applied sciences

MDPI

Article

Classification and Predictions of Lung Diseases from Chest X-rays Using MobileNet V2

Abdelbaki Souid [1], Nizar Sakli [2] and Hedi Sakli [1,2,*]

[1] MACS Laboratory (Systems Modeling, Analysis and Control), RL16ES22, National Engineering School of Gabes, University of Gabes, Gabes 6029, Tunisia; SouidAbdelbaki@gmail.com
[2] EITA Consulting, 5 Rue du Chant des Oiseaux, 78360 Montesson, France; nizar.s@eitaconsulting.fr
* Correspondence: hedi.s@eitaconsulting.fr

Featured Application: The method presented in this paper can be applied in medical computer systems for supporting medical diagnosis.

Abstract: Thoracic radiography (chest X-ray) is an inexpensive but effective and widely used medical imaging procedure. However, a lack of qualified radiologists severely limits the applicability of the technique. Even current Deep Learning-based approaches often require strong supervision, e.g., annotated bounding boxes, to train such systems, which is impossible to harvest on a large scale. In this work, we proposed the classification and prediction of lung pathologies of frontal thoracic X-rays using a modified model MobileNet V2. We considered using transfer learning with metadata leverage. We used the NIH Chest-Xray-14 database, and we did a comparison of performance of our approach to other state-of-the-art methods for pathology classification. The main comparison was by Area under the Receiver Operating Characteristic Curve (AUC) statistics and analyzed the differences between classifiers. Overall, we notice a considerable spread in the achieved result with an average AUC of 0.811 and an accuracy above 90%. We conclude that resampling the dataset gives a huge improvement to the model performance. In this work, we intended to create a model that is capable of being trained, and modified devices with low computing power because they can be implemented into smaller IoT devices.

Keywords: convolutional neural networks; deep learning; lung diseases X-ray images; mobileNeV2; multi-label classification

Citation: Souid, A.; Sakli, N.; Sakli, H. Classification and Predictions of Lung Diseases from Chest X-rays Using MobileNet V2. *Appl. Sci.* **2021**, *11*, 2751. https://doi.org/10.3390/app11062751

Received: 4 February 2021
Accepted: 16 March 2021
Published: 18 March 2021

1. Introduction

Different types of lung disease have affected many people around the world. Lung diseases make the lungs more prone to certain physical problems and air pollution. As a result, lung function is impaired. Some lung diseases, such as emphysema, asthma, pleural effusion, tuberculosis, and other diseases including aspiration fibrosis, pneumonia, and lung malignancies, lead to loss of versatility in the lungs, which causes a decrease in the total volume of air [1]. Lung disease spreads easily, so it is important to identify the problem and provide the patient with the appropriate treatment. To detect and diagnose lung diseases, radiologists mainly deal with chest X-rays. Radiologists can use chest X-rays to diagnose and detect diseases and many other conditions. These diseases include bronchitis, infiltrations, atelectasis, pericarditis, fractures, and many others [2]. Chest radiography is considered the most-used type of disease examination in the world, with more than 2 million procedures performed each year [3].

This technology is essential and critical for the examination, diagnosis, and treatment of chest diseases, which is one of the main causes of death worldwide [4]. Therefore, to improve workflow priority and clinical decision support in large-scale projections and global population health programs, computer systems must be used to interpret chest radiographs as effectively as radiologists. In practice, this could play an important

role in many clinical environments. There is a lot of work done using AI technology to detect and diagnose lung diseases [5,6]. As a kind of probabilistic neural network, multilayer perceptron, learning vector quantization [7], and recurrent neural network (RNN) have been used to diagnose lung disease [8]. For the detection of lung diseases such as tuberculosis Sahlol, Ahmed T et al. [9] proposed a new hybrid method for tuberculosis X-ray image classification by extracting features from X-ray images with transfer learning and then filtering the produced huge number of features using the recently proposed Artificial Ecosystem-based Optimization (AEO); this method surpassed 90% accuracy in two datasets. Chronic Obstructive Pulmonary Disease (COPD) and pneumonia were performed with a neural network to diagnose cancer [10]. He-xuan Hu et al. [11] exceeded the state-of-the-art in lung tumor segmentation by proposing parallel deep learning model a hybrid attention mechanism and DenseNet, which got 94% accuracy.

The advances in deep learning and large datasets allow the algorithm to match the performance of medical professionals in various medical imaging tasks, including the detection of diabetic retinopathy [12], the classification of skin cancer [13], and metastases in the detection of lymph nodes [14]. The automatic diagnosis of chest imaging has attracted increasing attention [15,16] and special algorithms have been developed for the detection of pulmonary nodules [17], but to detect other diseases, such as pneumonia and pneumothorax, a method to classify multiple pathologies at the same time is suggested. Nilanjan, Dey et al. [18] developed a new method to detect pneumonia by computing the handcrafted features from the chest X-ray with the use of a modified VGGNet19 (Visual Geometry Group Network) [19]. This method achieved a 97.94% classification accuracy (Check Table 1). Until recently, computing power and the availability of large datasets allowed this approach to flourish. The National Institute of Health published ChestX-ray14 [20], which has led to many studies using deep learning for diagnosis from chest X-rays [21–24].

Table 1. New classification method with an accuracy higher than 90%.

	Method	Accuracy (%)
Sahlol, Ahmed T et al. [9]	A Novel Method for Detection of Tuberculosis in Chest Radiographs Using Artificial Ecosystem-Based Optimisation of Deep Neural Network Features	Dataset 1: 90.2% Dataset 2: 94.1%
He-xuan Hu et al. [11]	Parallel Deep Learning Algorithms with Hybrid Attention Mechanism for Image Segmentation of Lung Tumors	94.61%
Nilanjan, Dey et al. [18]	Customized VGG19 Architecture for Pneumonia Detection in Chest X-rays	No Threshold filter: 95.70% Threshold filter: 97.94%

Convolutional neural networks are the most widely used methods from ImageNet [25]. We want to deliver high accuracy results while keeping the parameters and mathematical operations as low as possible to bring deep neural networks to mobile devices. The MobileNet V2 architecture [26] was designed to create small, low-latency applications such as computer vision and IoT applications in many fields, namely fruit image classification for higher fruit production rates [27]; an artificial intelligent diagnostic system was built for cholelithiasis [28].

In this work, we propose a new method to combine transfer learning with feature selection to improve the state-of-the-art medical classification methods. After this section, we list some of the previous state-of-the-art lung disease classification methods in the related work Section 2. Dataset and disease distribution are discussed in Section 3. We present our model architecture and evaluation procedure in Section 4. Experimental results and comparisons with other works are presented in Section 5, The conclusions are presented in Section 6.

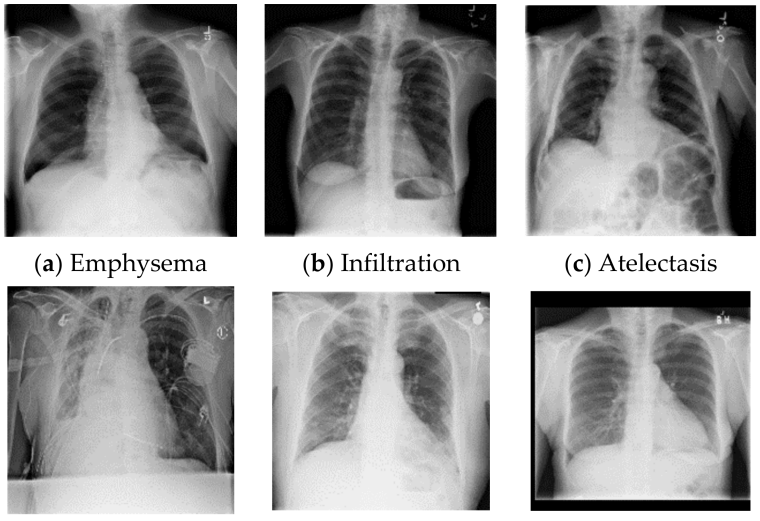
2. Related Work

As the spread of pulmonary disease increases, new, automatic, and reliable methods of accurate detection are essential to reduce the exposure of medical experts to an outbreak.

The ChestX-ray14, which consists of 112,120 frontal chest X-ray images of 30,805 unique patients, was triggered by the work of Wang et al. [20]. It has attracted a huge attention in the deep learning community, using convolutional neural networks (CNNs) in computer vision, and several research groups have started to focus on the application of CNNs for chest X-ray classifications. In the work of Yao et al. [21], they offered a combination of a CNN and RCNN (Recurrent Neural Network) to exploit the dependency on labels. As a CNN skeleton, they used a DenseNet [22] model that was fully adapted and trained on X-ray data. Li et al. [23] provided a framework for classifying and detecting pathologies using CNNs. More recently, Rajpurkar et al. [24] proposed refined transfer learning using a DenseNet-121 [26], which further increased the AUC results on ChestX-ray14 for multi-marker classification. In [2], various lung abnormalities, such as pulmonary nodules and diffuse lung, are detected and classified using CNN for feature extractions and R-CNN for the detection of lung disease. Author Yaniv Bar et al. [29] examined the power of deep learning methods on chest radiography data for the detection of pathologies. The dataset was trained using the ImageNet [25] competition, and mainly the descriptions DeCAF and PiCoDes (image codes) are extracted using the implementation of the Convolution Neural Network.

3. Dataset

The ChestX-ray14 database [20] was used to develop deep learning algorithms. This database is currently one of the largest public X-ray databases, containing 112 back-to-front and front-to-back thoracic films from 30,805 unique patients. Each image was annotated with as many as 14 different thoracic pathological ethics, which were selected based on the frequency of observation and diagnosis in clinical practice. The label of each image was obtained with the automatic extraction method in the radiological report, and each image produced 14 binary values (Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia, Pneumothorax). Some of them are provided in Figure 1, where 0 means that the pathology does not exist and 1 means that there are many pathologies in each image.

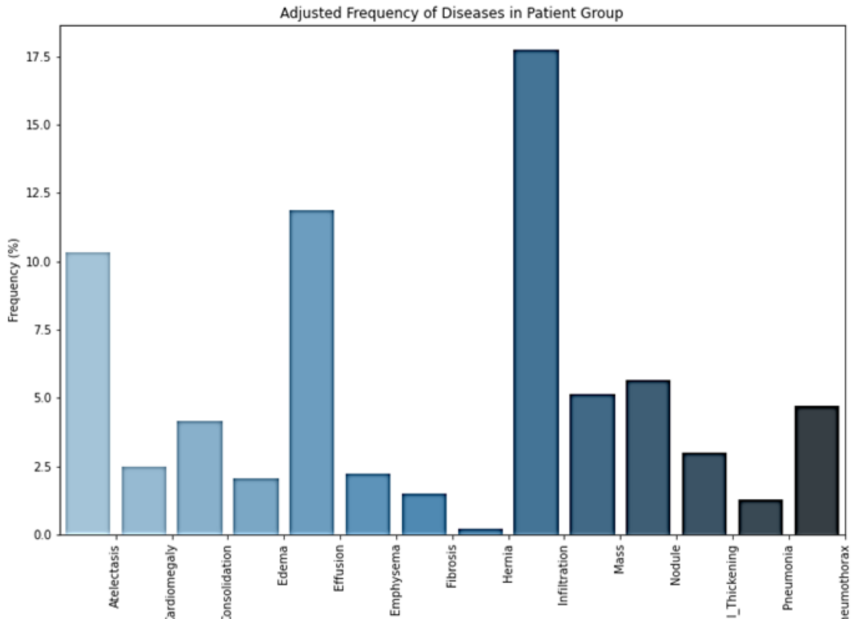


(**a**) Emphysema (**b**) Infiltration (**c**) Atelectasis

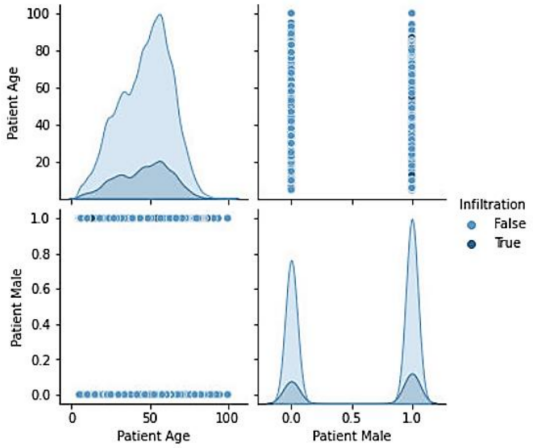
(**d**) Pneumothorax (**e**) No Finding (**f**) Nodule

Figure 1. Six examples of the ChestX-ray14 dataset. ChestX-ray14 consists of 112,120 frontal chest X-rays from 30,805 patients. All images are labeled with up to 14 pathologies or "No Finding". The dataset does include acute findings, such as the "Nodule" (**f**), and also treated patients with a drain, such as "Pneumothorax" (**d**).

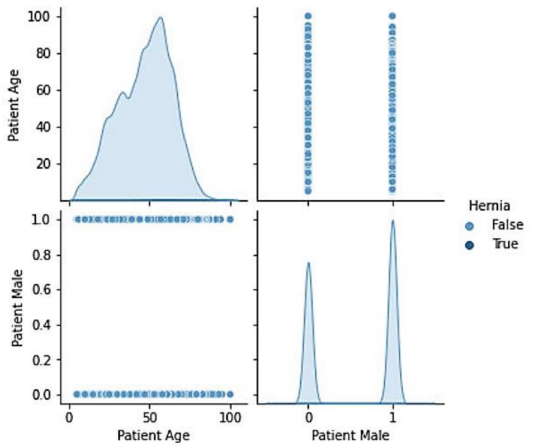
We show the distribution of each class illustrated in Figure 2. The prevalence of individual pathologies is generally low and varies between 0:2% and 17:74%, as shown in Table 2a, while the distribution of patient gender and view position is quite low. Even with a ratio of 1.3 and 1.5, respectively (see Table 2b), most of the data image samples were biased, as we got "Infiltration" as the most frequent with (9546) image samples, while the least frequent was "Hernia", with just 227 samples. "Emphysema"," Fibrosis"," Edema" Pneumonia", were close to each other, but were lower, with the number of samples between (322) and (727). For the other diseases, like in Figure 2a, the frequency of classes was over 1093 and under 4214. This class imbalance issue will have had a huge impact on the result. Next, we create the data generator to do data augmentation and split the dataset into training, validation, and testing. Figure 2b,c illustrates the distribution of the least frequent disease, "Hernia", and the most frequent, "Infiltration", with respect to the meta-information (age, gender, negative case, or positive case).



(**a**) Disease distribution in the dataset



(**b**) "Infiltration" distribution with meta-information features, including age, gender, and true/false disease (true with dark blue, false with light blue).



(**c**) "Hernia" distribution with meta-information features, including age, gender, and true/false disease (true with dark blue, false with light blue).

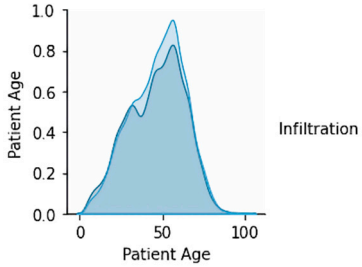
Figure 2. The distribution of diseases in the dataset: the figures (**a**,**b**) represent the least frequent disease, "Hernia", and the most frequent disease, "Infiltration", and include (**c**) the meta-information which are age, gender, case positive or negative.

Table 2. Overview of label distributions in the ChestX-ray14 dataset.

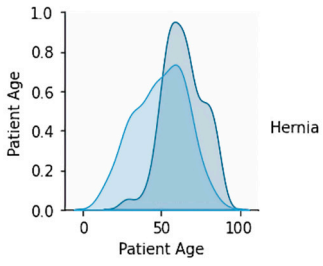
(a) Diseases			
Pathology	**TRUE**	**FALSE**	**Prevalence (%)**
Cardiomegaly	2776	109.344	2.48
Emphysema	2516	109.604	2.24
Edema	2303	109.817	2.05
Hernia	227	111.893	0.20
Pneumothorax	5302	106.818	4.73
Effusion	13.317	98.803	11.88
Mass	5782	106.338	5.16
Fibrosis	1686	110.434	1.50
Atelectasis	11.559	100.561	10.31
Consolidation	4667	107.453	4.16
Pleural Thicken.	3385	108.735	3.385
Nodule	6331	105.789	5.65
Pneumonia	1431	110.689	1.28
Infiltration	19.894	92.226	17.74
(b) Meta-information			
	Female	**Male**	**Ratio**
Patient Gender	63.340	48.780	1.3
	PA	**AP**	**Ratio**
View Position	67.310	44.810	1.50

For the split, we used a built function from the "sklearn" library called "train_test_split", and we got 38,819 image samples for the training, and for validation/testing we got, successively, 12,940 and 12,940.

The dataset was imbalanced and biased, as the difference between "Hernia" (227 samples) and "Infiltration" (more than 17,000 samples) shows. We tried to focus on the positive and negative frequency. As we can see in Table 2b, the false class samples were much higher than the true class samples (Ex: Edema True: 2303, False: 109,817 prevalence: 2.05%). To solve the class imbalance problem, we tried to include meta-information. At first, we used the patient age as the reference and then balanced the positive and negative samples for each disease by multiplying the weight of the positive cases with the negative sample frequency and did the same for the negative weight. The result is better illustrated in Figures 3 and 4 for both the least most disease frequencies.



(**a**) "Infiltration" after applying the true/false frequency balance (dark blue is true samples; light blue is false).



(**b**) "Hernia" after applying the true/false frequency balance (dark blue is true samples; light blue is false).

Figure 3. Distribution of "Hernia" and "Infiltration" based on age after affecting the balance.

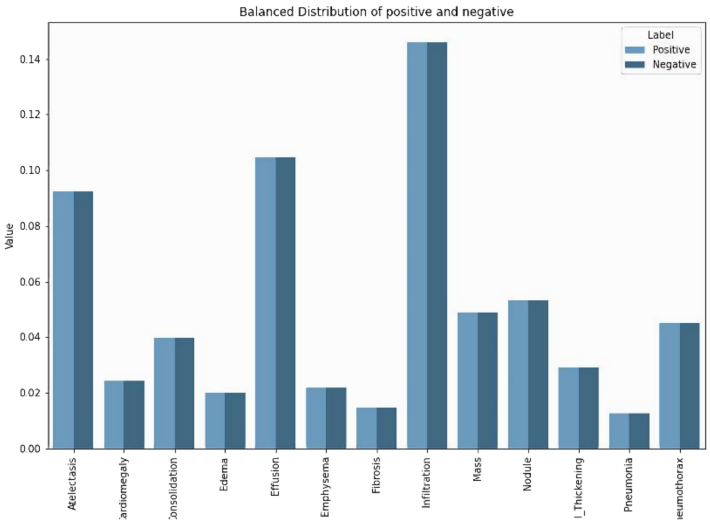


Figure 4. Distribution of all diseases based on the positive and negative weights for each disease.

4. Methodology

In this section, we propose the listed model in both Figures 5 and 6 to achieve the classification and prediction of 14 different lung diseases in chest X-rays using the Keras framework (which is an open-source software library that provides a Python interface for artificial neural networks) [30]. In this proposed work, we have identified and classified 112,120 chest X-ray images from the NIH dataset. The MobileNet V2 [26] model, and additionally the CNN layers, are employed in this work to predict and classify the chest thoracic diseases in the chest X-ray images.

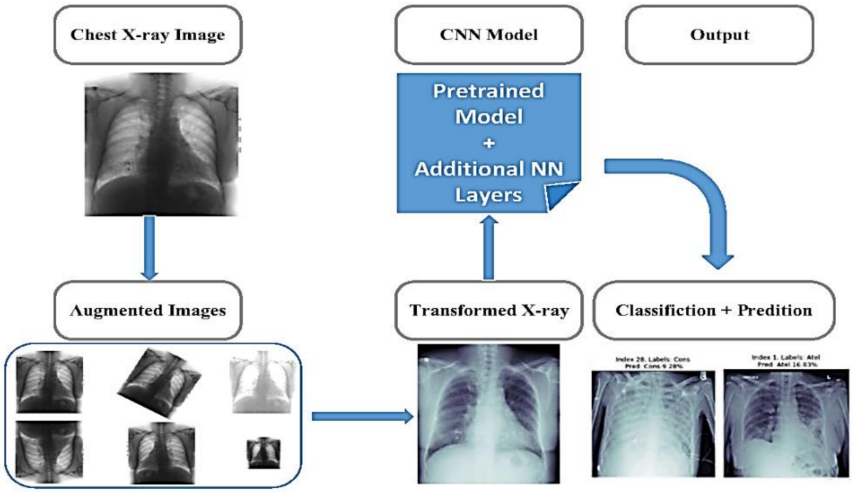


Figure 5. Diagram block for the proposed architecture: we started by loading the dataset and executed data sampling and image augmentation. Next, we loaded the CNN model and the pretrained MobileNet V2 on ImageNet plus and included additional layers on the train samples and the validation samples. Next, we loaded the new weight and started predicting the test samples. The result was a test image with a predicted label.

MobileNet V2 is an associated design suitable for basic mobile and on-board vision applications. Nowadays, deep learning methods are not only used to help computer vision, but also include various applications such as robotics, the Internet of Things (IoT), Natural Language Processing (NLP), and medical image processing application fields.

The architecture of the customized MobileNet V2 has a group of a hidden layers based on a bottleneck residual block and they have a depth-wise separable convolution that considerably reduces the number of parameters and leads to a lightweight neural network

differing from the normal convolution. The normal convolution is replaced by a depth-wise convolution and has a single filter which is followed by a pointwise convolution that is termed a depth-wise severable convolution [31].

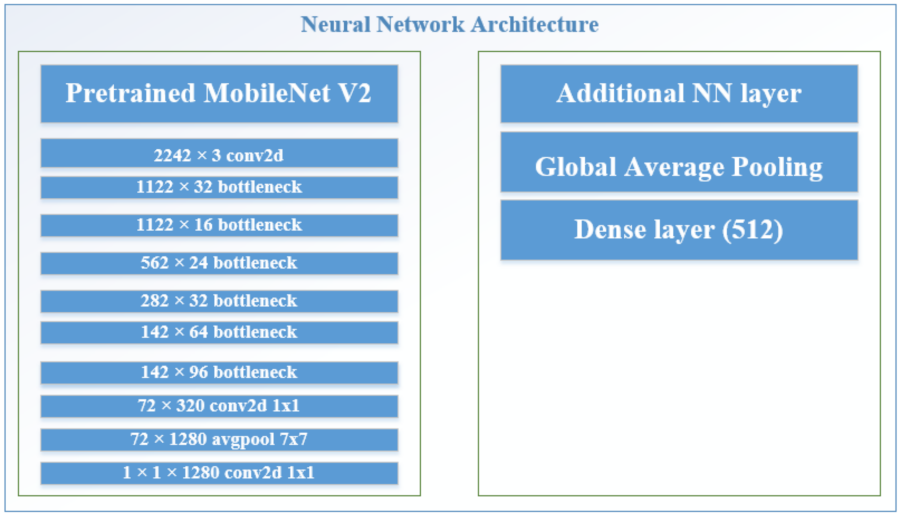


Figure 6. The convolutional neural network architecture: Constructed from an ImageNet [25] pretrained MobileNet V2 with an additional Global Average Pooling layer and one dense layer.

4.1. MobileNet V2 Architecture

The bottleneck residual block was mainly composed of three convolutional layers. As listed in Figure 6, the last two are the ones that already existed in the first generation of MobileNet: a depth-wise convolution that filters the inputs, followed by a 1×1 pointwise convolution layer. However, this 1×1 layer now has a different job.

MobileNet V1 [32] is one of the most popular mobile-centric deep learning architectures, which is not only small in size but also computationally efficient while achieving a high performance. The main idea of MobileNet is that instead of using regular 3×3 convolution filters, the operation is split into depth-wise separable 3×3 convolution filters followed by 1×1 convolution. While achieving the same filtering and combination process as a regular convolution, the new architecture requires a lower number of operations and parameters. In MobileNet V1, the pointwise convolution had to either double the number of channels or keep them the same. In MobileNet V2, the pointwise convolution does the opposite: the pointwise convolution makes the number of channels smaller. This is why this layer is now known as the projection layer, because it projects data with a high number of dimensions (channels) into a tensor while lowering the dimension, as illustrated in Table 3.

The first new feature that came with MobileNet V2 is the expansion layer. The expansion layer is a 1×1 convolution. Its role is to expand the number of channels in the image data before going into the depth-wise convolution. Hence, this expansion layer always has more output channels than input channels, as it does the opposite of the projection layer.

The second new feature in the MobileNet V2's building block is the residual connection presented in Figure 7. This works like in the ResNet [31] and helps with the flow of gradients through the network. The feature channels are increased by an expansion factor t. In our experiments, we used MobileNet V2 with $0.5\times$ and a $1\times$ channel multiplier with an input size of 224×224 for testing.

Table 3. MobileNet V2 original architecture.

Input	Operator	*t*	*c*	*n*	*s*
2242 × 3	conv2d	-	32	1	2
1122 × 32	Bottleneck	1	16	1	1
1122 × 16	Bottleneck	6	24	2	2
562 × 24	Bottleneck	6	32	3	2
282 × 32	Bottleneck	6	64	4	2
142 × 64	Bottleneck	6	96	3	1
142 × 96	Bottleneck	6	160	3	2
72 × 160	Bottleneck	6	320	1	1
72 × 320	conv2d 1 × 1	-	1280	1	1
72 × 1280	avgpool 7 × 7	-	-	1	-
1 × 1 × 1280	conv2d 1 × 1	-	k	-	

MobileNet V2: A sequence of 1 or more identical layers (modulo stride) repeated n times is represented by each line, and an expansion factor of t. Both layers in the same sequence have the same output channel number c. A stride s is on the first layer of each sequence and all others use one stride. 3 × 3 kernels are used for all spatial convolutions.

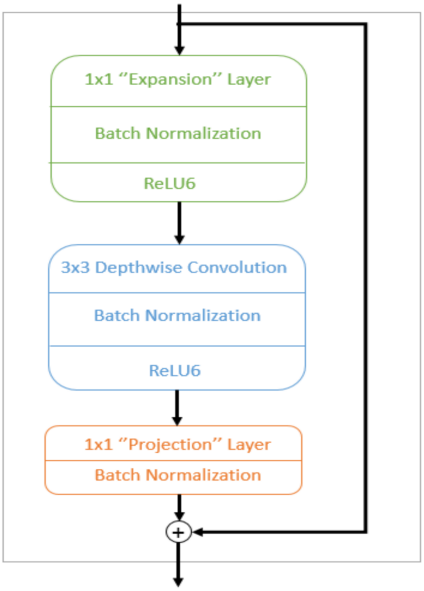


Figure 7. Bottleneck residual block: Three blocks are inside, the first bloc is the new feature of MobileNet V2 [26] architecture.

The MobileNet V2 focuses on optimizing latency, but at the same time, it also enables small networks to operate efficiently and support any size input, which can provide a better performance from the implementation of ReLU6 as the activation function in each layer, and also batch normalization. In the experimental results, the ImagNet [25] classification task shows that the MobileNet V2 outperforms MobileNet V1 and ShuffleNet [33] (width multipliers 1.5) with a comparable model size and computational cost. For a width multiplier of 1.4, MobileNet V2 also outperforms ShufflNet [33] (width multiplier 2) and NASNet [34], with a faster inference time. The MS COCO Object Detection task [35], MobileNetV2 + SSDLite [36] is 20× more efficient and 10× smaller while it still outperforms YOLOv2 (YOLOv2, You Only Look Once version 2 is a state-of-the-art, real-time object detection system), [37] on the COCO dataset. In PASCAL VOC 2012 [38] Semantic Segmentation, MobileNet V2 outperforms ResNet-101 [31] by obtaining mIoU (mean intersection of union) of 75.32%, with a lower model size and computational cost (2.11M parameters less than 70% of the ResNet-101 [31] parameters). The full MobileNet V2 architecture as shown in Figure 6 consists of 17 of these building blocks in a row followed by a regular 1 × 1 convolution, a global average pooling layer, and a classification layer (the very first

block is slightly different and it uses a regular 3 × 3 convolution with 32 channels instead of the expansion layer).

4.2. Preparation for Training the Data

In this work, MobileNet deep neural network was used with the supervised learning algorithm. After balancing the disease samples, we used a data generator function to divide the entire dataset into three groups, i.e., training (60%), validation (20%), and testing (20%).

We also use the sparse technique (also known as one hot encoding, to generate a column for each disease, this column will be given a "1" if the image in question has the discussed disease else, we affect to it a "0").

In the processing steps the datasets are loading and transforming images randomly by using TensorFlow and Keras packages. The image batch size is 32 for the three stages.

- Proposed model was trained in TensorFlow and Keras.
- First, we give inputs (size 224 × 224) to the MobileNet V2 as a base and additionally a global average pooling (GAP) layer (2 × 2) filter is performed for spatial data and reduce the dimension of data, and finally the Dense layer is an FC (fully connected) layer for all neurons connected to the next output nodes. For the activation function for the fully connected layer, we used the sigmoid function because of its robustness in the classification machine learning model.

For the training model, we applied several stages by applying fine-tuning the number of epochs and the batch size:

- At first, we trained the model for the epochs, we floor divided (floor division is a normal division operation except that it returns the largest possible integer), the total number of the train samples by their batch size. We used early stopping to avoid overfitting due to a big load of data in the step per epoch.
- In the second phase, we increased the number of epochs to 15 and divided the steps per epochs into half. To improve the accuracy, we reduced the learning rate.
- Lastly, we added 10 epochs and returned the step per epochs to the initial phase, and to minimize the overfitting, we both reduced the learning rate and early stopping and we configured the patience of the early stopping to force him to finch the 10 epochs.

We proposed a model performing under different hyper-parameters, namely filter size, number of kernels in all input channels, number weights, and pooling. The activation function translates the input range and calculates the loss function with the help of forward propagation on the training images. Batch normalization (BN) and ReLU were applied after each convolution layer for preventing overfitting problems, while epochs were also responsible for early stopping and data augmentation techniques were implemented (ImageDataGenerator). GAP layers act as a more depth type of reducing dimensionality: $H_{im} * W_{im} * D_{im}$, where H_{im} is height, W_{im} weight and D_{im} is Dimension. By handling all $H * W$ average values, the global average pooling (GAP) layer decreased by a single number ($1 \times 1 \times D$) of each $H*W$ feature map. The dense layer extracts the features from the convolution layer and was down-sampled by the GAP layer. In a fully connected layer (dense layer), every input node is connected to the output node.

The Adam optimizer algorithm finds the individual learning rate in every parameter for calculating the loss by binary cross-entropy and MAE (mean absolute error). We measured the performance of the classification and prediction probability value to be between 0 and 1.

4.3. Evaluation

We used accuracy (Acc), sensitivity (Sens), specificity (Spec), AUC and F1-score and time consumption as fitness measures. These are defined as follows:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$Sensitivity = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \tag{6}$$

Where TP (true positives) is the number of the disease samples that were labeled correctly, TN (true negatives) is the number of none of the 14 diseases samples that were labeled correctly. FP (false positives) is the number of the disease samples that were labeled incorrectly as they did not contain disease in their samples, and "FN" (false negatives) is the number of the not-diseases samples that were miss-classified as disease samples.

The proposed approach was implemented in Python 3.7 on Windows 10 bit using intel Xeon and 16 GB RAM, besides Google Colaboratory "Kaggle". The model was developed using Keras library [30] with a Tensorflow backend [39].

5. Experimental Results

The lack of computing power had a big role in limiting the experiment. In this work, we used the "Kaggle" open free cloud to edit and experiment with our model. Although this cloud offers a free GPU and fifteen gigabytes of ram, it was not enough to train our model over ten epochs. However, we managed to get a good result (Check Figures 8 and 9). We started by viewing the accuracy of the model and in the first epoch we had (0.951) a training accuracy and (0.913) a validation accuracy, with a difference of 3.74% between the training and validation. In the epoch number, the fourth gap between the training and validation accuracy gets much smaller with a training accuracy of 0.953 and a validation accuracy of 0.941. The difference decreases to 1.18%. In the last epoch, and we got some overfitting signs, with an increase in the difference between the training accuracy and validation accuracy with more than 2% of a difference.

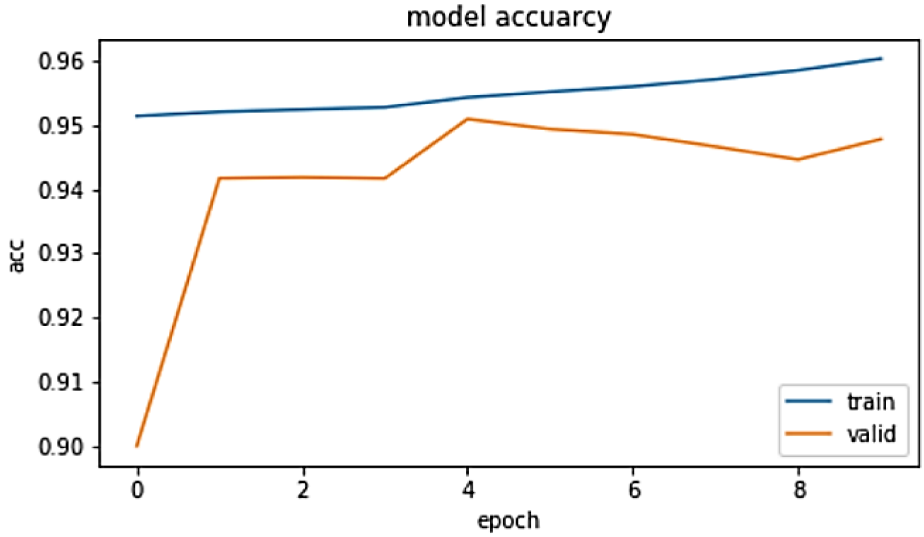


Figure 8. Model accuracy: a comparison between the training accuracy and the validation accuracy over a ten epoch.

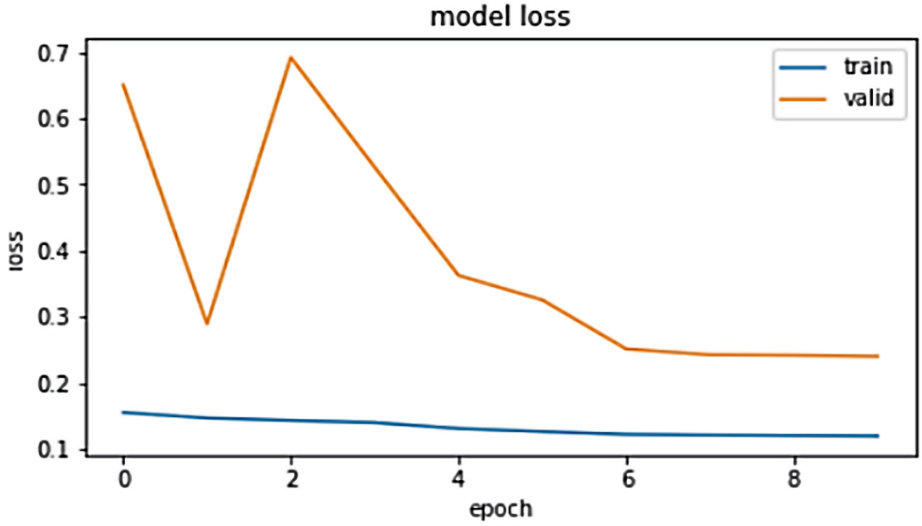


Figure 9. Model loss: the difference between the loss under the training phase and loss under validation loss.

The same case with the other metric, which is the loss function. In the first epoch, the difference between train loss and validation loss was about 4%. In the middle epoch, the gap starts to get bigger to pass 6% of a difference. In the last epoch, it starts to settle at 4%.

The specificity did show some promising results averaging 97%. The sensitivity on the other side did not provide a good result as it shows an average sensitivity of 45%. However, the F1-score got 55% (68% best and 40% lowest). The detail of the 14 disease class is provided in Table 4.

Table 4. Experimental result of our method.

Pathology	Accuracy	Sensitivity	Specificity	F1-Score
Atelectasis	0.796	0.507	0.968	0.683
Cardiomegaly	0.964	0.435	0.989	0.521
Consolidation	0.919	0.335	0.991	0.653
Edema	0.948	0.500	0.991	0.635
Effusion	0.794	0.405	0.924	0.510
Emphysema	0.958	0.635	0.996	0.478
Fibrosis	0.966	0.472	0.991	0.652
Hernia	0.997	0.600	0.991	0.400
Infiltration	0.735	0.328	0.948	0.560
Mass	0.880	0.445	0.972	0.432
Nodule	0.863	0.443	0.952	0.423
Pleural Thickening	0.962	0.460	0.991	0.607
Pneumonia	0.970	0.380	0.991	0.558
Pneumothorax	0.882	0.404	0.933	0.675
Average	0.902	0.453	0.973	0.556

We used the ROC curve to extract more detail about the result of our model. The ROC curve illustrated in Figure 10 is a plot which compares the False Positive Rate over the True Positive Rate of a predicted image.

For the “Emphysema”, the AUC is (0.891), and the “Effusion” and “Edema” got 0.876 and 0.884, the “Pneumothorax” and “Cardiomegaly” also had a good AUC, with 0.880 and 0.885 and “Mass”, “Hernia” and “Atelectasis” had a good AUC (0.826, 0.811 and 0.794). For the other diseases—"Consolidation”, “Fibrosis”, and “Pleural Thickening”—the was similar (0.790, 0.762, 0.763). The last three diseases got the lowest AUC: “Nodule” and AUC of 0.743; “Pneumonia” an AUC of 0.733; and “Infiltration” an AUC of (0.701).

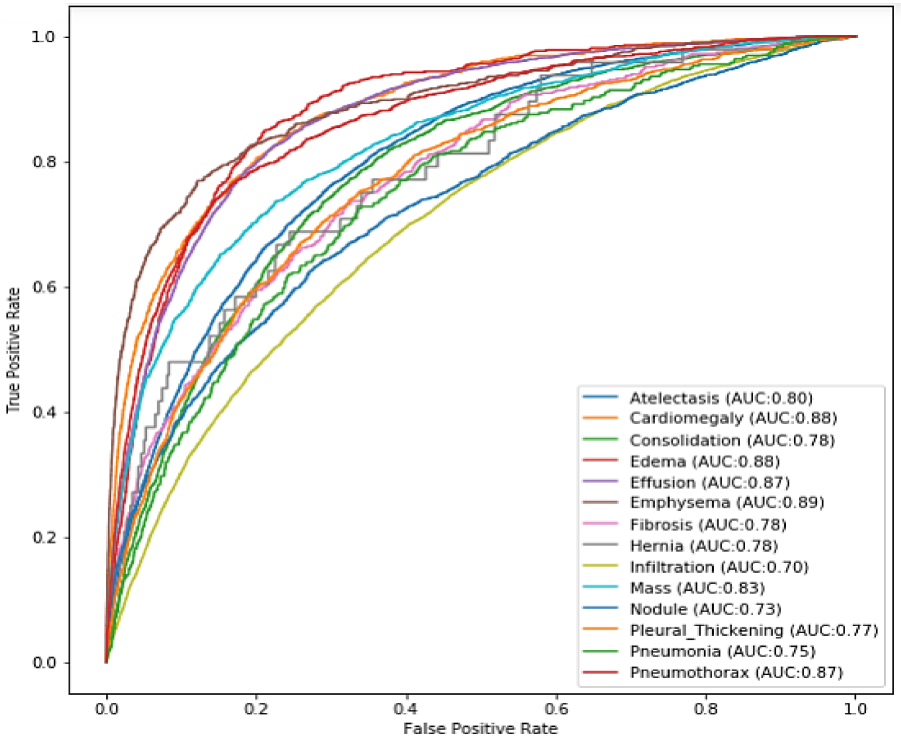


Figure 10. ROC curves for our model over ten epochs: AUC 0.89 for "Emphysema" and AUC 0.70 for "Infiltration".

As we can see in the image with the index zero from Figure 9, the ground truth label was "Effusion". Knowing that the model had AUC 0.89, the result was not surprising in that it gets 89.22% as a prediction. For the second image, we can see that the model starts getting a compromise and that the true diseases in this picture was "Mass", and the prediction of the models of two diseases got a prediction rate higher than 97%.

6. Comparison to Other Approaches

In our assessment, a significant distribution of results relative to AUC values was observed. In addition to the data split used, this might flow from the random initialization of the models and the stochastic nature of the optimizer.

When ChestX-ray14 was made publicly available, only images and no official data splitting was released. As a result, the researchers began training and testing their proposed methods on their own split database. With a different split to our re-sampling, we found a wide diversity in performance. Therefore, as shown in Table 5, direct comparisons with other results may fail in the sense of leading results. For example, Rajpurkar et al. [24] reported excellent results for the 14 categories in their own division. We compare the best performance of our model architecture (i.e., MobileNet V2) with experimental sampling with Li et al. [23] and Baltruschat et al. [40]. Baltruschat et al. in his paper compared the state-of-the-art of three models; the ResNet-38-large-meta had the best performance, and we will only include this model in the comparison. For our model, we recorded the minimum and maximum AUC.

We present the outcomes for our best performing architecture in this table and compare them to other categories. In addition, an AUC average of overall pathologies was provided in the last row, and we emphasize the overall highest AUC value by putting it in bold text.

We published results for our best performing architecture on the official released split by Wang et al. [20] in order to provide a fair comparison with other classes. First, we compared our results to Wang et al. [20] and Yao et al. [21] While the MobileNet V2 already has a higher average AUC, with 0.810 and in 6 out of 14 classes, there was a higher individual AUC with a plus-30 percent improvement on more than three disease classes

over Baltruschat et al. [40], with a ResNet-38-large-meta by Yao et al. [21]. Li et al. [23] also reported state-of-the-art results for the official split in all 14 classes, with an average AUC of 0.812. While our MobileNet V2 is trained with fewer images, it still achieved state-of-the-art results for "Effusion", "Edema", "Infiltration", "Pneumonia", and "Pneumothorax" and a slightly less average AUC of 0.810.

Table 5. AUC result overview for our experiment.

Pathology	Wang et al. [20]	Baltruschat et al. [40]	Yao et al. [21]	Li et al. [23]	Our Model
Atelectasis	0.700	0.755	0.733	0.800	0.794
Cardiomegaly	0.810	0.875	0.856	0.871	0.885
Consolidation	0.703	0.749	0.711	0.801	0.790
Edema	0.805	0.846	0.806	0.881	0.884
Effusion	0.759	0.828	0.806	0.859	0.876
Emphysema	0.833	0.895	0.842	0.870	0.891
Fibrosis	0.786	0.818	0.743	0.901	0.762
Hernia	0.872	0.937	0.775	0.773	0.811
Infiltration	0.661	0.709	0.673	0.701	0.711
Mass	0.693	0.821	0.777	0.831	0.826
Nodule	0.669	0.758	0.724	0.751	0.743
Pleural Thickening	0.669	0.761	0.724	0.791	0.763
Pneumonia	0.658	0.731	0.684	0.671	0.733
Pneumothorax	0.799	0.846	0.805	0.871	0.880
Average	0.745	0.806	0.761	0.812	0.810

7. Conclusions

A convolutional neural network (CNN) is developed for the diagnosis of thoracic (pulmonary) diseases in an X-ray image. The outperformance achieved is mainly due to the deep structure of CNN, which uses the mining power of different level features and resulted in a better generalization ability (Figure 11 shows some prediction result on the test-set). CNN models such as the MobileNet V2 network have superior generalization capabilities and precision compared to other networks and the results obtained demonstrate the high recognition rates of the proposed CNN and show that it gave very precise results in the diagnosis of the disease in the chest X-ray and then in classification.

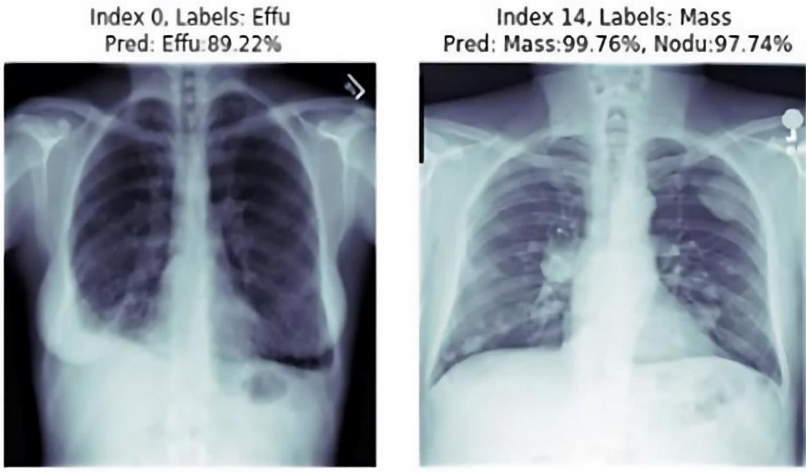


Figure 11. Result of the prediction attempt from the testing set: two other samples with different index labels. Each one could have one or more diseases.

Our optimized MobileNet V2 architecture achieves state-of-the-art results in six out of fourteen classes compared to Baltruschat et al. [40] (who had state-of-the-art results in five out of fourteen classes on the official split). For other approaches, even higher scores are reported in the literature. However, a comparison of the different CNN methods

with respect to their performance is inherently difficult, as most evaluations have been performed on individual and random partitions of the datasets. The model also had some failure when inspecting the F1-score, which could open the way to modify the model or to implement cross-validation.

In this work, the proposed model opens the way for light neural network architecture to be implemented in the medical field, which is very severe and depends on high quality resources and performs well under these conditions (AUC result was good, averaging 0.811, accuracy 95%) and can be more implemented in future experiments. However, the sensitivity of the model is rather low, and it could raise the false negative classification of samples in the prediction. The low sensitivity rate is the result of the biased dataset (the difference in the class distribution).

While the results obtained suggest that deep neural network training in the medical field is a viable choice, as more and more public databases become available, the practical application of deep learning in clinical practice remains an open topic. Especially for the ChestX-ray14 databases, the rather high tag noise of 10% makes it difficult to estimate the real network. Therefore, a clean test set without a label is necessary for a clinical efficiency evaluation.

Future work may include the investigation of other model architectures, new architectures for leveraging label dependencies and the incorporation of segmentation information.

Author Contributions: Conceptualization, A.S., H.S.; methodology, A.S., H.S.; software, A.S, N.S.; validation, A.S., H.S. and N.S.; formal analysis, A.S.; investigation, A.S.; resources, A.S., N.S.; data curation, A.S., N.S.; writing—original draft preparation, A.S.; writing—review and editing, H.S.; visualization, H.S., N.S.; supervision, H.S.; project administration, H.S., N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://nihcc.app.box.com/v/ChestXray-NIHCC (accessed on 1 September 2017).

Acknowledgments: This work was supported in part by the Research and Development Division of the society of EITA Consulting in France.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. National Institutes of Health. Conditions & Diseases. Available online: https://www.niehs.nih.gov/health/topics/conditions/index.cfm (accessed on 12 September 2020).
2. Kido, S. Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN). In Proceedings of the 2018 International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, Thailand, 7–10 January 2018; pp. 1–4.
3. Raoof, S. Interpretation of plain chest roentgenogram. *Chest* **2012**, *141*, 545–558. [CrossRef] [PubMed]
4. Mathers, C.D.; Loncar, D. Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Med.* **2006**, *3*, e442. [CrossRef]
5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
6. Pereira, S.; Pinto, A.; Alves, V.; Silva, C.A. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1240–1251. [CrossRef]
7. Sheikh Abdullah, S.N.H.; Bohani, F.A.; Nayef, B.H.; Sahran, S.; Al Akash, O.; Iqbal Hussain, R.; Ismail, F. Round Randomized Learning Vector Quantization for Brain Tumor Imaging. *Comput. Math. Methods Med.* **2016**, *2016*, 8603609. [CrossRef] [PubMed]
8. Sasikumar, S.; Renjith, P.N.; Ramesh, K.; Sankaran, K.S. Attention Based Recurrent Neural Network for Lung Cancer Detection. In Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 7–9 October 2020; pp. 720–724. [CrossRef]

9. Sahlol, A.T.; Abd Elaziz, M.; Tariq Jamal, A.; Damaševičius, R.; Farouk Hassan, O. A Novel Method for Detection of Tuberculosis in Chest Radiographs Using Artificial Ecosystem-Based Optimisation of Deep Neural Network Features. *Symmetry* **2020**, *12*, 1146. [CrossRef]
10. Er, O.; Yumuşak, N.; Temurtas, F. Chest diseases diagnosis using artificial neural networks. *Expert Syst. Appl.* **2010**, *37*, 7648–7655. [CrossRef]
11. Hu, H.; Li, Q.; Zhao, Y.; Zhang, Y. Parallel Deep Learning Algorithms with Hybrid Attention Mechanism for Image Segmentation of Lung Tumors. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2880–2889. [CrossRef]
12. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef] [PubMed]
13. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
14. Bejnordi, B.E.; Veta, M.; Van Diest, P.J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J.A.; Hermsen, M.; Manson, Q.F.; Balkenhol, M.; et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **2017**, *318*, 2199–2210. [CrossRef] [PubMed]
15. Qin, C.; Yao, D.; Shi, Y.; Song, Z. Computer-aided detection in chest radiography based on artificial intelligence: A survey. *BioMed. Eng. OnLine* **2018**, *17*, 113. [CrossRef] [PubMed]
16. Cicero, M. Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classi-fication of Abnormalities on Frontal Chest Radiographs. *Investig. Radiol.* **2017**, *52*, 281–287. [CrossRef] [PubMed]
17. Setio, A.A.A. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Trans. Med. Imaging* **2016**, *35*, 1160–1169. [CrossRef] [PubMed]
18. Dey, N.; Zhang, Y.-D.; Rajinikanth, V.; Pugalenthi, R.; Raja, N.S.M. Customized VGG19 Architecture for Pneumonia Detection in Chest X-Rays. *Pattern Recognit. Lett.* **2021**, *143*, 67–74. [CrossRef]
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
20. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the IEEE CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 2097–2106. [CrossRef]
21. Yao, L.; Poblenz, E.; Dagunts, D.; Covington, B.; Bernard, D.; Lyman, K. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv* **2018**, arXiv:1710.10501v2.
22. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
23. Li, Z.; Wang, C.; Han, M.; Xue, Y.; Wei, W.; Li, L.J.; Fei-Fei, L. Thoracic Disease Identification and Localization with Limited Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
24. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **2018**, *15*, e1002686. [CrossRef]
25. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009. [CrossRef]
26. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNet V2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
27. Xiang, Q.; Wang, X.; Li, R.; Zhang, G.; Lai, J.; Hu, Q. Fruit Image Classification Based on MobileNetV2 with Transfer Learning Technique. In Proceedings of the 3rd International Conference on Computer Science and Application Engineering (CSAE 2019), Sanya, China, 22–24 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; Article 121, pp. 1–7. [CrossRef]
28. Pang, S.; Wang, S.; Rodríguez-Patón, A.; Li, P.; Wang, X. An artificial intelligent diagnostic system on mobile Android terminals for cholelithiasis by lightweight convolutional neural network. *PLoS ONE* **2019**, *14*, e0221720. [CrossRef]
29. Bar, H.; Diamant, I.; Wolf, L.; Lieberman, S.; Konen, E.; Greenspan, H. Chest pathology detection using deep learning with non-medical training. In Proceedings of the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, USA, 16–19 April 2015; pp. 294–297. [CrossRef]
30. François, C. Keras. Available online: http://keras.io/ (accessed on 5 October 2020).
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385.
32. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
33. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *arXiv* **2017**, arXiv:1707.01083.
34. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning Transferable Architectures for Scalable Image Recognition. *arXiv* **2018**, arXiv:1707.07012.

35. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* **2015**, arXiv:1405.0312.
36. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *ECCV*; Springer: Cham, Switzerland, 2016.
37. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242.
38. Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. Available online: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html (accessed on 12 October 2012).
39. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: tensorflow.org (accessed on 31 May 2020).
40. Baltruschat, I.M.; Nickisch, H.; Grass, M.; Knopp, T.; Saalbach, A. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *Sci. Rep.* **2019**, *9*, 6381. [CrossRef] [PubMed]