2021

Abstract

Background: over the last year, the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) and its variants have highlighted the importance of screening tools with high diagnostic accuracy for new illnesses such as COVID-19. To that regard, deep learning approaches have proven as effective solutions for pneumonia classification, especially when considering chest-x-rays images. However, this lung infection can also be caused by other viral, bacterial or fungi pathogens. Consequently, efforts are being poured toward distinguishing the infection source to help clinicians to diagnose the correct disease origin. Following this tendency, this study further explores the effectiveness of established neural network architectures on the pneumonia classification task through the transfer learning paradigm.

Methodology: to present a comprehensive comparison, 12 well-known ImageNet pre-trained models were fine-tuned and used to discriminate among chest-x-rays of healthy people, and those showing pneumonia symptoms derived from either a viral (i.e., generic or SARS-CoV-2) or bacterial source. Furthermore, since a common public collection distinguishing between such categories is currently not available, two distinct datasets of chest-x-rays images, describing the aforementioned sources, were combined and employed to evaluate the various architectures.

Results: the experiments were performed using a total of 6330 images split between train, validation and test sets. For all models, common classification metrics were computed (e.g., precision, f1-score) and most architectures obtained significant performances, reaching, among the others, up to 84.46% average f1-score when discriminating the 4 identified classes. Moreover, confusion matrices and activation maps computed via the Grad-CAM algorithm were also reported to present an informed discussion on the networks classifications.

Conclusion: this paper examines the effectiveness of well-known architectures on a joint collection of chest-x-rays presenting pneumonia cases derived from either viral or bacterial sources, with particular attention to SARS-CoV-2 contagions for viral pathogens; demonstrating that existing architectures can effectively diagnose pneumonia sources and suggesting that the transfer learning paradigm could be a crucial asset in diagnosing future unknown illnesses.

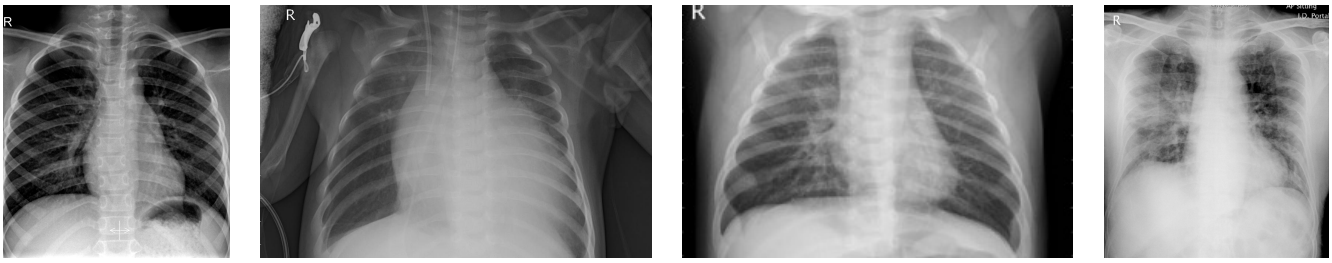
Keywords: Pneumonia classification, Deep learning, Transfer learning, Explainable AI

1. Introduction

Pneumonia is an acute respiratory infection caused by viral, bacterial or fungal pathogens. This infection can affect either a single or both lungs and can cause mild to life-threatening effects in people of all ages. Notably, there are increased risks for those with preexisting health conditions or in adults with more than 65 years, and is the largest worldwide infectious cause of death for children; accounting for 15% of children deaths in 2017, according to World Health Organization (WHO) statistics. Among the possible sources, viruses and bacteria are the most common pneumonia infection causes [1] and, depending on the pathogen, the illness shows a different behavior. In particular, for viral agents, the infection is generally diffused across all lungs [2], while for bacterial ones it usually concentrates on given areas [3]. Therefore, due to intrinsic differences between viruses and bacteria, as well as their potential treatment, a key aspect is to correctly diagnose the pneumonia source. The latter has become even more crucial since the discovery of a novel coronavirus in 2019 [4], that has resulted in the SARS-CoV-2 virus and related COVID-19 disease pandemic outbreak. Indeed, while many works are addressing this diagnosis task [5], there is still a need for increasingly more general research that might be used to oppose the possibly already endemic SARS-CoV-2 [6], its future variants [7], or completely new diseases that might result in other pandemics [8].

To address the pneumonia classification task and help clinicians with their diagnoses, an effective solution is to analyze chest-x-rays images. Indeed, as can be observed in Fig. 1, the infection spread can be clearly visible through the x-ray technology, especially for severe conditions such as the one associated to the SARS-CoV-2 virus. To fully exploit information found in such images, deep learning approaches are being explored since they can already obtain significant performances in several heterogeneous medical fields [9, 10, 11, 12]. In detail, many solutions focus on the binary classifica-

*Corresponding author: Tel.: +39 06 8576 8425;
Email address: fagioli@di.uniroma1.it (Alessio Fagioli)



(a) (b) (c) (d)

Figure 1: Chest-x-ray samples showing a healthy patient (a), pneumonia from a bacteria infection (b), pneumonia from a virus infection (c), and pneumonia from a COVID-19 case (d).

tion between COVID-19 patients and healthy people [13, 14, 15, 16, 17, 18]. For instance, the authors of [15] design a custom deep neural network (DNN) model with a hybrid diagnose strategy (HDS) and fuzzy inference engine to identify COVID-19 patients, while in [18] transfer learning is applied to 4 pre-trained networks to distinguish between SARS-CoV-2 infected and healthy people. Although a binary classification allows a given model to completely focus on the COVID-19 disease, several works try to build more general models by also differentiating pneumonia patients with other sources (i.e., bacterial or virus) through the use of deep learning methods [19, 20, 21, 22, 23, 24]. In [21], for example, the authors devise a stacking ensemble method to leverage several ImageNet pre-trained architectures and discern between healthy, COVID-19 and other pneumonia-affected patients; while a custom architecture based on long-short term memory (LSTM) units is trained from scratch in [19] to achieve a similar purpose. Finally, different works propose yet another abstraction step by recognizing 4 different classes, i.e., healthy, COVID-19, as well as pneumonia patients infected via bacterial or other viral pathogens [25, 26, 27, 28, 29, 30]. In particular, the authors of [25] introduce a convolutional neural network (CNN), called CoroNet, specifically designed and trained to discriminate between the aforementioned categories. In [28], instead, 5 ImageNet pre-trained architectures are combined via the ensemble technique to increase their system classification performance.

While most literature works tend to introduce specific architectures to handle the classification task, others simply focus on transfer learning capabilities by fine-tuning known architectures on pneumonia datasets [31, 13]. Following this methodology, we performed a transfer learning study by fine-tuning 12 models on an extended collection derived by combining two public datasets presented, respectively, in [32] and [13]. In particular, we froze all Imagenet pre-trained models up to their respective classification component, as summarized in Fig. 2, which was instead modified to handle the classification task of the 4 available categories, i.e., healthy and pneumonia derived from bacterial, viral and SARS-CoV-2 pathogens. This allows to train even complex models in less time and enables the presented architectures to be ready to handle a complex task such as medical image classification. Furthermore, we also present qualitative experiments to explain the networks choices by directly observing what they focus on inside the chest-x-ray images via the Grad-CAM algorithm [33]. Presenting such a comprehensive transfer learning overview helps to choose better backbone architectures to build custom models upon, and offers a strong methodology for possible future diseases which might require to be diagnosed as early as possible, and for which it might be difficult to design good performing neural networks right away.

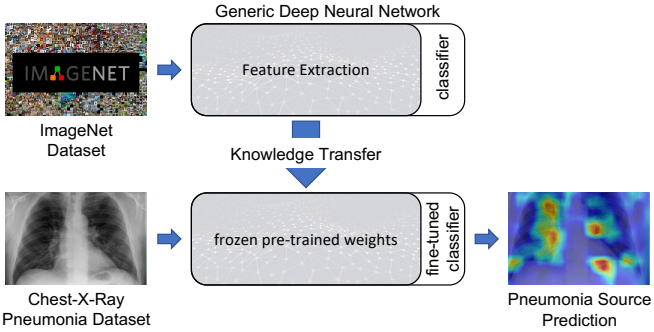


Figure 2: Transfer learning paradigm scheme for pneumonia classification.

The rest of this paper is structured as follows. A brief introduction on the rationale behind each pre-trained model used in this work is provided in Section 2. The quantitative and qualitative experimental results, as well as a discussion on the obtained performances, are presented in Section 3. Finally, work conclusions and future plans are drawn in Section 4.

2. Materials and Methods

The intuition behind transfer learning is to use a model pre-trained on a large and heterogeneous dataset, so that its generic world representation can be exploited on a new and different task [34]. In particular, when analyzing images, this learning approach takes advantage of feature maps learned by a network, without requiring to train the model from scratch and, therefore, avoiding a time-consuming and resource-intensive procedure. To transfer previous knowledge to the new task, there are two possible approaches. In the first one, a pre-trained model is treated as a feature extractor by freezing its internal weights. A classifier is then trained on top of this frozen architecture to achieve the knowledge transferal, and retain as

	Bacteria	Normal	Virus	COVID-19		Total
		[32]		[13]	[48, 49][1]	
Train	2249	1060	1056	84	-	4449
Val	289	289	289	-	289	1156
Test	242	234	148	101	-	725

[1]https://github.com/ieee8023/covid-chestxray-dataset

Table 1: Final merged dataset distribution.

much information as possible from the previous field. In the second approach, instead, the whole architecture, or a subset thereof, is fine-tuned on the new task. In this case, the pre-trained model weights are used as a baseline instead of starting from a random initialization. This generally allows a network to reach training convergence faster, and to specialize its entire structure to the new field. While both methodologies can be effective, in this work we explore the former approach and assess the capabilities of various networks used as feature extractors. Specifically, this allows to emulate the scarcity of data, typical when new illnesses are first discovered, but still enables to correctly train otherwise data hungry networks [35]. In more detail, we performed experiments on a combined collection derived from datasets presented in [32] and [13] to evaluate 12 well-known ImageNet pre-trained models briefly discussed later in this section, i.e., AlexNet [36], DenseNet [37], GoogleNet [38], MnasNet [39], MobileNetv2 [40], MobileNet v3 (Large) [41], ResNet50 [42], ResNeXt [43], ShuffleNet [44], SqueezeNet [45], VGG16 [46] and Wide ResNet50 [47]. Moreover, each model $\mathcal{M}$ was modified and adapted to the new task by changing the output layer $\mathcal{M}_o$ nodes number, by exploiting the following equation:

$$\mathcal{M}_{logits} = \mathbf{W}_o * Dropout(ReLU(\mathbf{W}_h\mathbf{x})), \tag{1}$$

where $\mathcal{M}_{logits}$ and $\mathbf{W}_o$ represent the model output logits and $\mathcal{M}_o$ weights, respectively; $Dropout(\cdot)$ is a regularization technique employed to avoid overfitting; $ReLU(\cdot)$ corresponds to the chosen activation function applied to the incoming inputs; $\mathbf{W}_h$ indicates weights between $\mathcal{M}_o$ and its preceding hidden layer $\mathcal{M}_h$; while $\mathbf{x}$ denotes the $\mathcal{M}_h$ output. In particular, given the model-dependent input size $n = |\mathbf{x}|$, $\mathbf{W}_h$ and $\mathbf{W}_o$ will have a dimension of $\mathbb{R}^{n\times256}$ and $\mathbb{R}^{256\times4}$, respectively. Finally, to train a model $\mathcal{M}$, we employed the categorical cross-entropy loss $\mathcal{L}$ defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=0}^{N-1} y_i \log \hat{y}_i, \tag{2}$$

where N corresponds to the number of classes, i.e., normal, bacteria, virus, and COVID-19; while $\hat{y}_i$ and y_i indicate, respectively, the predicted output and ground truth labels.

2.1. AlexNet

The AlexNet [36] architecture had a significant impact in the computer vision field over the last decade. Its success can be related to the ImageNet LSVRC-2012 [50] competition, which it won by a large margin compared to the second classified, i.e., with an increase of 10.9%. The architecture comprises 5 convolutional layers using kernels with size 11×11, 5×5, and 3×3 for the last three layers. Moreover, max pooling, dropout, and ReLU are applied after the first, second, and fifth convolution to extract relevant features. Finally, the latter are classified through 3 fully connected layers with a shape of 4096, 4096, and 1000, respectively.

2.2. DenseNet

A more recent model is the DenseNet [37], developed to mitigate the vanishing/exploding gradient problem which is a typical issue in deep neural networks with many layers. To this end, the authors organized their network into dense blocks interleaved by transition layers that can act as a bottleneck to reduce the architecture parameters number as well as the input size. Moreover, a dense block can contain several convolutions (e.g., 6, 12, 24), and each of these operations receives as input all of the feature maps produced by its preceding layers. This allows to forward more information inside a given block, and results in an overall more compact configuration, via this dense blocks subdivision, as well as in improved performances.

2.3. GoogLeNet

Another key architecture is GoogLeNet [38], organized into 27 layers containing convolutions as well as 5 max pooling operations and 9 inception modules. The latter are designed to allow a more efficient computation through a dimensionality reduction achieved via 1×1 convolutions. What is more, on top of these convolutions, GoogLeNet also retains crucial spatial information by analyzing the input at different scales by using kernel sizes of 1×1, 3×3, and 5×5 inside its inception modules. Finally, intermediate classifications are also performed to reinforce the final network output and improve the network gradient throughout its training.

2.4. MnasNet

A network specifically optimized to obtain significant performances on mobile devices through the Mobile Neural Architecture Search (MNAS) approach, is the MnasNet [39]. In particular, the architecture latency, i.e., the time required to perform inference on a given image, is incorporated directly into the main objective to be minimized at training time; enabling the model search to find a good trade-off between accuracy and latency. Moreover, the MnasNet employs 9 layers comprising convolutions as well as depthwise separable convolutions and mobile inverted bottlenecks, which exploit inverted residual connections. Notice that both techniques were defined by the MobileNet v2, presented in the following section.

2.5. MobileNet v2

An earlier network devised for mobile devices is the MobileNet v2 [40]. This model implements a fully convolution layer followed by 19 residual bottleneck layers that extend the depth-wise separable convolution designed for the MobileNet [51]. In detail, this operation applies convolutions on the input channels in a disjoint fashion, then merges the resulting outputs

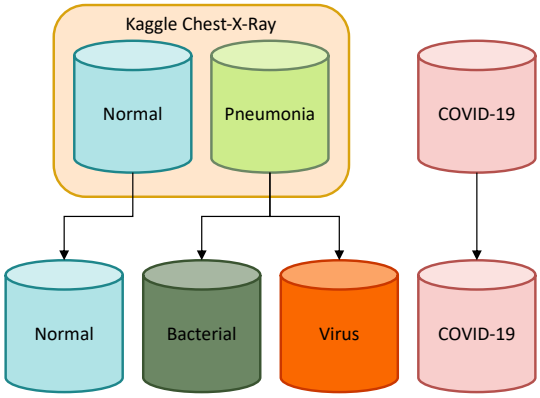


Figure 3: Dataset subdivision schema.

via 1×1 point-wise convolution. In MobileNet v2, this operation is further refined through an inverted residual block that expands, performs a depth-wise separable convolution, then compresses the output to filter features; preventing overfitting and reducing the model parameter number.

2.6. MobileNet v3

A further advancement for mobile networks is also provided by MobileNet v3 [41], where features extracted by inverted residual blocks with depth-wise separable convolutions, i.e., MobileNet v2 blocks, are improved through the squeeze-and-excitation technique [52]. Specifically, the latter explicitly models interdependencies between channels by recalibrating channel-wise feature responses through adaptive weights associated to each feature map. Furthermore, AutoML [53] is also employed to find the best neural network architecture using the extended inverted residual blocks mentioned above.

2.7. ResNet50

An effective architecture focusing on the extracted features is the ResNet50 [42] which, as the name implies, contains 50 convolutional layers. In more detail, this architecture applies a first convolution and a max pooling operation to consistently reduce the input size. Subsequently, residual blocks are implemented to reach the desired size of 50 convolutional layers. Moreover, a key component of this architecture lies in these residual blocks, which leverage residual connections. The latter define an identity shortcut that skips 3 layers and ensures input data is feed-forwarded throughout the model, enabling for a higher information retention.

2.8. ResNeXt

The ResNeXt [43] architecture extends the ResNet residual blocks with a split-transform-merge paradigm similar to the GoogLeNet inception module. Differently from the latter, that concatenates depth-wise feature maps generated using different kernel sizes, ResNeXt merges the different paths by summation. What is more, each path will employ identical kernels for its convolutions, effectively designing an architecture by repeating a building block with a specific set of transformations; resulting in a simple and homogeneous design and reducing the number of hyper-parameters to be set.

Hyperparameter	Value
Max Epochs	40
Optimizer	Adam
Learning Rate	1e-3
Batch size	100
Loss function	Categorical Cross-Entropy
Model Selection	High Accuracy

Table 2: Model training hyperparameter configuration.

2.9. ShuffleNet

The ShuffleNet model [44] is another network optimized for mobile devices that requires low computational resources. To lower the latter, generally measured in Mega Floating-point Operations per Second (MFLOPs), this model defines two operations that reduce the computational costs while still preserving accuracy, i.e., point-wise group convolution and channel shuffle. These two procedures rearrange feature maps channels to increase the architecture abstraction level, achieving high performances on mobile devices (e.g., 18x faster execution with respect to an AlexNet on ARM-based devices, for similar accuracy) even though the structure contains 50 layers organized into 4 stages that apply the aforementioned mechanisms.

2.10. SqueezeNet

Differently from other networks, SqueezeNet [45] focuses on significantly reducing the model size, resulting in a compact CNN with 18 layers that has similar performances to AlexNet on the ImageNet dataset [36], but with 50x less parameters. To obtain these results, SqueezeNet authors describe Fire layers implementing an alternating squeeze-expand strategy, where parameters are constrained by reducing the convolution filter size. In particular, 1×1 as well as 1×1 and 3×3 kernels are used in the squeeze and expand components, respectively. As a result, the parameter number reduction allows for more feature maps to be generated in deeper layers to enhance the model accuracy.

2.11. VGG16

One of the first networks to include a deeper architecture is the OxfordNet model, also known as VGG16 [46]. The Oxford Visual Geometry Group achieved remarkable performances by stacking up to 13 convolutional layers, interleaved by 5 max pooling operations applied at the second, fourth, seventh, tenth, and thirteenth convolutions to reduce the input size. Through this structure, meaningful features can be extracted from the input image and used to discriminate between the ImageNet 1000 classes. Moreover, the VGG16 is one of the first models applying smaller kernels in its convolutions, i.e., with a shape of 3×3, used to capture details in extremely small receptive fields.

2.12. Wide ResNet

A different approach to typical deep neural network design is explored by the Wide ResNet model [47]. In this architecture, ResNet residual blocks, which were otherwise devised to

lessen the number of parameters of deep neural networks by allowing for deep, thin and effective structures, are modified to reduce their depths and increase their width. In particular, this is achieved by limiting the amount of layers in residual blocks, while employing bigger kernels and generating more feature maps per convolution. A strategy that enabled the Wide ResNet with its 16-layer structure to outperform other deeper models on the ImageNet Large Scale Visual Recognition challenge [50].

3. Experimental Results and Discussion

Extensive experiments were carried out to assess the transfer learning capabilities of all networks. The dataset employed for evaluation and its merging procedure are described in Section 3.1. Implementation details and the chosen testing protocol are instead presented in Section 3.2. Finally, quantitative and qualitative experimental results are shown and discussed in Section 3.3.

3.1. Dataset

To explore transfer learning capabilities on pneumonia classification of chest-x-rays images, with particular attention to the SARS-CoV-2 infection, we performed experiments on a specifically designed dataset containing different public collections addressing the pneumonia classification task. In detail, the merged dataset contains 4 categories, namely: normal (i.e., healthy), and pneumonia deriving from bacteria, generic viruses or COVID-19, as depicted in Fig. 3.

Regarding the first three classes, they are taken from a well-known Kaggle pneumonia challenge, and split into training, development, and test sets in accordance with [32]. Observe that this dataset, which comprises more than 5000 images pre-COVID-19 era, is annotated as a binary classification task, i.e., images either containing pneumonia symptoms or not. Nevertheless, thanks to its metadata it is possible to further split the images according to the illness source, i.e., either bacterial or virus, therefore resulting in 3 of the aforementioned classes.

Concerning chest-x-rays images showing a COVID-19 contagion, we further extend the training and test sets with 185 images by following the same arrangement proposed in [13]. Moreover, since the latter does not have any validation data, we also expand the corresponding set through the remarkable works of [49, 48], which assembled ≈400 chest-x-rays COVID-19 samples from different publications. From this collection, 289 images were retained to ensure there were non-overlapping images with respect to sets defined by [13].

Summarizing, the final dataset, fully described in Table 1, comprises the bacteria, virus, and normal classes as per [32], while for COVID-19 it contains train and test sets from [13], and a validation set taken from [48, 49], for a total of 6330 chest-x-ray images.

3.2. Implementation Details

All experiments performed across the various architectures make use of a standard set of hyperparameters, summarized in Table 2, to have a comparable environment across the models.

Model	Prec.%	Sens.%	Spec.%	F1%	Params
AlexNet	79.23	73.48	91.16	73.14	58M
DenseNet	84.70	83.01	94.34	84.00	27M
GoogleNet	83.09	80.52	93.51	81.33	5M
MnasNet	69.50	56.49	85.50	54.60	3M
MobileNet v2	83.78	81.91	93.97	82.48	2M
MobileNet v3	**84.92**	**83.43**	**94.48**	**84.36**	3M
ResNet50	80.58	76.38	92.13	77.49	24M
ResNext	79.55	77.49	92.50	78.50	23M
ShuffleNet	83.19	78.04	92.68	79.33	1M
SqueezeNet	20.73	27.49	75.83	19.70	0.7M
VGG16	77.39	74.72	91.57	75.35	135M
Wide ResNet50	79.83	74.45	91.48	76.23	67M

Table 3: Performance comparison computed on the test set for precision, sensitivity, specificity, and F1-score metrics.

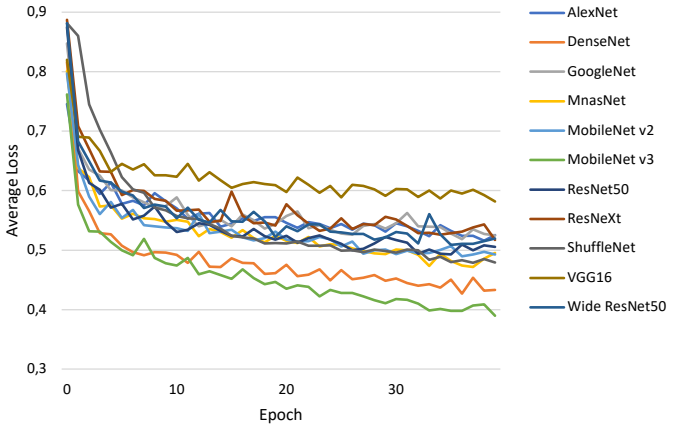


Figure 4: Average loss evolution during training time.

In detail, each network was trained for up to 40 epochs using the Adam optimizer, a learning rate of 1e-3, and a batch size of 100, which allowed for a good sample mixture. Moreover, for each model, weights associated to the highest performing epoch, with respect to the development set accuracy, were selected as a final configuration to be used for inference. Furthermore, several common classification metrics, i.e., precision, sensitivity, specificity, and F1-score, were used to evaluate the networks performances. Notice that, being in a multiclass classification scenario, a weighted average of the aforementioned metrics was used to estimate the models performances so that the dataset distribution would also be taken into account.

Concerning the architectures implementation, we employed the PyTorch framework and PyTorch Lightning library, with a 16-bit floating-point precision to speed up the computation. Furthermore, ImageNet pre-trained models were imported from the TorchVision library to exploit the transfer learning paradigm. Finally, all tests were performed on the Google Colab infrastructure with an Intel CPU x86-64 architecture using 25GB of RAM, along an NVIDIA Tesla V100 with 16GB of VRAM.

3.3. Performance Evaluation

In this section, we evaluate all of the presented models both quantitatively and qualitatively by means of sheer performances,

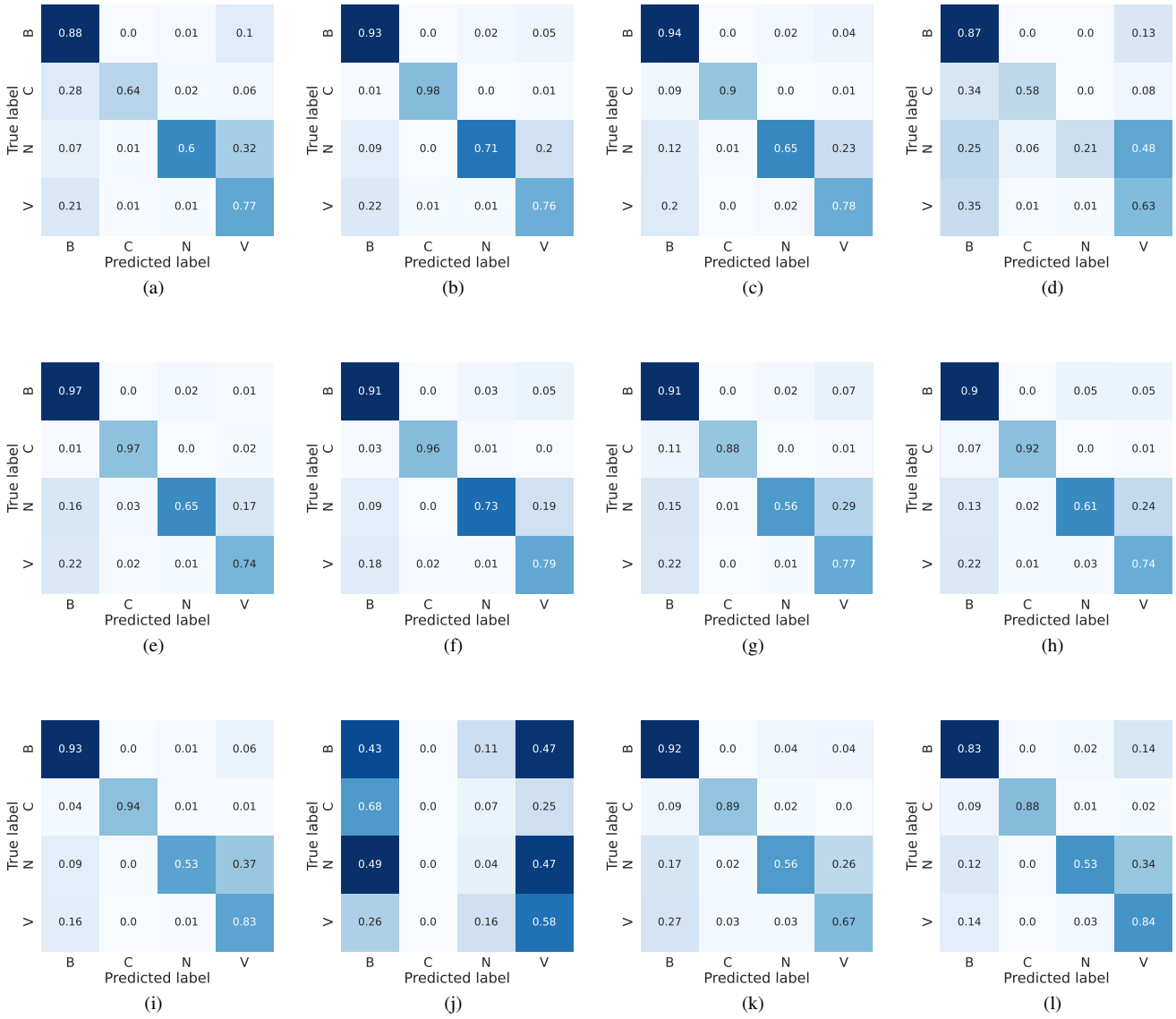
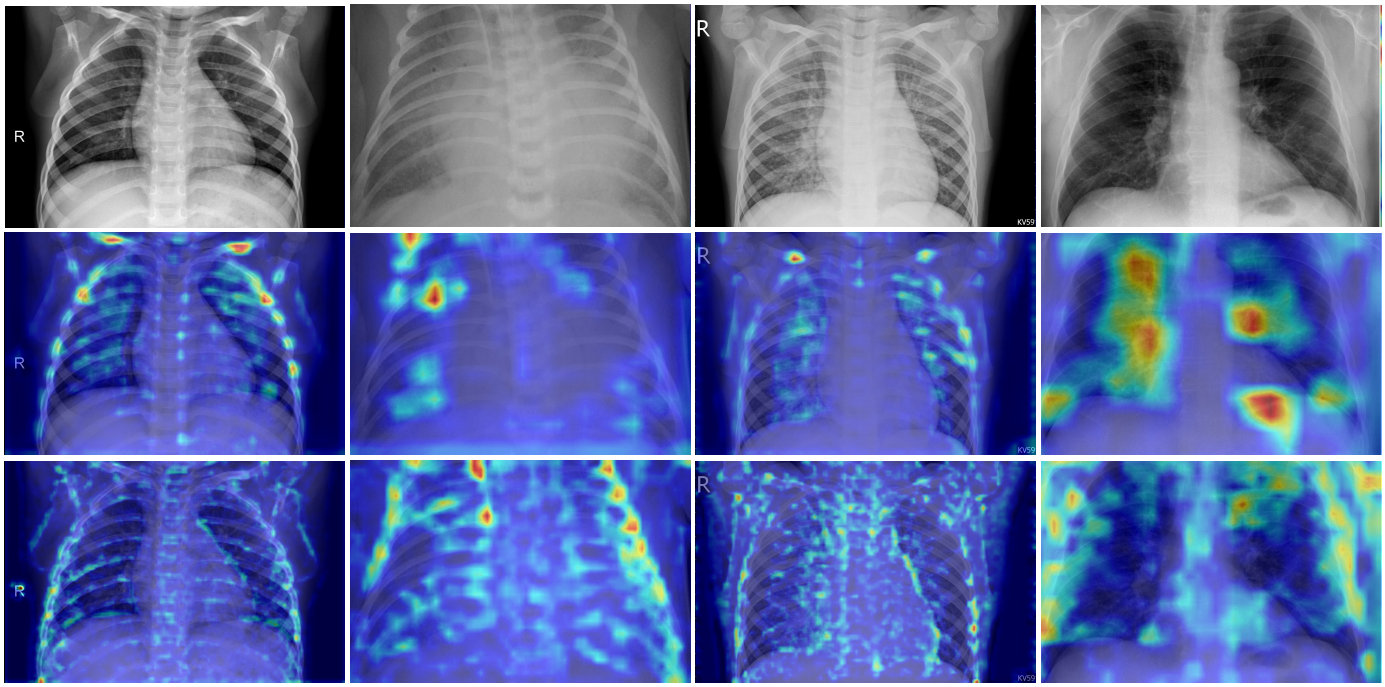


Figure 5: Confusion matrices for AlexNet (a), DenseNet (b), GoogleNet (c), MnasNet (d), MobileNet v2 (e), MobileNet v3 (Large) (f), ResNet50 (g), ResNext (h), ShuffleNet (i), SqueezeNet (j), VGG16 (k) and Wide ResNet50 (l), computed on the test set. Labels N, B, V, C correspond to normal, bacteria, virus and COVID-19 classes, respectively.

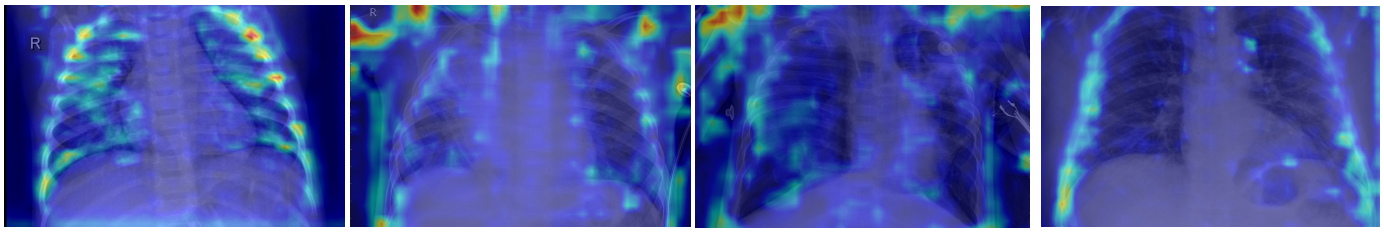
for the former, and through an explainable visualization of the output, for the latter.

Concerning the quantitative assessment, relevant classification metrics are reported in Table 3. As shown, almost all networks can achieve satisfactory results with a score higher than 70% across all metrics, indicating that transfer learning can be an important tool for the early screening of new diseases. Moreover, all models achieve high specificity scores, implying that, for each class, there generally is a low number of false positives, i.e., pneumonia sources tend not to be mixed during predictions. More interestingly, the top performing architecture is the MobileNet v3, even though it is one of the models with the lowest number of parameters since it is optimized for mobile devices. This outcome highlights that more recent and advanced architectures can have an edge over older ones (e.g., AlexNet) and, as a consequence, may perform better using significantly fewer parameters without loss of generality. While networks with less parameters can still obtain remarkable performances, the opposite result can also be observed. Examples of this behavior are the MnasNet and, in particular, SqueezeNet, which do not perform well during inference as they cannot effectively transfer previous knowledge to the new task due to both their architecture and parameters number; an expected outcome since these networks are mainly focused on significantly reducing pre-existing models sizes and can have a lower accuracy as a trade-off. Furthermore, the reduced performances are also confirmed by the average losses computed at training time, which are reported in Fig. 4. As can be observed, most net-



(a) (b) (c) (d)

Figure 6: Chest-x-ray inference samples showing a healthy patient (a), pneumonia from a bacteria infection (b), pneumonia from a virus infection (c), and pneumonia from a COVID-19 case (d). Input images are reported in the first row, while Grad-CAM overlaid images for MobileNet v3 and SqueezeNet are reported in the second and third row, respectively.



(a) (b) (c) (d)

Figure 7: Chest-x-ray misclassified samples showing a healthy patient (a), pneumonia from a bacteria infection (b), pneumonia from a virus infection (c), and pneumonia from a COVID-19 case (d), overlaid with Grad-CAM algorithm.

works start converging well before the 40-th epoch, apart from the best performing DenseNet and MobileNet v3 that kept improving across the entire training procedure. This implies that previous knowledge, although useful, can saturate fast during its transferal. What is more, the SqueezeNet architecture actually diverges when adapting its classifier to the new task, thus resulting in loss values outside the reported range, and further justifying the low performance obtained on the pneumonia classification task. This outcome suggests that using an extremely low number of parameters can result in issues when exploiting the transfer learning paradigm, especially in the case of complex tasks such as the classification of chest-x-ray images. To conclude this quantitative evaluation and better appreciate the capabilities of knowledge transfer, confusion matrices for all models are shown in Fig. 5. As can be observed, excluding the diverged SqueezeNet architecture, i.e., Fig. 5.(j), predictions from the various models concentrate, as expected, on the corresponding matrix diagonal, indicating the right recognition of the four classes. On a different note, most of the errors are associated to the misclassification of the normal healthy class into the viral one. A behavior that can most likely be associated to artifacts present in the chest-x-ray, which might be confused for the diffused aspect of mild pneumonia cases from viral sources, in accordance with the findings of [2]. Nevertheless, there is a robust COVID-19 recognition across the architectures, even though viruses, and especially other SARS pathogens, fall into a different category. This outcome highlights the effectiveness of transfer learning, and indicates that, for early diagnosis of new illnesses, it can be a practical strategy thanks to the pre-existing models prowess.

Regarding the qualitative evaluation, Grad-CAM overlaid images for the best and worst performing models, i.e., the Mo-

bileNet v3 and SqueezeNet architectures, are shown in Fig. 6. As can be observed, the former model, reported on the second row, concentrates on the lungs for pneumonia cases deriving from bacteria, virus, and COVID-19, i.e., Fig. 6.(b-d); while the architecture internal weights have a higher response, i.e., red areas on the Grad-CAM image, on both collar as well as rib cage bones for a normal chest-x-ray, i.e., Fig. 6.(a). This outcome is most likely due to the pneumonia absence from the input, which resulted in the model focusing on other points of interest to learn the right category. Even more interesting, the activated areas in pneumonia images seem to confirm the illness propagation described in [2, 3]. Specifically, bacterial pathogens present a concentrated activation on specific spots, while viral sources present a diffused activation throughout the lungs. This effect is especially visible on COVID-19 pneumonia, i.e., Fig. 6.(d), where bigger lungs areas have a higher response during inference, indicating that the model could also learn differences between viral and SARS-CoV-2 pathogens. Differently from the MobileNet v3, the SqueezeNet architecture tends to focus on bones and edges from all chest components by focusing on smaller details, most likely due to the smaller kernel filters of the network. Consequently, this model mixes normal chest-x-rays with either a bacterial or other generic viral pathogens, as expected from the diverged confusion matrix presented in Fig.5.(j). A similar issue for this second network, albeit to a lesser extent, is also presented when examining activated areas in the bacterial and COVID-19 images, i.e., Fig. 6.(b) and Fig. 6.(d). In both cases, the model responds to wider regions of the chest-x-ray in conjunction with rib cage bones, suggesting that the architecture is not analyzing pneumonia effects on lungs and resulting in the sensibly lower reported performances as well as training divergence. What is more, a comparable scenario can also be observed in Fig. 7, where misclassified images for the MobileNet v3, i.e., the best architecture on the presented dataset, are reported. As shown, for each category, the model has a strong response on other points of interest, such as collar or rib cage bones, as well as external lungs locations. Furthermore, for the healthy chest-x-ray, i.e., Fig. 7.(a), there are active areas within the lung in correspondence with image artifacts. Such behaviors indicate that, as mentioned, the architecture fine-tuning could be further improved, especially since it did not reach convergence on the training dataset. Moreover, it highlights the complexity of the task since even proficient networks have difficulties handling these information-rich images. Regardless of these issues, qualitative experiments corroborate the quantitative effectiveness of transfer learning to diagnose pneumonia sources.

4. Conclusion

In this paper, we presented a benchmark with 12 renowned deep neural network architectures modified to exploit the transfer learning paradigm. In particular, a dataset was organized from several public collections of chest-x-rays for the pneumonia classification derived by bacteria, generic viral or SARS-CoV-2 pathogens. Moreover, quantitative and qualitative experiments were reported to assess all models, which managed to reach up to 84.46% average f1-score at inference time. As shown by the results, one of the smallest architectures, i.e., MobileNet v3, achieves the best performances and, when applying the Grad-CAM algorithm, it exhibits similar patterns with respect to the corresponding pneumonia source, i.e., high internal weights responses in concentrated or diffused areas for bacterial or viral and SARS-CoV-2 sources, respectively; demonstrating that previous knowledge does indeed help to address a new task, and suggesting that transfer learning can become a fundamental tool when diagnosing future unknown illnesses.

As future work, an even larger collection will be organized to account for more variegate pulmonary diseases by merging other relevant public collections. In addition, further inquiries will be made by also evaluating other models as well as preprocessing techniques that could improve a given architecture performances by mitigating misclassification through the reduction, for instance, of artifacts in the input.

Acknowledgments

This work was supported by the MIUR under grant "Departments of Excellence 2018–2022" of the Sapienza University Computer Science Department and the ERC Starting Grant no. 802554 (SPECGEO).

References

[1] S. B. Smith, G. W. Ruhnke, C. H. Weiss, G. W. Waterer, R. G. Wunderink, Trends in pathogens among patients hospitalized for pneumonia from 1993 to 2011, JAMA Internal Medicine 174 (11) (2014) 1837–1839. doi:10.1001/jamainternmed.2014.4344.

[2] S. N. Shah, R. G. Bachur, D. L. Simel, M. I. Neuman, Does this child have pneumonia?: the rational clinical examination systematic review, Jama 318 (5) (2017) 462–471. doi:10.1001/jama.2017.9039.

[3] J. Reynolds, G. McDonald, H. Alton, S. Gordon, Pneumonia in the immunocompetent patient, The British Journal of Radiology 83 (996) (2010) 998–1009. doi:10.1259/bjr/31200593.

[4] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in wuhan, china, The Lancet 395 (10223) (2020) 497–506. doi:10.1016/S0140-6736(20)30183-5.

[5] S. Bhattacharya, P. K. R. Maddikunta, Q.-V. Pham, T. R. Gadekallu, C. L. Chowdhary, M. Alazab, M. J. Piran, et al., Deep learning and medical image processing for coronavirus (covid-19) pandemic: A survey, Sustainable Cities and Society 65 (2021) 102589. doi:10.1016/j.scs.2020.102589.

[6] N. Phillips, The coronavirus is here to stay-here's what that means, Nature 590 (7846) (2021) 382–384. doi:10.1038/d41586-021-00396-2.

[7] A. Fontanet, B. Autran, B. Lina, M. P. Kieny, S. S. A. Karim, D. Sridhar, Sars-cov-2 variants and ending the covid-19 pandemic, The Lancet 397 (10278) (2021) 952–954. doi:10.1016/S0140-6736(21)00370-6.

[8] S. Michie, R. West, Sustained behavior change is key to preventing and tackling future pandemics, Nature Medicine 27 (5) (2021) 749–752. doi:10.1038/s41591-021-01345-2.

[9] W. William, A. Ware, A. H. Basaza-Ejiri, J. Obungoloch, A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images, Computer Methods and Programs in Biomedicine 164 (2018) 15–22. doi:10.1016/j.cmpb.2018.05.034.

[10] R. Karthik, R. Menaka, A. Johnson, S. Anand, Neuroimaging and deep learning for brain stroke detection-a review of recent advancements and future prospects, Computer Methods and Programs in Biomedicine 187 (2020) 105728. doi:10.1016/j.cmpb.2020.105728.

[11] G. Martí-Juan, G. Sanroma-Guell, G. Piella, A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in alzheimer's disease, Computer Methods and Programs in Biomedicine 189 (2020) 105348. doi:10.1016/j.cmpb.2020.105348.

[12] D. Avola, L. Cinque, A. Fagioli, G. Foresti, A. Mecca, Ultrasound medical imaging techniques: A survey, ACM Computing Surveys (CSUR) 54 (3) (2021) 1–38. doi:10.1145/3447243.

[13] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, G. J. Soufi, Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning, Medical Image Analysis 65 (2020) 101794. doi:10.1016/j.media.2020.101794.

[14] L. Brunese, F. Mercaldo, A. Reginelli, A. Santone, Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays, Computer Methods and Programs in Biomedicine 196 (2020) 105608. doi:10.1016/j.cmpb.2020.105608.

[15] W. M. Shaban, A. H. Rabie, A. I. Saleh, M. Abo-Elsoud, Detecting covid-19 patients based on fuzzy inference engine and deep neural network, Applied Soft Computing 99 (2021) 106906. doi:10.1016/j.asoc.2020.106906.

[16] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. L. Suárez, J. Luengo, M. Valero-González, et al., Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images, IEEE Journal of Biomedical and Health Informatics 24 (12) (2020) 3595–3605. doi:10.1109/JBHI.2020.3037127.

[17] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, et al., Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection, IEEE Transactions on Medical Imaging 40 (3) (2020) 879–890. doi:10.1109/TMI.2020.3040950.

[18] A. Shamsi, H. Asgharnezhad, S. S. Jokandan, A. Khosravi, P. M. Kebria, D. Nahavandi, S. Nahavandi, D. Srinivasan, An uncertainty-aware transfer learning-based framework for covid-19 diagnosis, IEEE Transactions on Neural Networks and Learning Systems 32 (4) (2021) 1408–1417. doi:10.1109/TNNLS.2021.3054306.

[19] F. Demir, Deepcoronet: A deep lstm approach for automated detection of covid-19 cases from chest x-ray images, Applied Soft Computing 103 (2021) 107160. doi:10.1016/j.asoc.2021.107160.

[20] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, U. R. Acharya, Automated detection of covid-19 cases using deep neural networks with x-ray images, Computers in biology and medicine 121 (2020) 103792. doi:10.1016/j.compbiomed.2020.103792.

[21] A. Gupta, S. Gupta, R. Katarya, et al., Instacovnet-19: A deep learning classification model for the detection of covid-19 patients using chest x-ray, Applied Soft Computing 99 (2021) 106859. doi:10.1016/j.asoc.2020.106859.

[22] S. Tang, C. Wang, J. Nie, N. Kumar, Y. Zhang, Z. Xiong, A. Barnawi, Edl-covid: Ensemble deep learning for covid-19 cases detection from chest x-ray images, IEEE Transactions on Industrial Informatics 17 (9) (2021) 6539–6549. doi:10.1109/TII.2021.3057683.

[23] A. Saleh, A. Nasir, S. Muhammad, Coronavirus disease (covid-19) detection using x-ray images and enhanced densenet, Applied Soft Computing preprint (2021) 107645. doi:10.1016/j.asoc.2021.107645.

[24] M. Nour, Z. Cömert, K. Polat, A novel medical diagnosis model for covid-19 infection detection based on deep features and bayesian optimization, Applied Soft Computing 97 (2020) 106580. doi:10.1016/j.asoc.2020.106580.

[25] A. I. Khan, J. L. Shah, M. M. Bhat, Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images, Computer Methods and Programs in Biomedicine 196 (2020) 105581. doi:10.1016/j.cmpb.2020.105581.

[26] Y. Oh, S. Park, J. C. Ye, Deep learning covid-19 features on cxr using limited training data sets, IEEE Transactions on Medical Imaging 39 (8) (2020) 2688–2700. doi:10.1109/TMI.2020.2993291.

[27] J. Li, Y. Wang, S. Wang, J. Wang, J. Liu, Q. Jin, L. Sun, Multiscale attention guided network for covid-19 diagnosis using chest x-ray images, IEEE Journal of Biomedical and Health Informatics 25 (5) (2021) 1336–1346. doi:10.1109/JBHI.2021.3058293.

[28] F. Ahmad, M. U. G. Khan, K. Javed, Deep learning model for distinguishing novel coronavirus from other chest related infections in x-ray images, Computers in Biology and Medicine 134 (2021) 104401. doi:10.1016/j.compbiomed.2021.104401.

[29] T. Mahmud, M. A. Rahman, S. A. Fattah, Covxnet: A multi-dilation convolutional neural network for automatic covid-19 and other pneumonia detection from chest x-ray images with transferable multi-receptive feature optimization, Computers in Biology and Medicine 122 (2020) 103869. doi:10.1016/j.compbiomed.2020.103869.

[30] R. Karthik, R. Menaka, M. Hariharan, Learning distinctive filters for covid-19 detection from chest x-ray using shuffled residual cnn, Applied Soft Computing 99 (2021) 106744. doi:10.1016/j.asoc.2020.106744.

[31] G. Marques, D. Agarwal, I. de la Torre Díez, Automated medical diagnosis of covid-19 through efficientnet convolutional neural network, Applied Soft Computing 96 (2020) 106691. doi:10.1016/j.asoc.2020.106691.

[32] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131. doi:10.1016/j.cell.2018.02.010.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1007/s11263-019-01228-7.

[34] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2009) 1345–1359. doi:10.1109/TKDE.2009.191.

[35] V. Cheplygina, M. de Bruijne, J. P. Pluim, Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Medical Image Analysis 54 (2019) 280–296. doi:10.1016/j.media.2019.03.009.

[36] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25 (2012) 1097–1105. doi:https://doi.org/10.1145/3065386.

[37] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708. doi:10.1109/CVPR.2017.243.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.

[39] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, Q. V. Le, Mnasnet: Platform-aware neural architecture search for mobile, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2820–2828. doi:10.1109/CVPR.2019.00293.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520. doi:10.1109/CVPR.2018.00474.

[41] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314–1324. doi:10.1109/ICCV.2019.00140.

[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[43] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1492–1500. doi:10.1109/CVPR.2017.634.

[44] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6848–6856. doi:10.1109/CVPR.2018.00716.

[45] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer, Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size, arXiv preprint arXiv:1602.07360 preprint

(2016) 1–13.

[46] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 preprint (2014) 1–14.

[47] S. Zagoruyko, N. Komodakis, Wide residual networks, arXiv preprint arXiv:1605.07146 preprint (2016) 1–15.

[48] J. P. Cohen, P. Morrison, L. Dao, Covid-19 image data collection (2020). URL https://github.com/ieee8023/covid-chestxray-dataset

[49] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, M. Ghassemi, Covid-19 image data collection: Prospective predictions are the future (2020).
URL https://github.com/ieee8023/covid-chestxray-dataset

[50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252. doi:10.1007/s11263-015-0816-y.

[51] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 preprint (2017) 1–9.

[52] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141. doi:10.1109/CVPR.2018.00745.

[53] C. Thornton, F. Hutter, H. H. Hoos, K. Leyton-Brown, Auto-weka: Combined selection and hyperparameter optimization of classification algorithms, in: Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2013, pp. 847–855. doi:10.1145/2487575.2487629.