DETECT PNEUMONIA IN CHEST X-RAYS USING DEEP LEARNING MODELS

Chrysoula Dontaki
cdontaki@ihu.edu.gr
Advanced Machine Learning
School of Technology
International Hellenic University

Christos Gkouramanis
cgkouramanis@ihu.edu.gr
Advanced Machine Learning
School of Technology
International Hellenic University

Eleftheria Trigeni
etrigeni@ihu.edu.gr
Advanced Machine Learning
School of Technology
International Hellenic University

ABSTRACT

More than 1 million adults are diseased with pneumonia and around 50,000 die every year in the United States [1]. Early diagnosis, vaccines and appropriate treatment could prevent many of these deaths [2]. Currently the best imaging examination tool for diagnosing pneumonia is Chest X-rays [3]. However, it is a demanding task that radiologists have to consider. Therefore, there is a need to develop reliable and interpretable deep learning models that can make Chest X-ray diagnoses more quickly and accurately. In this study, the **classification of Chest X-rays images** into **no disease**, **bacterial** and **viral pneumonia** was conducted. Various image preprocessing steps and augmentation were applied, so as the machine learning models could benefit from these improved images. To build the models, transfer learning was used along with fine-tuning of hyper parameters. The performance of **EfficientNetB0, B4, B7, V2-L V2-B0 and MobileNetV2** was examined in terms of accuracy score. The results clearly revealed that **EfficientNetB0** model outperforms all of them with an accuracy of **86.5%.**

1. RELATED WORK

Since pneumonia is an aggressive disease, early diagnosis is deemed necessary. Therefore, many scientists have tried to facilitate the diagnosis by taking advantage of the virtue of deep learning. Several remarkable surveys have been conducted that can be a medical breakthrough.

An interesting approach has been applied by Antin, B., Kravitz, J., & Martayan, E. [4]. Their task was to make a binary classification of pneumonia or non-pneumonia regardless of the type of pneumonia. Their first step was a thorough exploration and visualization of data using PCA and t-SNE in order to better understand the data. Then, as a baseline model, they trained Logistic Regression with a L2 regularization term on 32x32 images due to memory limitation. Due to the significant class imbalance, they evaluated the model by measuring AUC (Area Under the Curve) metric which was 0.60. An attempt to improve their results forced them to perform a Logistic Regression of 128x128 images. However, the test score dropped even further. They concluded that Logistic Regression could not cope with the stringent demands of the task and thus led to the deep learning approach. Initially, as part of the data augmentation they flipped the images horizontally and then trained a DenseNet using transfer learning and the Adam optimizer. They achieved a test score of 0.609 which was a slight improvement compared to Logistic Regression. Even the test scores remained low, their approach was very interesting and based on these remarkable results can be obtained.

A very exhaustive research was accomplished in paper [5]. The goal was to determine if a patient was healthy or had viral or bacterial pneumonia. To address this task, a transfer learning approach was vital. AlexNet, DenseNet121, InceptionV3, GoogleNet and ResNet18 were trained using the Adam optimizer. The best performance achieved by ResNet18 which reached a test accuracy of 92.86%. Then an ensemble method that combined the predictions of the five models regarding the majority voting, improved the results and achieved 96.39% accuracy.

The paper [6] presents another noteworthy research. They had a dataset of 5247 X-ray images and to be more specific, 2561 showed bacterial pneumonia, 1345 viral pneumonia and the rest no disease. The main goals were to detect pneumonia and they performed three different experiments to treat it. The first experiment was considered as a classification task with two classes: a patient had pneumonia or not. The second approach concerned a classification problem with three categories: a patient either had viral pneumonia, had bacterial pneumonia or was healthy. The last approach was to delve deeper into the differences between viral and bacterial pneumonia and this is because sometimes it is not easy to determine the type of pneumonia. Thus, in this approach they considered the challenge as a binary classification with classes: bacterial pneumonia or viral pneumonia. Regardless of how they dealt with the problem, they implemented data augmentation, transfer learning and trained four CNN models: AlexNet, ResNet18, DenseNet201 and SqueezeNet. They compared the results and DenseNet201 was the winner achieving the highest accuracy in all experiments. To be more detailed, it achieved a test accuracy of 98% in the first case that

mentioned above, 93.3% in the second case and 95% in the third case.

During the pandemic, Covid-19 could also cause pneumonia. El Asnaoui, K [7] conducted a study focusing on the power of ensemble learning models to detect the type of pneumonia. While he was going to classify the X-ray images in Covid-19 bacterial, viral pneumonia and no diseases, three questions arose. The first was about the accuracy that a deep learning model can offer in such problems. He also wondered if ensemble methods could improve it. Finally, he wanted to examine whether the number of combined models affects accuracy. Similar to the above works, he implemented data augmentation, transfer learning and trained InceptionResNetV2, ResNet50 and MobileNetV2. They all performed very well with InceptionReNetV2 to achieve the highest accuracy. He then created ensemble models that contained all possible combinations of individual models. He noticed that ensemble models perform better. He cannot answer with strong evidence if the number of combined models positively or negatively affects the accuracy. He noticed that it depends on how one would evaluate the model. However, for this classification task the ensemble ResNet50 with MobileNetV2 was the winner if one takes into account the F1 score and training time.

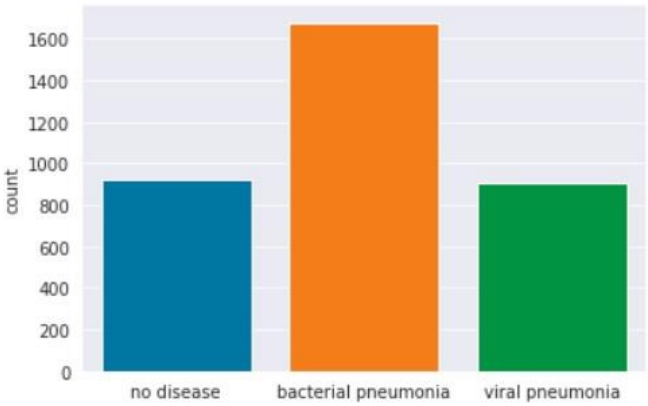
2. DATA & PROBLEM DESCRIPTION

The data are mostly taken from the dataset Chest X-Ray Images (Pneumonia) [8] which is available on Kaggle [9]. Each image is labeled with 0 if it corresponds to a patient without disease (normal) or 1 if the patient suffers from bacterial pneumonia or 2 if the patient diseased with viral pneumonia. The normal chest X-ray depicts clear lungs without any areas of abnormal opacification in the image. Bacterial pneumonia typically exhibits a focal lobar consolidation, whereas viral pneumonia manifests with a more diffuse "interstitial" pattern in both lungs. There are 5,840 X-Ray images (JPEG) of which **4,672** are **train images** and the remaining **1,168** are **test images**. We organized the dataset into 3 folders (train, val & test) and split them to subfolders for each image class. The problem consists of classification of Chest X-rays on three different classes of pneumonia.

Table 1: Dataset categorization

Images	Type	No disease	Bacterial Pneumonia	Viral Pneumonia	Total
Train	orig	1043	1902	1026	3971
	augm	983	1701	956	3640
	total	**2026**	**3603**	**1982**	**7611**
Valid	orig	184	336	181	701
	augm	174	300	169	643
	total	**358**	**636**	**350**	**1344**
Test	orig	?	?	?	**1168**
Total	orig	1227	2238	1207	4672
	augm	1157	2001	1125	4283
	total	**2384**	**4239**	**2332**	**8955**
					10123

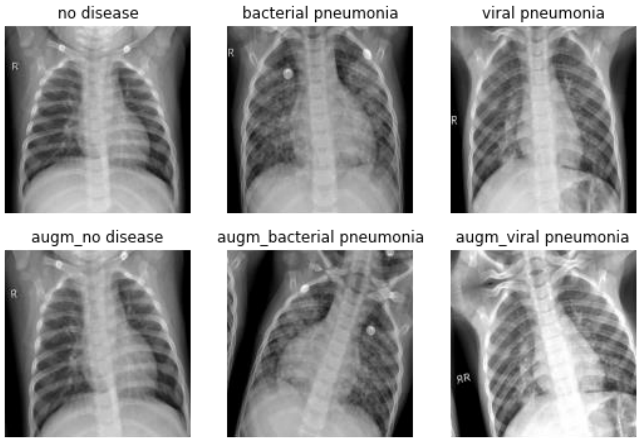
Figure 1: Number of images per class



The above figure shows that class imbalances are present in the dataset. To overcome this issue, we used class weights. Giving different weights to both the majority and minority classes will affect the classification of the categories during the training phase. The whole purpose is to penalize the wrong classification made by the minority class by setting a higher class weight and at the same time reducing the weight for the majority class. To increase the number of training examples, **data preprocessing** and **data augmentation** were applied. The first step was the transformation of grayscale images to **arrays** and the conversion of them into **RGB channels**. Next, the **resize** of the images to fewer pixels was implemented for example 224 × 224, 174 × 174. Moreover, to reduce the effect of illumination's differences the **normalization** of the images was conducted. Thus, most of the models will converge faster at [0,1] data than at [0,255]. In other words, we will get better results faster.

In order to avoid overfitting problem, the artificial expansion of the dataset was crucial. The idea was to alter the training data with small transformations to **reproduce variations of the existing dataset** [10]. Making it even larger, by randomly rotating the training images by 30 degrees, zooming by 20%, shifting horizontally by 10% of the width and vertically by 10% of the height. Furthermore, the brightness was increased and the images was randomly flipped horizontally using the **ImageDataGenerator** functionality from the TensorFlow Keras framework [11].

Figure 2: Chest X-Ray original images vs augmented images for each class



In order to accelerate the training of the deep convolutional neural networks and to overcome the problems of insufficient data **Transfer Learning** was used [12]. Instead of going through the long process of training models from scratch, this method of transferring learning from one predefined and trained model to some new domain by reusing the network layer weights was applied [5]. More explicitly, it extracts the knowledge from multiple source tasks and this knowledge is applied to a different target task. It is different from multitasking learning, despite learning from source and target at the same time, transfer learning focuses on the target task only [13]. As a result, the learning performance was improved by discarding data labeling efforts which are time-consuming [14]. Having significantly better accuracy when using **ImageNet pre-trained weights**, this was a good indication of success in classifying properly the Chest X-Rays.

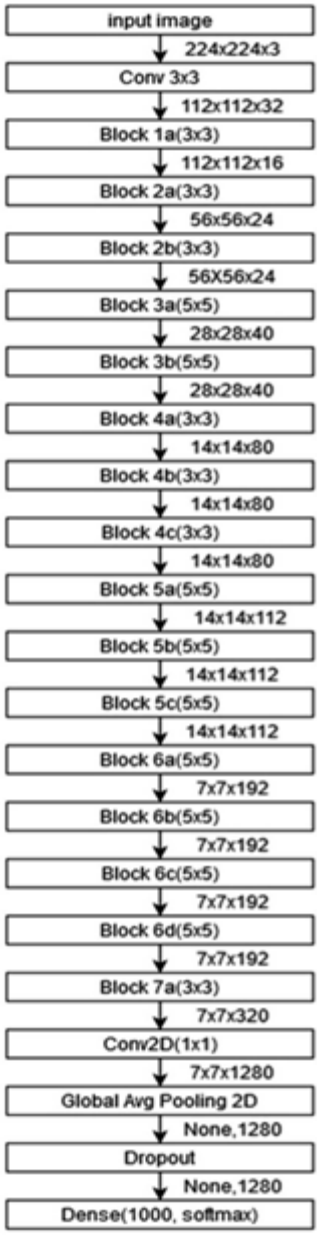
3. DESCRIPTION OF MODELS USED IN EACH SUBMISSION

To determine the classes (no disease, bacterial and viral pneumonia) regarding the Chest X-Rays images, six models, already trained on the ImageNet dataset, were implemented (EfficientNetB0, EfficientNetB4, EfficientNetB7, EfficientNetV2-L, MobileNetV2 and EfficientNetV2-B0). For pneumonia diagnosis, the Python programming language with the help of TensorFlow, Keras, Scikit-learn and OpenCV libraries was used to implement the framework. Google Colab was utilized in the GPU runtime in the training and validation phases.

EfficientNet is a family of Convolutional Neural Networks (CNNs) that achieved high performance and accuracy in the ImageNet Challenge [15]. This model significantly outperforms other ConvNets. It is about 8 times smaller and 6 times faster to inference compared to the best existing groups such as SENet [16] and GPipe [17]. EfficientNet uses a composite scaling method to create different models in the family that trade volume for accuracy. The composite scaling regularly measures the depth, width and resolution of the network. Network depth corresponds to the number of layers in a network. Width is related to the number of filters in a convolutional layer. The resolution is the height and width of the input image [18]. EfficientNets require fewer parameters and less computation time than most other models. In this competition, we utilized the **EfficientNetB0** as our baseline model, because it has less parameters than the rest models (B1-B7) of the EfficientNet family. Moreover, it is more cost-efficient for training and tuning than the more advanced EfficientNetB1-B7 models, as it does not require much computational power [19].

First, we **froze the weights** of the earlier layers of the pre-trained model to help us extract the generic low-level patterns (**feature extraction transfer learning**) from the Chest X-Ray image data. This stage is critical as the pre-trained model was trained for a different classification task [20]. The input of the EfficientNetB0 model was the images of size **224 × 224 × 3** after applying augmentation. Then, the features extracted automatically utilizing this architecture of EfficientNetB0 with transfer learning. The structure of EfficientNetB0 comprises **a convolutional layer** and **seven blocks**. With the **global average pooling layer**, the aggregation of the most important outputs of the base model was achieved so as to boost translational invariance and lessen feature map extent [21]. The next layer was a **dropout layer** with dropout rate 0.2 which means 20% of the neurons will output 0 to reduce the overfitting. Then, a **dense** layer of neuron numbers 1280 and Rectified Linear unit (**ReLu**) used as an activation function in the dense layer to strengthen the network for solving nonlinear problems. For training the model, **categorical cross-entropy loss function** is minimized by using 'Adam' optimizer. The **total layers** in our base model were **237** and the total number of **epochs** for training was **20** with a **batch size** of **16**. The base learning rate was chosen to be 0.0001. The next step was the classification of Chest X-ray images based on the extracted features. In the testing phase, the test dataset was evaluated using the trained model based on the validation accuracy and other performance measures like precision, recall, F1-score etc.

Figure 3: Distribution of layers in the EfficientNetB0 model



Our algorithmic approach for identifying potential pneumonia using the EfficientNetB0 model achieved **the highest accuracy** in this prediction task, thus we proceed with this model and tried to optimize it. We **fine-tuned** our EfficienNetB0 by unfreezing a few of the top layers (**fine-tuning transfer learning**) and jointly training both the unfrozen layers and the two layers that we added. By training some of the top layers, we adjusted the presentations of the pre-trained model that were more abstract to make them more relevant to our sample. In this way, the performance of our model improved.

Several other variants of this algorithm such as **EfficientNetB4, EfficientNetB7** and **EfficientNetV2-L** were also tried, but they failed to produce better predictions from our tests. It is worth mentioning for these algorithms that have extensive width, depth and resolution. However, while experimenting with the Chest X-ray images, the accuracy gain was observed stable. So, we insisted on trying **MobileNetV2** since it is a model specially designed for images and used for classification and feature generation tasks. MobileNetV2 was released in early 2018 [24] and its architecture adds two new modules, reverse residual and linear bottleneck, which can achieve higher accuracy. The total parameters of the MobileNetV2 amount to 2,430,403, while the trainable parameters amount to 2,240,131. Experiments were also undertaken with **EfficientNetV2-B0** which proposed by Tan & Le [22] in 2021. It possesses faster training speed and better parameter efficiency. Compared to original EfficienNetB0, EfficienNetV2-B0 uses both MBConv and the new fused-MNConv in the early layers. The expansion ratio for MBConv in EfficientNetV2 is 4, which is smaller than the expansion ratio used in EfficientNetB0. The advantage of smaller expansion ratio is to reduce the memory access overhead [23]. This model after fine-tuning had lower performance than tuned EfficientNetB0. Generally, in all models we had higher results after fine-tuning.

4. COMPARATIVE EXPERIMENTS AND RESULTS

To analyze and compare the results of our six classification models, different metrics were used as evaluation criteria. But the metric that we want to examine in this competition is the test accuracy. Accuracy is the overall percentage of correctly no disease, bacterial and viral pneumonia images.

Accuracy = number of correctly predicted test images / total number of test images

Initially, **random guessing** was performed. For every image in the testing set, 0 or 1 or 2 was randomly selected with 0 representing an image with no disease, 1 a patient with bacterial pneumonia and 2 with viral pneumonia. When this method was applied, the accuracy score was very low as expected, due to the randomness of the process. The goal was to overcome random guessing and find the maximum accuracy.

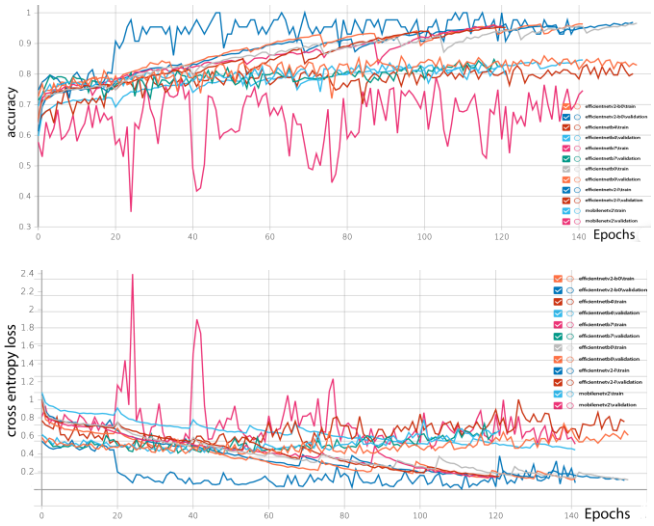
Table 2: Comparison of classification metrics between models

Model	Validation accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
EfficientNetB0	**84.77**	**85**	85	85
EfficientNetB4	79.80	80	80	81
EfficientNetB7	82.85	82	83	82
EfficientNetV2-L	80.36	79	80	79
MobileNetV2	76.38	76	77	76
EfficientNetV2-B0	81.93	82	83	81

As observed at Table 2, the accuracy score ranges from 76.38% to 84.77%. The difference between the models regarding accuracy is at most around 8%. The **EfficientNetB0** model received **the highest level of accuracy (84.77%)**. The output shows that we are able to successfully classify an image as "no disease", "bacterial" or "viral" with 84.77% accuracy. The **F1-score** of our model was **85%** considering both precision and recall, for times when we want a compromise between the two. It represents the harmonic mean (average) of precision and recall and will be high if both are high. In our model, both precision and recall were high. **EfficientNetB7, V2-L** and **V2-B0** had similarly high F1-scores and accuracy. However, they did not perform as well in the test set, so we did not trust these models.

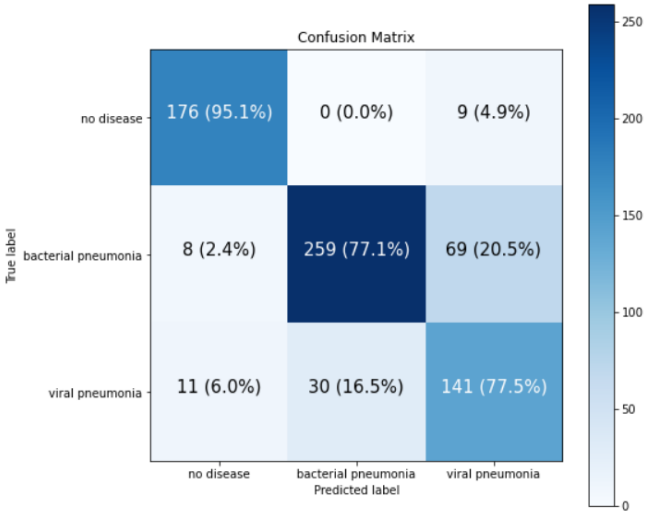
We preferred to train again the two most accurate models: **EfficientNetB0 and V2-B0** with the whole training dataset. The advantage of this approach was the accuracy in prediction, which improved to **85.22%** for **EfficientNetV2-B0**. Tuning the **EfficientNetB0** model, it gave us the **optimal** accuracy score of **86.5%.**

Figure 4: Accuracy and Cross entropy loss against epoch for all trained models



In Figure 4 plots of accuracy and loss against epochs are provided. The **best results** were obtained by the **EfficientNetV2-B0** model both in terms of accuracy and loss values. EfficientNetB0 had also a good performance trained for 162 epochs, while EfficientNetV2-B0 for 151 epochs.

Table 3: Confusion matrix of **EfficientNetB0** model



In the above table of confusion matrix, we observed that our model almost perfectly predicted the "no disease" patients (95.1%), while only 4.9% was misclassified as having bacterial or viral pneumonia. However, for images with bacterial pneumonia, our model got wrong having **20.5% incorrect predictions**, considering them as viral. Respectively regarding viral pneumonia, the **model errors reached 16.5%**. Thus, an important observation from this table was that our model finds it difficult to distinguish bacterial from viral Chest X-Rays, which in fact is difficult for scientists as well.

Thereafter, we proposed an ensemble model that combines outputs from all pre-trained models, which outperformed individual models, reaching the state-of-the-art performance in pneumonia recognition. More specifically, we ensembled our **EfficientNetB0, B4** and **B7** models using majority vote. The bagged classifier counted the votes of each individual classifier and assigned the class with the most votes. Our ensemble model reached an accuracy of **83.72%** on unseen data.

5. CONCLUSIONS

In this competition, we presented our deep learning-based approach to the detection of pneumonia in X-Ray images of patients' chests. While pneumonia diagnoses are commonly confirmed by a doctor, allowing for the possibility of error, deep learning methods can be regarded as a confirmation system. Therefore, we adopted the transfer learning approach and used the pre-trained architectures, EfficientNetB0, B4, B7, V2-L V2-B0 and MobileNetV2 trained on the ImageNet dataset, to extract features. These features were passed to the classifiers of respective models, and the output was collected from individual architectures. Before employing these models, data augmentation steps were implemented to increase the dataset size and improve

the performance of the models. Our results suggest that **EfficientNetB0** model **performed best** in categorizing images into no disease, bacterial or viral in all metrics, although EfficientNetV2-B0 had a similar performance with lower variability.

By taking into consideration the limitations of the present study, it should be referred that the dataset was imbalanced. Although, it has handled properly and implemented some techniques, there is still the possibility of imperfect data. The lack of good data can cause our algorithms to perform poorly and hence limit the capabilities of our model. It is also clear that the predictions on a Chest X-ray given by the trained deep neural networks cannot be relied upon alone. Its best use is in conjunction of clinical tests or professional diagnosis based on other clearer medical imaging methods.

REFERENCES

[1] CDC, 2017. https://www.cdc.gov/nchs/fastats/pneumonia.htm

[2] Aydogdu, M., Ozyilmaz, E., Aksoy, H., Gursel, G., & Ekim, N. (2010). Mortality prediction in community-acquired pneumonia requiring mechanical ventilation; values of pneumonia and intensive care unit severity scores. Tuberk Toraks, 58(1), 25-34.

[3] World Health Organization. (2001). Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children (No. WHO/V&B/01.35). World Health Organization.

16-[4] Antin, B., Kravitz, J., & Martayan, E. (2017). Detecting pneumonia in chest X-Rays with supervised learning. Semanticscholar. org.

9-[5] Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., ... & De Albuquerque, V. H. C. (2020). A novel transfer learning based approach for pneumonia detection in chest X-ray images. Applied Sciences, 10(2), 559.

[6] Rahman, T., Chowdhury, M. E., Khandakar, A., Islam, K. R., Islam, K. F., Mahbub, Z. B., ... & Kashem, S. (2020). Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. Applied Sciences, 10(9), 3233.

[7] El Asnaoui, K. (2021). Design ensemble deep learning model for pneumonia disease classification. International Journal of Multimedia Information Retrieval, 10(1), 55-68.

[8] https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[9] https://www.kaggle.com/competitions/detect-pneumonia-spring-2022/data

[10] Jaiswal, A. K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., & Rodrigues, J. J. (2019). Identifying pneumonia in chest X-rays: a deep learning approach. Measurement, 145, 511-518.

[11] Gulli, A., & Pal, S. (2017). Deep learning with Keras. Packt Publishing Ltd.

[12] Dai, W., Chen, Y., Xue, G. R., Yang, Q., & Yu, Y. (2008). Translated learning: Transfer learning across different feature spaces. Advances in neural information processing systems, 21.

[13] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[14] Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.

[15] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[16] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[17] Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., ... & Wu, Y. (2019). Gpipe: Efficient training of giant neural networks using pipeline parallelism. Advances in neural information processing systems, 32.

[18] El Zein, O. M., Soliman, M. M., Elkholy, A. K., & Ghali, N. I. (2021). Transfer Learning Based Model for Pneumonia Detection in Chest X-ray Images. International Journal of Intelligent Engineering and Systems, 14(5), 56-66.

[19] Nikolaou, V., Massaro, S., Fakhimi, M., Stergioulas, L., & Garn, W. (2021). COVID-19 diagnosis from chest x-rays: developing a simple, fast, and accurate neural network. Health information science and systems, 9(1), 1-11.

[20] Zebin, T., & Rezvy, S. (2021). COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization. Applied Intelligence, 51(2), 1010-1021.

[21] Shalbaf, A., & Vafaeezadeh, M. (2021). Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. International journal of computer assisted radiology and surgery, 16(1), 115-123.

[22] Tan, M., & Le, Q. (2021, July). Efficientnetv2: Smaller models and faster training. In International Conference on Machine Learning (pp. 10096-10106). PMLR.

[23] Huang, M. L., & Liao, Y. C. (2022). A lightweight CNN-based network on COVID-19 detection using X-ray and CT images. Computers in Biology and Medicine, 105604.

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