Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/18/20230980

The transferability of transfer learning model based on ImageNet for medical image classification tasks

Zhanhao Zhang

The Department of Control Science and Engineering, Shandong university, Jinan, 250001, China

202000172010@mail.sdu.edu.cn

Abstract. Transfer learning with pretrained weights is commonly based on the ImageNet dataset. However, ImageNet does not contain medical images, leaving the transferability of these pretrained weights for medical image classification an open question. The core purpose of this study is to investigate the impact of transfer learning on the accuracy of medical image classification, utilizing ResNet18, VGG11, AlexNet, and MobileNet, which are four of the most widely used neural network models. Specifically, this study aims to determine whether the incorporation of transfer learning techniques leads to significant improvements in the performance of image classification tasks, as compared to traditional methods that do not utilize transfer learning. The dataset consists of approximately 4,000 chest X-ray images with labels of healthy, COVID, or Viral Pneumonia. The final layer's output neurons of the network's architecture were revised to three to accommodate the ternary classification task. Preprocessing techniques include downsampling and normalization of the pixel values. By maintaining the same dataset and preprocessing methods, this study compares the performance of the models with and without pretrained weights. The results demonstrate that, compared to not using transfer learning, all four network models converge more quickly and achieve higher validation accuracy in the initial epochs when transfer learning is employed. Furthermore, the models exhibit higher prediction accuracy in the final test set. This study suggests that using transfer learning with pretrained weights based on ImageNet can boost the efficiency of medical image classification tasks.

Keywords: transfer learning, ImageNet pretrained weights, medical image classification.

1. Introduction

Pneumonia manifests as a pathological state involving inflammation of alveoli, the tiny air sacs within the lungs. Generally, bacterial or viral infections lead to this condition, which can sometimes necessitate hospitalization and even result in death. The COVID-19 pandemic, a novel coronavirus disease originating from the SARS-CoV-2 virus, has impacted millions of individuals across the globe. Presently, COVID-19 pneumonia diagnosis predominantly depends on Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests and radiological imaging, including chest X-rays and Computed Tomography (CT) scans [1]. Nonetheless, RT-PCR tests can be laborious, time-intensive, and produce false-negative outcomes [2]. Additionally, CT scans have limitations, such as delayed results, substantial labor expenses, and occasional misdiagnoses [3]. Given the constraints of the present

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/18/20230980

detection approaches, it is imperative to devise a more efficient and precise method for detection in an urgent manner, and Artificial Intelligence (AI) methods might offer a potential solution.

AI, particularly deep learning methodologies, have shown considerable potential in numerous medical imaging tasks, including pneumonia diagnosis [4]. Deep learning algorithms like Convolutional Neural Networks (CNNs) possesses the capability to significantly augment the pace and accuracy of pneumonia detection through automated medical image analysis. By incorporating appropriate preprocessing techniques and classifiers, neural networks can exhibit improved disease classification performance [5]. This, in turn, could reduce the burden on healthcare professionals and minimize misdiagnoses.

Deep learning has undergone rapid development in recent years, with notable advancements in computer vision domains. Among various deep learning architectures, CNNs have been particularly successful in image classification tasks due to their ability to capture hierarchical patterns in images [6]. Deep learning has been applied to address a broad spectrum of issues in medical imaging, such as tumor detection, organ segmentation, and disease diagnosis [7]. The adoption of transfer learning, which involves utilizing pre-trained weights from extensive datasets like ImageNet, has been proven to boost the accuracy of deep learning models in medical imaging tasks [8]. However, the suitability of using ImageNet pre-trained weights for medical image classification remains debatable, as the dataset lacks medical images. While several studies have reported favorable outcomes using transfer learning with ImageNet pre-trained weights for medical image classification, the fundamental suitability of this approach is still uncertain. In a study conducted by Basil Mustafa et al., the authors employed transfer learning with ImageNet pre-trained weights to create a deep learning model for mammography and dermatology detection [9]. The model achieved remarkable accuracy and surpassed multiple baseline models, showcasing the potential advantages of transfer learning in medical imaging tasks. Nonetheless, the absence of medical images in the ImageNet dataset raises concerns about the generalizability of this approach to other medical image classification tasks, such as pneumonia detection in chest X-rays.

To address the aforementioned concerns and evaluate the effectiveness of using ImageNet pre-trained weights for medical image classification such as chest x-rays, this study aims to compare the performance of classical deep learning models (i.e., AlexNet, MobileNetV3, VGG11, and ResNet18) with and without the application of ImageNet pre-trained weights in the context of pneumonia detection from chest X-rays. This study extracted X-ray images from three distinct categories: normal, COVID-19, and viral pneumonia cases from the COVID-19 Radiography Database on the Kaggle platform [10]. Each category was trained using the aforementioned neural network models. The experimental results demostrate that using pre-trained weights from ImageNet in transfer learning can enhance the network's performance in chest X-ray image classification tasks. Specifically, training with ResNet18 using pre-trained weights resulted in an increase in the classification accuracy of the test set from 93% to 96%. Similar improvements in accuracy were observed when using pre-trained weights in three other network structures, all of which resulted in a 3-4% increase in accuracy and achieved high accuracy in the first few epochs of training. Moreover, from the confusion matrix and ROC curve, it can be observed that using pre-trained weights improves the prediction accuracy of each class, with the Area Under the Curve (AUC) approaching a value of 1.

2. Method

2.1. Dataset description and preprocessing

In this study, dataset was obtained from the COVID-19 Chest X-ray Images and Lung Masks Database on Kaggle [10]. It comprises chest X-ray images classified into three distinct categories: COVID-19, NORMAL, and Viral Pneumonia. Although additional information is available within this database, this study focused solely on the aforementioned categories for the purposes of conducting a multi-class classification task. To ensure a balanced representation of each category, the dataset was modified by randomly discarding images from the larger categories until an equal number of images (i.e., 1345) were

DOI: 10.54254/2755-2721/18/20230980

present in each category. The images have dimensions of 299×299 and are in RGB format. A sample of these images is demonstrated in Figure 1.

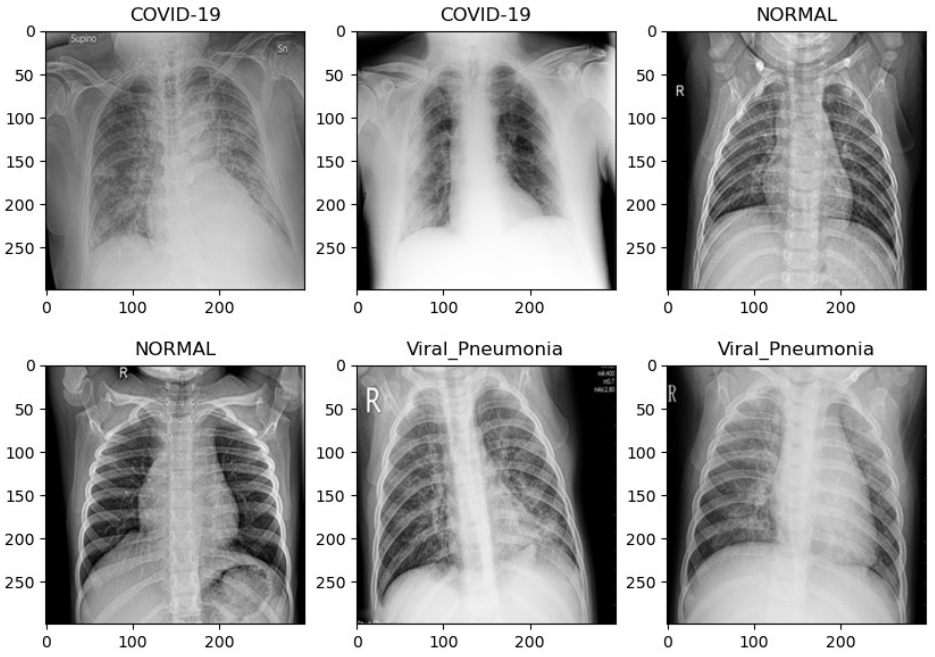


Figure 1. A selection of images from the dataset.

Prior to their utilization within the models, the input underwent a preprocessing stage. First, the image dimensions were resized to 128x128 pixels as well as the conversion of the images into tensors for training purposes. Additionally, normalization was employed to improve the contrast and distinctness of the features in the images. This involved setting the mean value of each channel to 0.5144 and the standard deviation to 0.2258, as part of the normalization process. Figure 2 shows some visualizations of the image after data normalization.

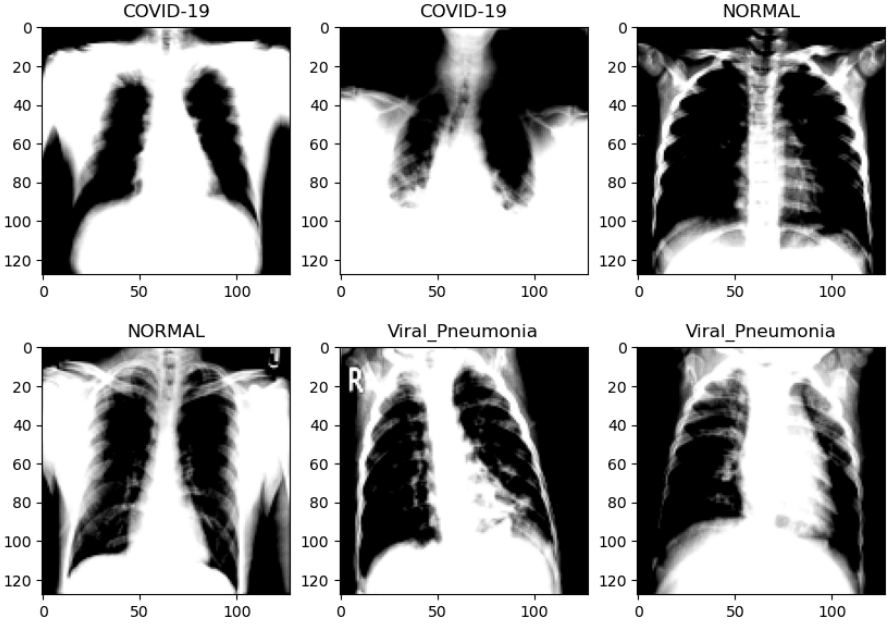


Figure 2. The preprocessed sample images of the collected dataset.

DOI: 10.54254/2755-2721/18/20230980

2.2. CNN models

The Convolutional Neural Network (CNN) has a powerful network architecture, which has made it widely used in the field of image classification since its proposal [11]. The number of layers in a CNN can be flexibly increased or decreased. In particular, convolutional layers utilize convolutional kernels of different sizes to extract features, pooling layers serve to blur the image and reduce computational complexity, and fully connected layers connect the preceding network layers to predict outcomes. In this study, four classic CNN architectures: AlexNet, MobileNetV3, VGG11, and ResNet18 were adopted.

AlexNet, introduced by Krizhevsky et al. [12], is a CNN architecture that has been demonstrated to exhibit excellent performance and high efficiency in image recognition tasks. Five convolutional layers are used as the front-end followed by three fully connected layers as the back end in its architecture, and it utilizes the ReLU activation function. To reduce spatial dimensions, the model also incorporates max-pooling layers between certain convolutional layers.

MobileNetV3, proposed by Howard et al. [13], is designed to be an efficient architecture suitable for deployment on mobile and embedded devices. To achieve this, the network incorporates depthwise separable convolutions and efficient network blocks, which reduce both computation and memory requirements. MobileNetV3 incorporates neural architecture search techniques and hardware-aware optimizations for resource-constrained devices.

VGG11, presented by Simonyan and Zisserman [14], is a variation of the VGG architecture with 11 weight layers. VGG networks use small (3x3) convolutional filters and multiple stacked convolutional layers for efficient feature extraction. Their straightforward design makes them popular for various computer vision tasks.

ResNet18, introduced by Him et al. [15], is a residual learning-based network with 18 layers. ResNet architectures use residual connections to learn residual functions and alleviate the vanishing gradient problem. These connections allow for accurate gradient flow during backpropagation, enabling the training of deeper models and successfully surpassing previous benchmarks and achieving superior performance in a wide range of tasks.

In this study, the CNN models were implemented using the torchvision.models library in PyTorch, which provided pre-built versions of AlexNet, MobileNetV3, VGG11, and ResNet18. The pretrained parameter was modified to control whether the models used the pre-trained ImageNet weights or were trained from scratch. By default, the output layer of these models comprised of 1000 neurons, corresponding to the 1000 classes within the ImageNet dataset. Nonetheless, given the three-class classification focus of the present study, the output layer was altered to comprise of three neurons to appropriately reflect the target classification problem.

2.3. Implementation details

Based on a split ratio of 6:2:2, the dataset was partitioned into training, validation, and test sets. Cross-Entropy loss was the chosen loss function, which is suitable for classification problems with multiple classes. An Adam optimizer was utilized during training, with a learning rate of 0.0001, as it has been shown to work well in various deep learning tasks. The batch size was 16, which balances the trade-off between computational efficiency and convergence speed. The training process for each network lasted for 20 epochs and the performance metrics used to evaluate the models included training and validation loss curves for identifying overfitting or underfitting, accuracy on test images to assess correct classification proportion, confusion matrix to summarize correct and incorrect classifications, Receiver Operating Characteristic Curve (ROC) curves to depict the precision at different thresholds, and Area Under the Curve (AUC) as a single value indicating overall classifier performance.

3. Result and discussion

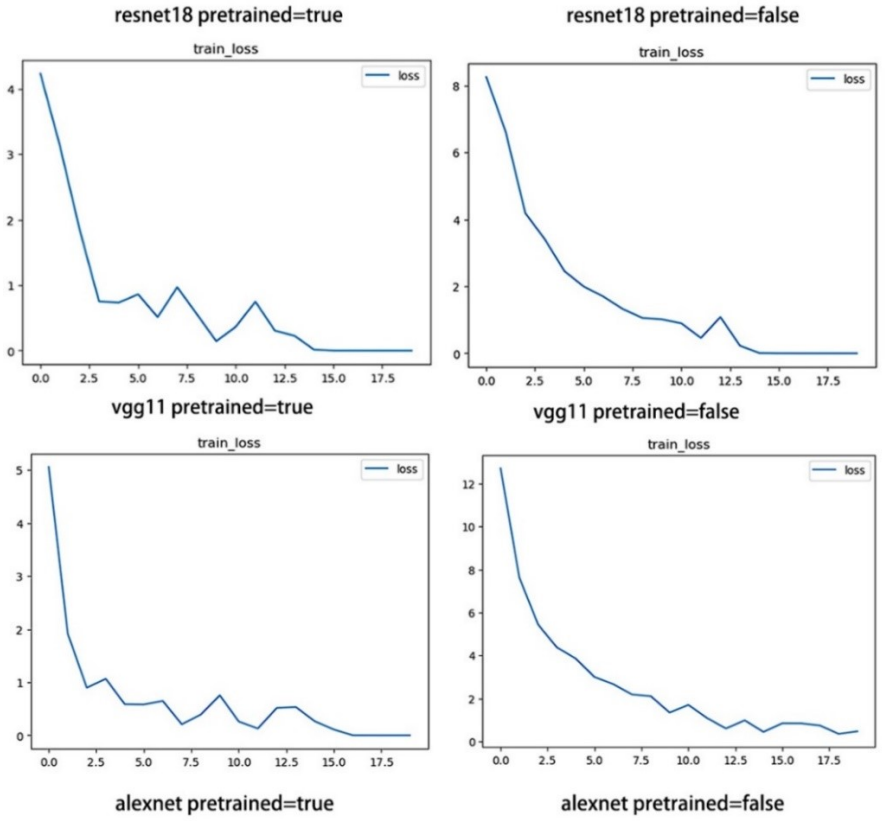
The study constructs four sets of distinct network models, and within each network, a comparison is conducted between using weights obtained from utilizing pretrained weights from ImageNet and not using them According to experiment results, the VGG11 model performs best when given consistent

DOI: 10.54254/2755-2721/18/20230980

training datasets and training settings, as shown in Table 1. The accuracy achieved with pretrained weights is 97.2%, while without pretrained weights, the accuracy is still the highest among the four networks at 94.4%. In contrast, MobileNet performs the worst, with accuracies of 95.5% with pretrained weights and 91.2% without pretrained weights.

Table 1. The performance of various models with/without pre-trained weights.

Performance	Model							
	Pretrained=false				Pretrained=true			
	Resnet18	Vgg11	Alexnet	MobileNet	Resnet18	Vgg11	Alexnet	MobileNet
Training loss of the first epoch	8.247	12.714	13.879	13.931	4.228	5.056	5.969	8.590
Validation accuracy of the first epoch	0.865	0.732	0.749	0.706	0.926	0.933	0.906	0.711
Training loss	0.001	0.345	1.123	1.452	0.016	0.110	0.205	0.353
Validation loss	6.688	4.634	5.239	7.418	2.070	2.570	5.0975	2.586
Highest validation accuracy	0.927	0.939	0.915	0.901	0.970	0.968	0.9294	0.964
Testing accuracy	0.923	0.944	0.933	0.912	0.971	0.972	0.963	0.955



DOI: 10.54254/2755-2721/18/20230980

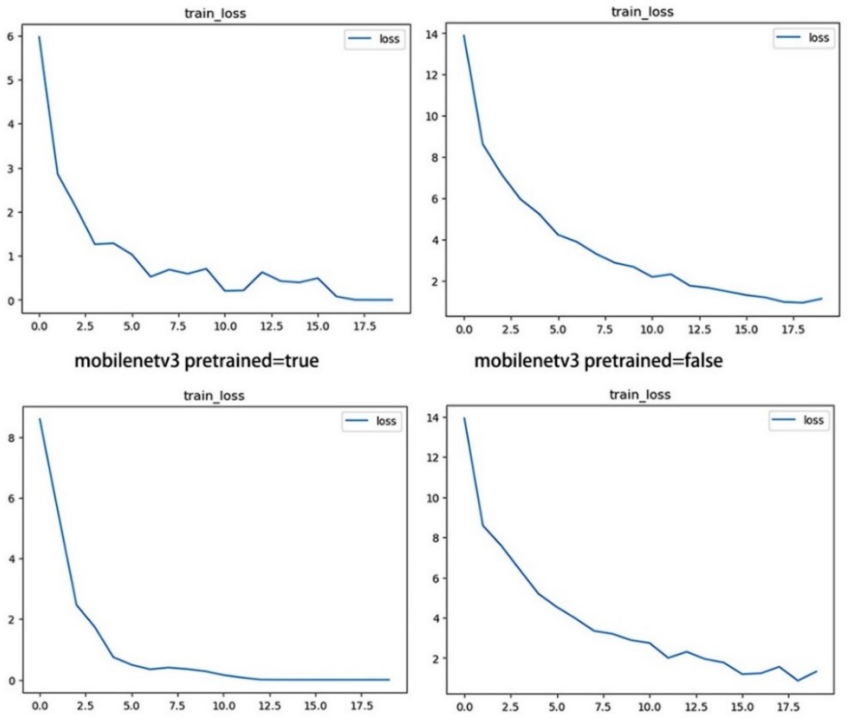
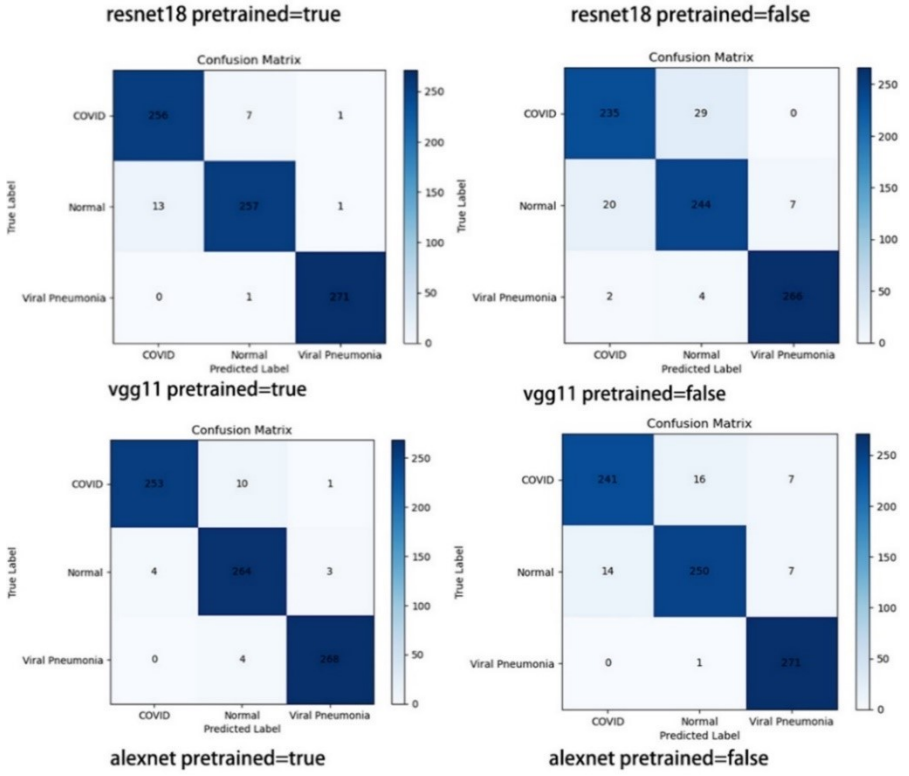


Figure 3. The training loss curve for different models, comparing those with and without pre-trained weights.



DOI: 10.54254/2755-2721/18/20230980

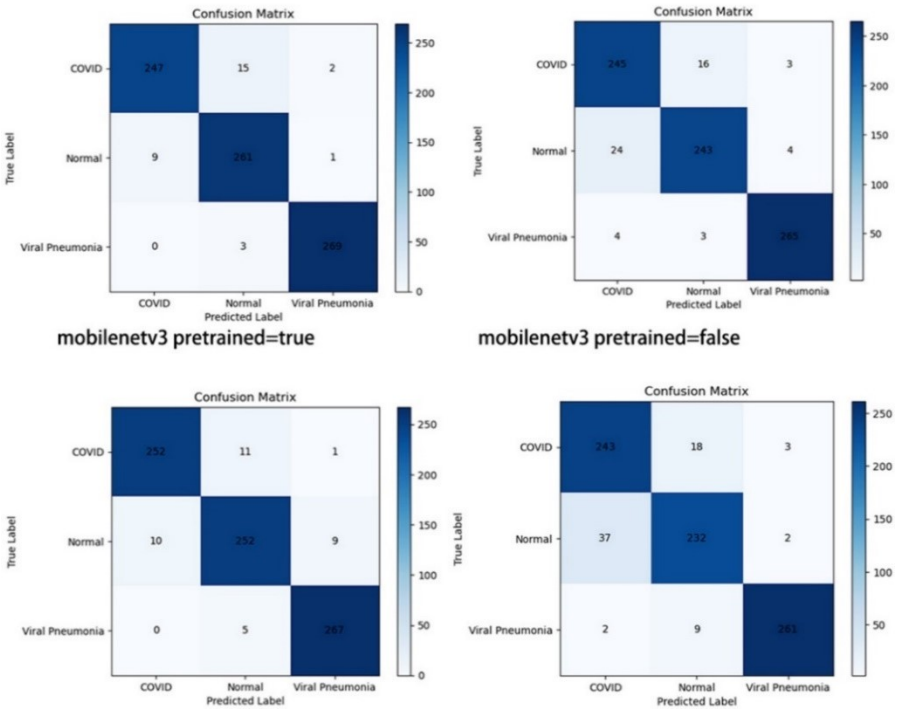
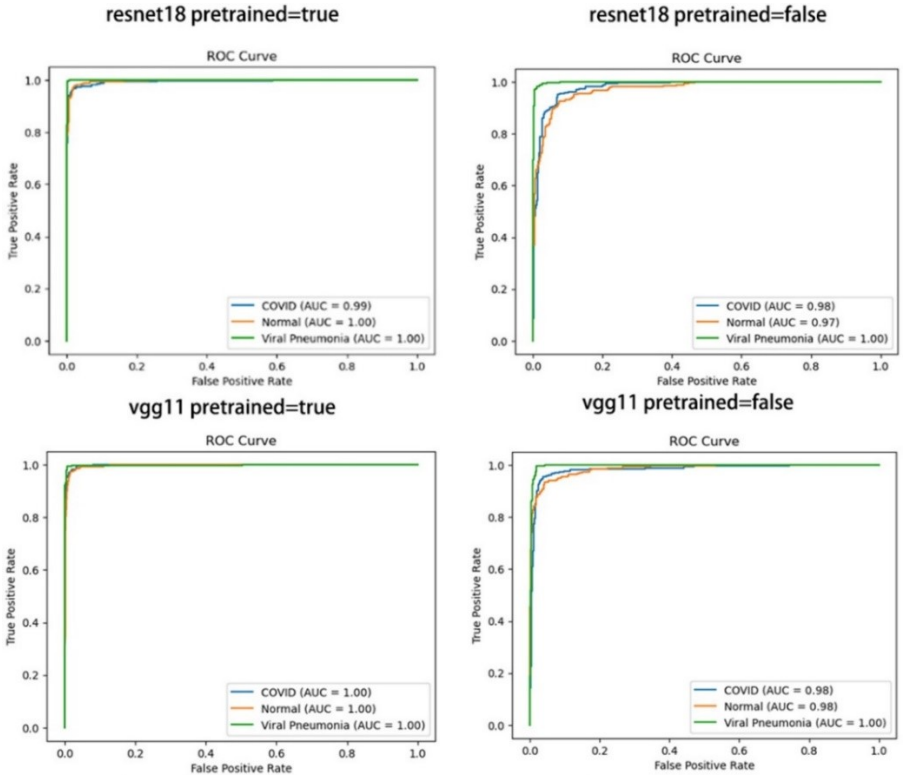


Figure 4. The confusion matrix for different models, comparing those with and without pre-trained weights.



DOI: 10.54254/2755-2721/18/20230980

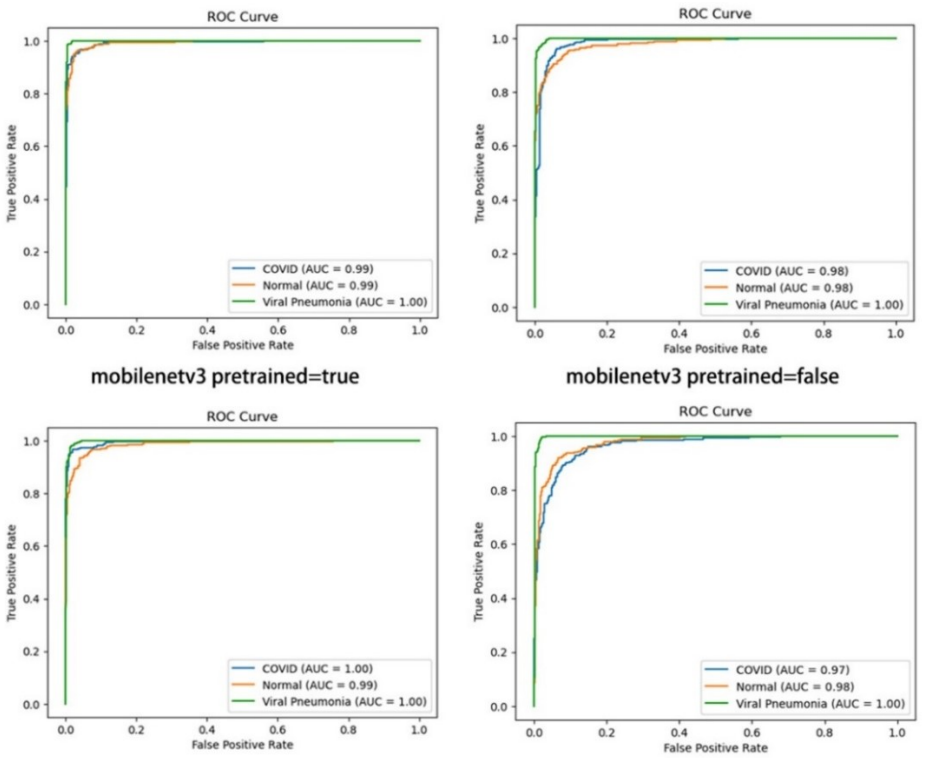


Figure 5. The ROC curve for different models, comparing those with and without pre-trained weights.

In addition to testing accuracy, other evaluation metrics are presented, including loss curves, confusion matrices, and ROC curve plots, which encompass the performance of all networks using different pre-trained weights, as depicted in Figure 3, Figure 4, and Figure 5. Figure 3 indicates that using pretrained weights allows the networks to converge more rapidly, and both ResNet18 and VGG11 demonstrate superior performance with minimal loss in the first epoch. Figure 4 illustrates that among the four networks, the distinction between Viral Pneumonia and the other two conditions is exceptional, with high recognition accuracy and low misclassification rates. The differentiation between COVID and normal cases is, in comparison, substantially less strong. The VGG11 model attains the highest classification accuracy for each category. Finally, Figure 5 reveals that, across any classification threshold, the performance of all four networks is better when using pretrained weights compared to not using pretrained weights. This phenomenon may be attributed to the presence of image features in ImageNet resembling medical images, such as black and white images similar to the X-ray images. Furthermore, the features learned from the extensive ImageNet dataset are sufficiently generic, with some fundamental features applicable to medical image classification tasks. During ImageNet pre-training, models learn to recognize various textures, shapes, and patterns found in a wide range of images, including medical images. These features provide a robust foundation, enabling improved performance when fine-tuned on smaller medical image datasets. By adopting this step, faster convergence and superior performance can be attained relative to commencing the training process from scratch.

The better performance achieved by the VGG model might be ascribed to VGG11's deeper architecture compared to AlexNet, allowing it to learn more intricate features and hierarchies from chest X-ray images. ResNet18's residual connections might not offer significant advantages in this task, while MobileNet is designed for resource-constrained environments, trading some accuracy for computational efficiency. VGG11's small-sized (3x3) convolutional filters enable learning a richer set of local features,

DOI: 10.54254/2755-2721/18/20230980

potentially beneficial for chest X-ray images where local patterns and textures are vital for accurate classification.

4. Conclusion

This study investigated the transferability of ImageNet pre-trained weights for natural images in medical image classification models based on classic networks compared to not using pre-trained weights. ResNet18, VGG11, AlexNet, and MobileNet were employed in this research for comparison. All four networks utilized the same dataset and hyperparameters, and the performance of each network was recorded for both using pre-trained weights and not using initial weights.

The experimental results demonstrate that using ImageNet pre-trained weights for transfer learning in medical image classification indeed enhances the performance of various network models, yielding higher classification accuracy and faster convergence. Moreover, among the four networks, VGG11 exhibited the highest classification performance for this study. While the conclusion supports the benefits of using ImageNet weights for medical image training, the underlying reasons warrant further exploration and explanation. In the future, efforts will be directed towards elucidating why transferring ImageNet weights to medical image classification tasks is helpful and identifying the optimal network for medical image classification.

References

[1] Mair M D et al 2021 A Systematic Review and Meta-Analysis Comparing the Diagnostic Accuracy of Initial RT-PCR and CT scan in Suspected COVID-19 Patients Br J Radiol 94(1119) 20201039
[2] Woloshin S Patel N and Kesselheim a S 2020 False Negative Tests for SARS-CoV-2 Infection - Challenges and Implications N Engl J Med 383(6) e38
[3] Liu J Wang Y He G Wang X and Sun M 2022 Quantitative CT Comparison between COVID-19 and Mycoplasma Pneumonia Suspected as COVID-19: A Longitudinal Study BMC Med Imaging 22(1) 21
[4] Litjens G Kooi T Bejnordi B E et al 2017 A Survey on Deep Learning in Medical Image Analysis Medical Image Analysis 42 60-88
[5] Farhan A M Q and Yang S 2023 Automatic Lung Disease Classification from the Chest X-Ray Images Using Hybrid Deep Learning Algorithm Multimed Tools Appl
[6] LeCun Y Bengio Y and Hinton G 2015 Deep Learning Nature 521(7553) 436-444
[7] Suganyadevi S Seethalakshmi V and Balasamy K 2022 A Review on Deep Learning in Medical Image Analysis Int J Multimed Info Retr 11 19-38
[8] Tajbakhsh N Shin J Y Gurudu S R et al 2016 Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine-Tuning? IEEE Transactions on Medical Imaging 35(5) 1299-1312
[9] Mustafa B et al 2021 Supervised Transfer Learning at Scale for Medical Imaging ArXiv abs/2101.05913
[10] Kaggle 2022 Convid19 radiography database https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database
[11] LeCun Y et al 1989 Backpropagation Applied to Handwritten Zip Code Recognition Neural Computation 1(4) 541-551
[12] Krizhevsky A Sutskever I and Hinton G E 2012 ImageNet Classification with Deep Convolutional Neural Networks Advances in Neural Information Processing Systems (NIPS) 1097-1105
[13] Howard A G Sandler M Chu G Chen L C Chen B Tan M et al 2019 Searching for MobileNetV3 Proceedings of the IEEE International Conference on Computer Vision (ICCV) 1314-1324
[14] Simonyan K and Zisserman A 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv preprint arXiv:1409.1556
[15] He K Zhang X Ren S and Sun J 2016 Deep Residual Learning for Image Recognition Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 770-778