Article

Pneumonia Detection on Chest X-ray Images Using Ensemble of Deep Convolutional Neural Networks

Alhassan Mabrouk [1,*], Rebeca P. Díaz Redondo [2], Abdelghani Dahou [3], Mohamed Abd Elaziz [4,5,6] and Mohammed Kayed [7]

1 Mathematics and Computer Science Department, Faculty of Science, Beni-Suef University, Beni Suef 62511, Egypt
2 Information and Computing Lab, atlanTTic Research Center, Telecommunication Engineering School, Universidade de Vigo, 36310 Vigo, Spain; rebeca@det.uvigo.es
3 Mathematics and Computer Science Department, University of Ahmed DRAIA, Adrar 01000, Algeria; dahou.abdghani@univ-adrar.edu.dz
4 Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt; abd_el_aziz_m@yahoo.com
5 Artificial Intelligence Research Center (AIRC), Ajman University, Ajman P.O. Box 346, United Arab Emirates
6 Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt
7 Computer Science Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni Suef 62511, Egypt; mskayed@gmail.com
* Correspondence: alhassanmohamed@science.bsu.edu.eg

Abstract: Pneumonia is a life-threatening lung infection resulting from several different viral infections. Identifying and treating pneumonia on chest X-ray images can be difficult due to its similarity to other pulmonary diseases. Thus, the existing methods for predicting pneumonia cannot attain substantial levels of accuracy. This paper presents a computer-aided classification of pneumonia, coined Ensemble Learning (EL), to simplify the diagnosis process on chest X-ray images. Our proposal is based on Convolutional Neural Network (CNN) models , which are pretrained CNN models that have been recently employed to enhance the performance of many medical tasks instead of training CNN models from scratch. We propose to use three well-known CNNs (DenseNet169, MobileNetV2, and Vision Transformer) pretrained using the ImageNet database. These models are trained on the chest X-ray data set using fine-tuning. Finally, the results are obtained by combining the extracted features from these three models during the experimental phase. The proposed EL approach outperforms other existing state-of-the-art methods and obtains an accuracy of 93.91% and a F1-score of 93.88% on the testing phase.

Keywords: image processing; deep learning; medical image classification; ensemble deep learning; vision transformer

1. Introduction

Virus infection has been one of the most serious threats to human health throughout history. One of the most common viral infections is pneumonia [1]. Infections caused by viruses and bacteria harm the lungs [2]. Pneumonia symptoms are common, including pain, cough, shortness of breath, etc. Pneumonia affects approximately 7.7% of the world's population each year. As a result, early detection is critical for such illnesses. Thus, the task of automated medical image classification has grown significantly [3]. This task aims to diagnose medical images into predefined classes. Recently, Deep Learning (DL) has become one of the most common and widely used methods for developing medical image classification tasks [4]. Further, DL models produced more effective performance than traditional techniques using chest X-ray images from pneumonia patients [2,5].

The DL architectures illustrated effective predictive ability and outperformed physicians [6]. On chest X-ray images, multiple tasks were performed on DL models, including

tuberculosis identification [7], tuberculosis segmentation [8], large-scale recognition [9], COVID-19 detection [10,11], and Radiograph classification [12]. The automated classification of chest X-ray images using DL models is growing rapidly, and choosing an appropriate region of interest (ROI) on chest X-ray images was used to discover pneumonia [13]. Furthermore, applying the DL modes helps to avoid problems that take a long time to solve in traditional approaches. However, these models require large volumes of well-labeled training samples.

To solve this problem, Transfer Learning (TL) has been developed. Due to its capacity to effectively solve the shortcomings of reinforcement learning and supervised learning, TL is becoming more widespread [14,15]. TL has the following types: Unsupervised, inductive, transductive, and negative learning. These types have been demonstrated to be able to tackle the DL problems [16]. To enhance accuracy, TL is essential to have the most suggestive contextual and exclusionary capacity in the feature extraction stage for many fields [17]. For example, online scraping [18], social media [19], sentiment classification [20], and medical image classification [17]. Therefore, this paper applies TL approaches to improve diagnostic reliability and reduce time-consuming decisions for clinicians.

We have proposed a new NN, coined as EL, which is built by training CNN models using TL methods. To achieve this goal, CNN's MobileNet [21], DenseNet [22], and Vision Transformer [23] methods were trained to detect pneumonia in chest X-ray images. We have decided to use three models to generate the proposed ensemble learning method trying to combine two approaches that, separately, have obtained promising results: on the one hand, using the best CNN models for the training stage; on the other hand, applying a vision transformer. Regarding the former, the authors of [24] proposed to use the best two CNN models to develop an ensemble learning solution. They built a hierarchical stacked method by using the most relevant features extracted by the selected two convolution networks, and they obtained a good performance. Additionally, Vision Transformer (VIT) has recently achieved highly competitive performance for several computer vision applications [25]. Furthermore, VIT achieves remarkable results compared to CNN while it obtains fewer computational resources for pretraining [23]. Since the existing approaches merge CNN models for creating an ensemble method without using the transformers, as in [26], we have decided to combine both to obtain a methodology that simultaneously merges and improves each individual proposal—that is, we propose a method that joins VIT and the selection of the best CNN models for the training stage. In addition, the obtained features from the selected three models were combined using a probability-based ensemble approach to achieve good classification performance.

With the aforementioned in mind, we have developed a novel method to enhance the diagnosis of pneumonia. This method is based on three well-known CNN models, which significantly improve classification performance. As a result of these processes, the following contributions to the suggested method are listed:

- We suggested an ensemble method that uses forecasts from multiple CNN models to improve the classification results.
- Instead of training a CNN model from scratch, we looked at appropriate transfer learning and fine-tuning methods.
- The architecture of the proposed ensemble learning method is improved by using a batch normalization layer and a dropout layer.
- A comprehensive analysis of the developed method is compared with different state-of-the-art approaches using a real-world data set.

The rest of the paper is organized as follows: Section 2 provides a review of related works. In Section 3, the existing CNN models and the proposed method are presented. The pneumonia classification performance of the proposed method is given in Section 4. Lastly, the conclusion provides future scope in Section 5.

2. Related Works

Over the past decade, many researchers have automatically used deep learning to detect lung infections and diseases from chest X-ray. For example, CheXNet is a 121-layer CNN-based approach developed by Rajpurkar et al. [27]. This approach was trained using 100,000 chest X-ray images from 14 different diseases. The approach was also applied using 420 chest X-rays, and the results were compared with those of radiologists. Therefore, it was found that the DL-based CNN method outperformed the average performance of radiological pneumonia detection. In [28], they trained a CNN method from scratch to retrieve features from chest X-ray images to achieve excellent classifier performance and used it to detect whether or not a patient had pneumonia, in contrast with previous studies based on traditional manual features. Wu et al., in [29], suggested a method based on adaptive average filtration CNN and random forest to predict pneumonia using chest X-ray images. The adaptive filtration was applied to remove noise from the chest X-ray image, improving accuracy and making it easier to identify. Then, using dropout, a CNN model with two layers is created for extracting features. Nevertheless, more preprocessing with the adaptive filter is required to enhance CNN's classification accuracy. However, there are some issues with CNN models, which require a large amount of data with labels to be trained. Furthermore, learning a CNN architecture is computationally expensive and requires advanced machines. As a result, a transfer learning (TL) approach has been proposed to solve these problems.

Recently, the TL method has become very popular, mainly because it enables the CNN model to be more efficient, have reduced costs, and require fewer inputs [30]. Ayan and Ünver [31] used the Xception and VGG16 structures to fine-tune transfer learning. The design of Xception was substantially altered with the addition of two fully linked levels and multiple-output tiers with a SoftMax activation mechanism. As per the theory, the channel's initial layer has the greatest generality potential. The previous eight layers of the VGG16 architecture have been stopped, and the fully linked levels have been altered. Similarly, the test period for each image was 16 ms for VGG16 and 20 ms for the Xception network. In [32], the methods included InceptionV3, ResNet18, and GoogLeNet. Classifier results were merged using the strong majority in this method. This means that the diagnosis goes with the group with a high proportion of first-time voters. Averaging out the model's testing results, this approach took 161 milliseconds per image. On top of that, they were able to classify chest X-ray images with great accuracy. Pneumonia may be detected using deep CNNs, as per the results of this research. We use standard algorithms as a component in our approach to categorizing data to keep computation costs minimum. Rahman et al. [33] used transfer learning techniques on ImageNet to detect pneumonia using four pretrained CNN architectures. They used three classification strategies to classify chest radiography images. Togacar et al. [34] utilized three well-known CNN models for extracting features in the pneumonia classification task. They used the same data for training each model individually and acquired 1000 features from every CNN's last fully connected layer. For this task, these features are essential, which was reduced by the minimum redundancy maximum relevance (mRMR) feature selection method. Moreover, the selected features were fed into machine learning (ML) classification algorithms. Mittal et al. [35] suggested a CapsNet architecture for diagnosing pneumonia in chest X-ray images using multilayered capsules. Liang and Zheng in [36] suggested a new residual network-based trained TL approach for pneumonia diagnosis. Further, the DL model used in their study had 49 convolutional layers and 2 dense layers. Their model had a 90.05% test accuracy. However, because of the huge number of convolutional layers used, this technique had a long execution time. In addition, Octave-Like Convolutional Neural Network [37] are considered lightweight and low-computational-cost neural networks that can replace the vanilla convolution operation such as in driver distraction detection [38], document image segmentation [39], and tumor segmentation [40]. Compared to the vanilla convolution, the octave CNN uses multifrequency feature representation, which decomposes the input into low- and high-frequency maps (feature representations) rather than only using the high frequency.

Thus, the low-frequency feature maps represent a low-resolution representation of the input, which helps decrease unnecessary redundancy and the concept spatial dimensions.

To address this problem, several papers have been recently published that attempt to detect pneumonia using deep CNN methods with a lower number of convolutional layers, as in [41,42]. For example, Liang and Zheng [36] used a CNN approach with residual junctions and dilated convolutional methods to identify pneumonia. While selecting chest X-ray, they revealed the influence of TL on CNN's approach. Transfer learning was used by Kermany et al. [43] to learn a CNN method to identify pneumonia in chest X-ray images. For classifying chest X-ray as normal vs. pneumonia, Rajaraman et al. [44] developed a new CNN-based approach. They used a region of interest (ROI) that only included the lungs rather than the entire image to learn CNN architecture. However, these approaches are still unable to achieve a high degree of efficiency in detecting pneumonia.

To sum up, there are interesting approaches in the state-of-the-art, but we have tried to go one step further by proposing a method that combines two different techniques: using CNN models for the training stage and taking the best one for ensemble learning, and using a vision transformer (VIT), which obtain good results. Therefore, the main difference between our proposal (EL method) and the other previous approaches is that we use an ensemble method that combines three well-known CNN models, one of them being the most recent vision transformer. The obtained results are promising and lightly improve the state-of-the-art performance, with a small number of layers and features.

3. Methodology

3.1. Deep Convolutional Neural Networks (DCNN) Models

Recently, many DCNN models have been suggested, which have been shown to enhance the productivity and effectiveness of machine learning (ML) [20,45]. Moreover, the DCNN models are among the most studied DL methods due to their capability to extract features automatically, and their adjustable structures, as in [15]. Many DL algorithms, such as MobileNet [21] and DenseNet [46], have incorporated the concept of depthwise separable convolutions to address the disadvantages of traditional operation. In contrast with traditional convolution operations, depthwise separable convolutions are performed independently of each input. Consequently, the algorithms are cost-effective to run and can be trained with fewer parameters in a short time. Therefore, the ensemble method has been recently introduced to learn more complex feature representations compared to single network [47].

There are two kinds of ensemble techniques utilized in CNN architectures [26]. In the first technique, some researchers employed different CNN algorithms to obtain features from the medical images, as in [48]. The collected features are aggregated and used in various machine learning techniques for classification/categorization tasks. Two distinct training methods and sophisticated algorithms are some of the limitations of this technique. In the second technique, predicted values are merged using a computational formula, as suggested in [49]. The benefit of this technique is that the ensemble method correctly classifies the data due to the resulting performed by other CNN models' correct predictions. Therefore, this paper employs an ensemble technique to improve the performance of the classification task.

3.1.1. MobileNet

The MobileNet architecture was designed by Howard et al. [21]. The MobileNet design is based on separable convolution layers and consists of two components: (a) Depthwise convolution—a single filter is applied to each input channel; (b) Pointwise convolution—a 1×1 convolution aggregates the depthwise convolution's outcomes. In a typical convolution, we filter and aggregate input images into a new vector of features through one phase.

Depthwise convolution is used to cut down on calculation time and model size. Eventually, MobileNet employs batch normalization and ReLU as a nonlinearity activation function. Furthermore, before the fully connected layer, the last average pooling decreases the spatial resolution to just one.

3.1.2. DenseNet

The DenseNet was suggested by Huang et al. [46] for improving the depth of CNN. This approach was first implemented to address issues when CNNs became more complex in the model size. The authors solved the issue by linking each layer completely to the next, thus assuring maximal information and gradient transfer. One of the key benefits of adopting such a structure is that the DenseNet structure maximizes its capacity by reducing the usage of a deep or broad design via feature recycling. Unlike traditional CNNs, DenseNet does not train duplicate features. Thus, it requires fewer parameters. Moreover, since the structure has relatively thin layers, it only adds a tiny number of new feature maps. Further, the structure depends on each layer having immediate access to the gradients from the loss function and the input image during the training stage.

It is worth noting that the DenseNet concatenates the layer's return image features with the input feature maps; thus, there is no aggregate between them. The feature maps could be the same dimension to accomplish this combination in any instance. To overcome this issue, DenseBlocks are a concept introduced by DenseNet. DenseBlocks are used to ensure that the size of feature maps stays consistent inside a block while the number of filters varies among them. Layers of a particular sort (called transition layers) are put in the DenseBlocks. Moreover, downsampling is performed using batch normalization, a 1×1 convolution, and 2×2 layers in these layers.

3.1.3. Vision Transformer (VIT)

The VIT has successfully obtained perfect performance on different computer vision tasks, as discussed in [25]. The Vision Transformer (VIT) [23] divides an image into patches and uses a transformer to pattern the similarity among these patches as sequences, resulting in sufficient image classification performance. VIT's structure can be summarized as follows: (1) Divide the given image into patches. (2) Flatten patches and use these patches to produce lower-dimensional linear embeddings (Patch Embedding). (3) Add a class token and positional embedding. (4) Give the patch sequence to the transformer layer and use a class token to obtain the label. (5) To obtain the output prediction, transfer the class token values to the Multilayer Perception (MLP).

Regarding inserting a 112×112 image to generate the patch, we start with 16×16 nonoverlapping and overlapping patches. Therefore, generating 49 patches becomes easy as well as inserting them into the linear projection layer, taking into account that each patch has three color channels. Further, the patches are loaded into the linear projection layer to achieve a long vector representation of each patch.

The overall number of overlapping and nonoverlapping patches in the patch embedding is 49, and the patch size with the number of channels is $16 \times 16 \times 3$. As a result, each patch's long vector is 768, and the patch embedding matrix is 49×196. In addition, class tokens and position embedding have been added to the sequence of embedded patches. If positional encoding is not used, the transformer will not be able to retain the information. Due to the additional class token, patch embeddings are still sized 50. Lastly, the acquired representations of the class token are obtained by feeding patch embeddings with a positional encoding and a class token into the transformer layer. As a result, the transformer-encoding layer produces 1×768, which is then transferred to the MLP block to give an accurate prediction.

The transformer encoder, which contains the Multihead Self-Attention (MHSA) block and the MLP block, is the most important element in the VIT structure. The encoded layer has 50×768 as input, which this layer merged into patch embeddings, positional embeddings, and class tokens. In the VIT architecture, the previous layer's inputs and

outputs for the 12 layers are 50×768. Furthermore, the normalization layer normalizes the inputs before they are fed into the Multiheaded Attention (MHA) block. To obtain the query, key, and value matrix in MHA, the input data are adapted into a 50×2304 (768×3) shape using a linear layer. Then, these matrices are reshaped into $50 \times 3 \times 768$, where each one is represented as 50×768. These matrices are then reshaped once more to $12 \times 50 \times 64$. Once these matrices are obtained, the attention process for the MHA block is performed using the following equation:

$$\text{Attention(Query, Key, Value)} = \text{softmax}(\frac{\text{Query} \cdot \text{Key}^{\text{T}}}{\sqrt{\text{d}_{\text{Key}}}}) \cdot \text{Value} \qquad (1)$$

The outputs from the MHSA block are delivered as an input to the skip connection. Then, the outcome of the skip connection is sent to the normalization layer before being delivered to the MLP block for processing. Due to significant advancements in VIT, MLP includes a local mechanism to understand local features [50]. Furthermore, depthwise convolution is integrated into the MLP block during the first fully connected layer to reduce parameters and achieve better results. The output of the MLP block must eventually feed the skip connection to achieve the encoder layer's output.

In this paper, the vision transformer is used because it focuses on each independent patch of the image, as well as their relationships with other patches. In contrast, the convolutional network does not have this property because it uses convolutional filters to learn image features.

3.2. Proposed EL Method

This section describes the implemented DL architecture based on the ensemble learning technique. The objective of the proposed method is to learn and extract medical image representation using three well-known DL models, including MobileNetV2, DenseNet169, and Vision Transformer (VIT). As shown in Figure 1, the input image to the ensemble method is fed to three functional layers simultaneously. At this stage, each functional layer represents a pretrained model that relies on MobileNetV2, DenseNet169, and VIT, respectively. For dimensionality reduction, each functional layer's output (learned representations) is fed to the global average pooling layer. After applying the pooling operation on each parallel flow, the output is flattened and concatenated to generate a single feature vector of each inputted image. To fine-tune the overall network, overcome overfitting, and boost the classification accuracy, a sequential set of layers were placed on top of each other, including batch normalization (BN), fully connected layer (dense), and dropout layer, as shown in Figure 1. The final output of the ensemble method is generated using a fully connected layer to output the classification result.

Using chest X-ray image data sets, the ensemble method was fine-tuned to learn and extract feature vectors from input images of size 224×224. The three models MobileNetV2, DenseNet169, and VIT were pretrained on the ImageNet [51]. In our experiments, the pretrained ensemble method was employed and fine-tuned on the data sets having chest X-ray images. As an output, these models generate a feature vector after flattening of sizes 1280, 1664, and 768, respectively. Thus, the concatenated feature vector is of size 3712. During the fine-tuning of the ensemble method, the weights of the three models were fixed to accelerate the training process.
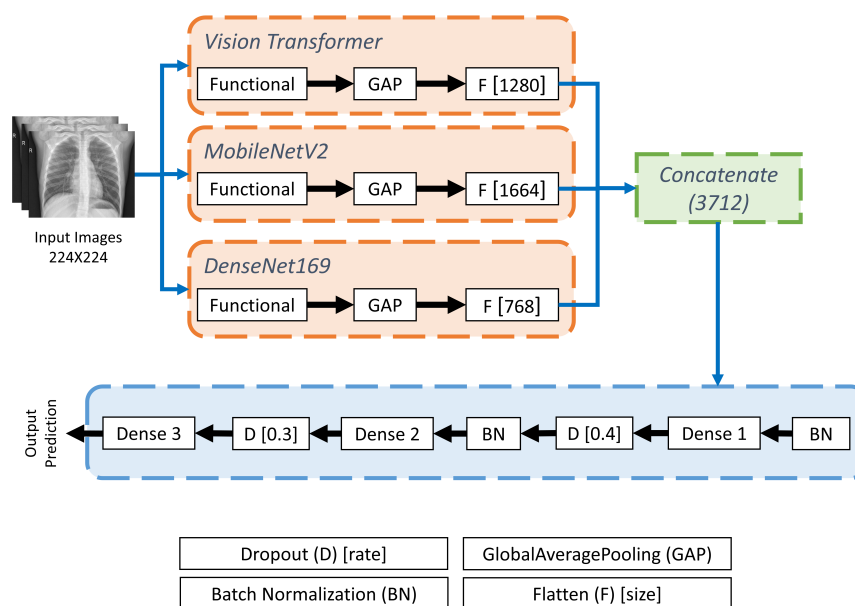
Figure 1. The structure of the proposed ensemble learning method.

4. Experimental Study

This study trained nine well-known CNN methods and the proposed EL method to classify pneumonia in chest X-ray images. In the training phase, different TL and fine-tuning techniques were attempted on these methods, and configurations ensuring excellent outcomes were utilized in the testing stage. A batch size of 32 and a learning rate of 1×10^{-4} were defined during this phase. We used various epoch sizes to train methods, but after 20 epochs, the methods began to overfit. To avoid overfitting of methods, early stopping was used. In addition, the Adam optimizer was applied to reduce the categorical cross-entropy loss function. For classifying, the softmax activation function was applied in the final layer.

As a result, this section describes the experimental study carried out. First, the data sets and the performance measures are portrayed; then, the experimental results and a discussion of them are presented. Finally, we compare our proposed method with state-of-the-art methods.

4.1. Data Set Description

Pneumonia diseases have been verified for our experimental assessment, including the diagnosis of pneumonia from chest X-ray images. Figure 2, for instance, displays chest X-ray images from the selected data set. This data set is split into two sets: normal and pneumonia. The data set was provided by Kermany and Goldbaum [43] and based on a chest X-ray scan images from pediatric patients from one to five years of age at the Guangzhou Women and Children's Medical Center. In the chest X-ray images (Pneumonia) data set, which is publicly available at https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed on 4 January 2022), there are in total 5856 normal and pneumonia chest X-ray images. To provide a fair comparison between our proposed method and the different methods, the training set, the validation set, and the test set were previously divided. The chest X-ray database was divided into two classes (normal and pneumonia). This data set contains two subsets for each class. The training subset consists of 1341 normal patients and 3875 pneumonia patients. Moreover, it contains 234 patients as normal and 390 pneumonia patients for the test subset. These data also consist of 16 validation data images, including eight pneumonia patients and eight normal patients. Samples of normal and pneumonia chest X-ray images can be seen in Figure 2.
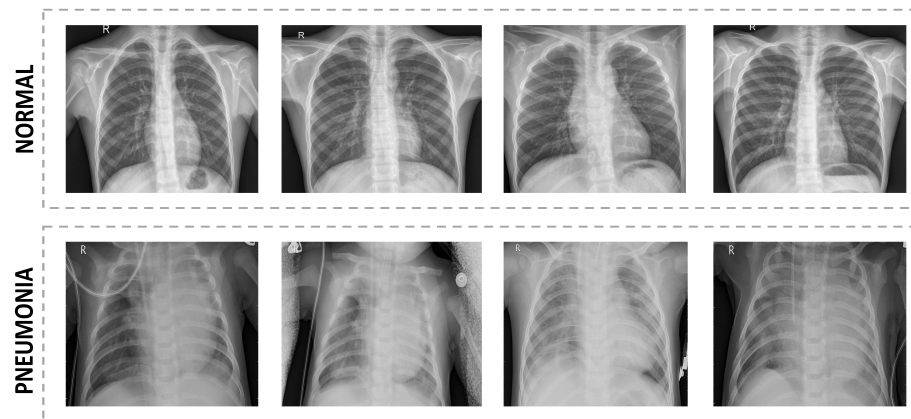
Figure 2. Samples of chest X-ray for pneumonia classification task from the selected database. The above row shows the normal images, and the bottom row shows the pneumonia images.

The performance of the proposed classification method was evaluated based on precision, recall, F1-score, and accuracy, as introduced in Equations (2)–(5), respectively. These metrics are the most popular in medical image classification [26,49]. The precision is measured as the percentage of exact data that conform to specified characteristics. The recall is measured as the percentage of actual statistics to quantities that should have been explicitly anticipated. The F1-score is an indication of imbalanced data between Recall and Precision. The amount produced across all expected amounts is known as accuracy.

4.2. Evaluation Metrics

According to the confusion matrix, the positive term denotes pneumonia, while the negative term denotes normal images. The true term denotes the proper classification, while the false term denotes the wrong classification. The number of normal images wrongly labeled as pneumonia is called False Positive (FP). The number of normal images accurately recognized as normal is referred to as True Negative (TN). The number of normal images wrongly labeled as pneumonia is known as False Negative (FN). The percentage of labels found by the system is measured by the recall. The percentage of labels correctly assigned by the system is measured by precision. For providing the correct results, the F1-score is dependent on precision and recall. From a different perspective, an accuracy metric is used to evaluate the baselines for each task in the two main phases. The system's recognition rate is defined by accuracy.

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

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

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

4.3. Results and Analysis

Initially, we chose the pretrained models from the previous research, as in [26,52]. The best three methods were selected after comparing their accuracy to other methods using the previously mentioned free dataset of chest X-ray images. In terms of testing accuracy, the MobileNetV2, VIT, and DenseNet169 models performed best, as shown in Table 1. Their characteristics are described above in Section 3.1, which also includes a summary of their structures.

Table 1. The results of well-known CNN models.

Model	Precision	Recall	F1-Score	Accuracy
Xception	0.7971	0.7676	0.7713	0.7676
VGG16	0.8126	0.8103	0.8087	0.8103
MobileNetV2	0.9003	0.9073	0.9034	0.9087
InceptionV3	0.8897	0.8871	0.8854	0.8871
ResNet50	0.8233	0.8222	0.8226	0.8222
DenseNet169	0.9133	0.9009	0.9063	0.9135
ResNet152V2	0.8702	0.8687	0.8673	0.8687
DenseNet121	0.8927	0.8922	0.8911	0.8922
VIT	0.9245	0.9247	0.9244	0.9247

The performance of MobileNetV2, DenseNet169, Vision Transformer (VIT), and the proposed ensemble learning (EL) method on training and validation losses and accuracy is compared in Figure 3. The proposed method reduced verification loss, as shown in the figure, which improved the accuracy results. DenseNet169's training loss is 0.1664, training accuracy is 0.9319, validation loss is 0.2408, and validation accuracy is 0.9103. Additionally, MobileNetV2 achieved a training accuracy of 0.9122, a training loss of 0.2096, a validation loss of 0.2072, and a validation accuracy of 0.9087. The VIT had a training loss of 0.1503, a training accuracy of 0.9421, a validation loss of 0.2071, and a validation accuracy of 0.9215. The proposed ensemble method had a training loss of 0.1361, a training accuracy of 0.9525, a validation loss of 0.0421, and a validation accuracy of 1.0.

The test set fulfills the requirements too. In reality, we obtain better accuracy results and a lower loss on the test set with the proposed method, illustrating a more reliable and robust approach. The accuracy of the test set can be seen in Table 2.
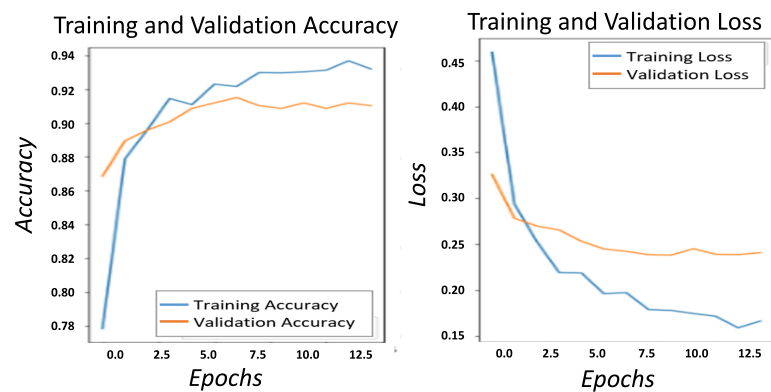
Table 2. Comparison of testing data results among the proposed Ensemble Learning (EL) and three well-known CNN models.

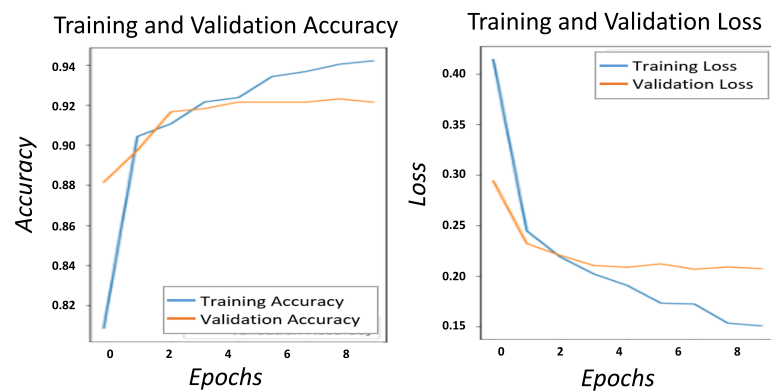| Model	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
DenseNet169	91.33	90.09	90.63	91.35
MobileNetV2	90.03	90.73	90.34	90.87
VIT	92.45	92.47	92.44	92.47
EL (Our)	**93.96**	**92.99**	**93.43**	**93.91**

A comparison of true and predicted labels can be seen in the confusion matrices for the proposed Ensemble Learning (EL) method, Vision Transformer (VIT), MobileNetV2, and DenseNet169, as shown in Figure 4, to better understand how the four approaches did in these binary classifications. This is done primarily to understand what a good classification approach should be, as well as how it could be enhanced while trying to deal with the diagnosis of diseases, which is often critical to a patient's survival. The confusion matrices for the MobileNetV2, DenseNet169, VIT, and the proposed EL method are shown in Figure 4. The confusion matrices in the figure contain actual and predicted labels for both normal (234) and pneumonia (390) on the chest X-ray images.
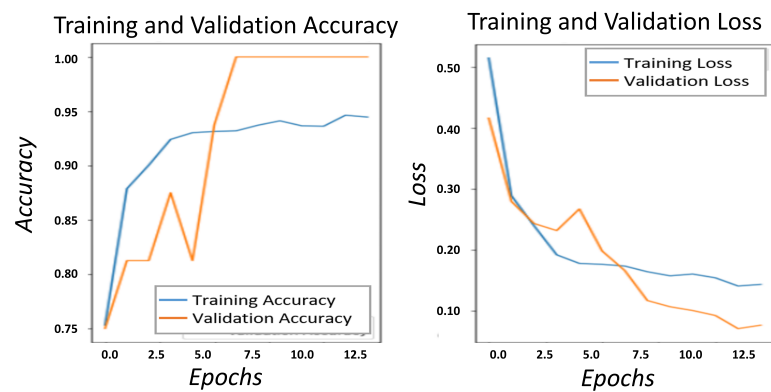
(a) MobileNetV2

(b) DenseNet169

(c) Vision Transformer (VIT)

(d) Ensemble Learning (EL)

Figure 3. The plots of the loss and accuracy on the training and validation sets of the chest X-ray images.
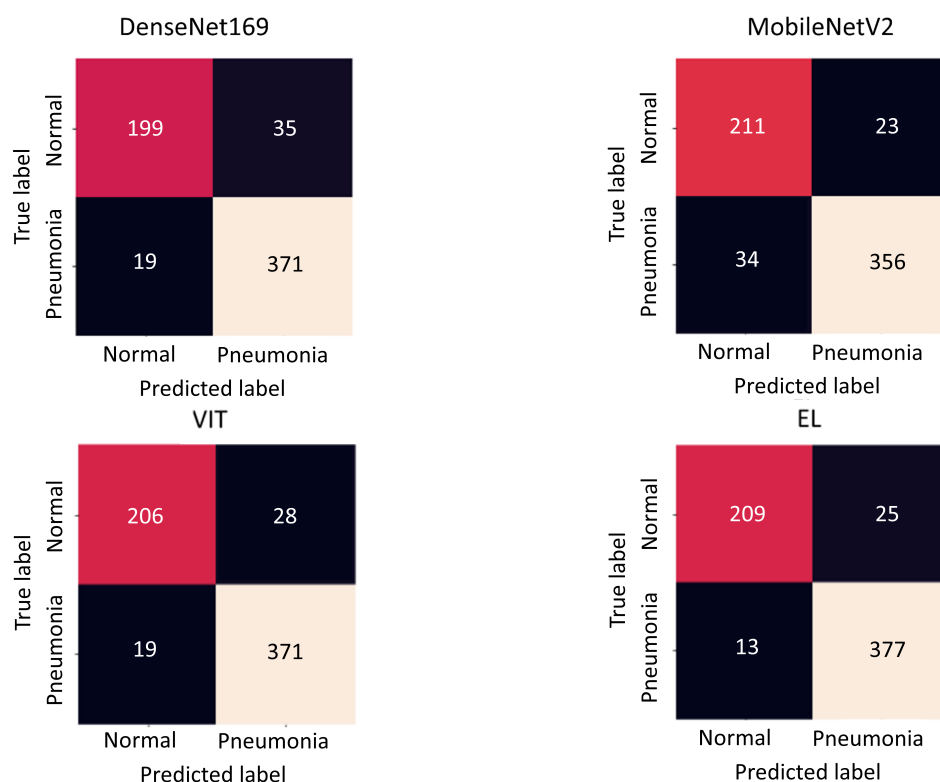
Figure 4. Confusion matrices of the chest X-ray data set.

4.4. Compared Methods

In this section, the proposed methodology is analyzed systematically, and its positive and negative aspects are discussed in comparison with other methods in the literature. The obtained results in this study were compared with other studies that achieved successful results in the literature. This comparison is given in Table 3. The chest X-ray data set was used to compare the various advanced methods for pneumonia detection. The state-of-the-art methods of this data set are discussed as follows:

- Madani et al. [53] examined using Generative Adversarial Networks (GANs) to enrich a data set by producing chest X-ray data samples. GANs offer a method to learn about the underlying architecture of medical images, which can subsequently be used to make high-quality realistic samples.
- Kermany et al. [43] used transfer learning, which allows them to learn a neural network with a portion of the data required by traditional methods. They also made the diagnosis more transparent and understandable by highlighting the neural network's known areas.
- Ayan and Ünver [31] employed two well-known CNN approaches, Xception and Vgg16. In the learning phase, they employed transfer learning and fine-tuning.
- Stephen et al. [28] proposed a CNN-based method. Unlike other methods based solely on transfer learning or traditional handcrafted techniques, they trained the CNN model from scratch to extract attributes from a given chest X-ray image to achieve remarkable classification performance. They used it to determine if a person was infected with pneumonia or not.
- Liang and Zheng [36] performed pneumonia detection with a CNN model architecture using residual connections and dilated convolution methods. They also discovered the transfer learning effect on CNN models when classifying chest X-ray images.
- Salehi et al. [54] proposed an automatic transfer-learning method based on CNN's using DenseNet121 pretrained concepts.

The proposed ensemble method showed better performance than a pretrained CNN model. In addition, designing a CNN model needs massive experiments and knowledge to pretrain the CNN model. According to our test results, the proposed ensemble method was shown to have better performance than a pretrained CNN model. Figure 4 shows a performance comparison of the proposed MobileNetV2, DenseNet169, VIT, and the proposed EL method. In addition, designing a CNN model needs massive experiments and domain knowledge to train a pretrained CNN model with transfer learning. Further, CNN models trained from scratch need more data, more training time, and more epochs to gain better generalization ability on input data.

However, the proposed method suffers from two drawbacks: The first is defining hyperparameters of pretrained CNN methods while applying TL and fine-tuning to a problem of one's own. TL requires determining an appropriate pretrained CNN method for a related issue, the size of fully connected layers, and the number of freezing layers. Many researchers use the trial-and-error approach or their own experiences to identify these parameters. Therefore, finding out TL parameters can reveal lengthy trial-and-error methods. The second drawback of the proposed EL method needs to have a lot of variance and bias.

Table 3. Comparative accuracy results for the state-of-the-art method on test set of the chest X-ray data set. The best results for each item are labeled in bold.

Method/Ref.	Accuracy (%)	Year
DCGAN/[53]	84.19	2018
[43]	92.80	2018
VGG16/[31]	87.00	2019
[28]	93.73	2019
[36]	90.50	2020
DenseNet121/[54]	86.80	2021
EL/Our	**93.91**	**2022**

5. Conclusions

This paper proposes a CNN Ensemble Learning (EL) method for automatically identifying normal and pneumonia patients in chest X-ray images. For this purpose, the three most successful CNN models (DenseNet169, MobileNetV2, and Vision Transformer) were selected for the proposed EL method from among the trained CNN models. The proposed method generated the results during the testing stage. Further, a global average pooling layer was merged after the convolutional layers to avoid losing spatial information in the image. In the classifier stage of the proposed method, fully connected layers were used. As a result, it was found that incorporating these capabilities enhances the classification performance of each CNN model. Consequently, the proposed EL method achieved satisfactory classifier performance using the chest X-ray data set. In future studies, we plan to build a weighted ensemble method based on the CNN model's accuracy.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The selected dataset in this study is available on this link: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed on 4 January 2022).

Conflicts of Interest: The authors declare that there are no conflict of interest regarding the publication of this paper.

References

1. Ortiz-Toro, C.; García-Pedrero, A.; Lillo-Saavedra, M.; Gonzalo-Martín, C. Automatic pneumonia detection in chest X-ray images using textural features. *Comput. Biol. Med.* **2022**, *145*, 105466. [CrossRef]
2. Ben Atitallah, S.; Driss, M.; Boulila, W.; Koubaa, A.; Ben Ghezala, H. Fusion of convolutional neural networks based on Dempster–Shafer theory for automatic pneumonia detection from chest X-ray images. *Int. J. Imaging Syst. Technol.* **2022**, *32*, 658–672. [CrossRef]
3. Wang, L.; Wang, H.; Huang, Y.; Yan, B.; Chang, Z.; Liu, Z.; Zhao, M.; Cui, L.; Song, J.; Li, F. Trends in the application of deep learning networks in medical image analysis: Evolution between 2012 and 2020. *Eur. J. Radiol.* **2022**, *146*, 110069. [CrossRef]
4. Singhal, A.; Phogat, M.; Kumar, D.; Kumar, A.; Dahiya, M.; Shrivastava, V.K. Study of deep learning techniques for medical image analysis: A review. *Mater. Today Proc.* **2022**, *56*, 209–214. [CrossRef]
5. Iori, M.; Di Castelnuovo, C.; Verzellesi, L.; Meglioli, G.; Lippolis, D.G.; Nitrosi, A.; Monelli, F.; Besutti, G.; Trojani, V.; Bertolini, M.; et al. Mortality prediction of COVID-19 patients using radiomic and neural network features extracted from a wide chest X-ray sample size: A robust approach for different medical imbalanced scenarios. *Appl. Sci.* **2022**, *12*, 3903. [CrossRef]
6. Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [CrossRef]
7. Kale, S.P.; Patil, J.; Kshirsagar, A.; Bendre, V. Early Lungs Tuberculosis Detection Using Deep Learning. In *Intelligent Sustainable Systems*; Springer: New York, NY, USA, 2022; pp. 287–294.
8. Bellens, S.; Probst, G.M.; Janssens, M.; Vandewalle, P.; Dewulf, W. Evaluating conventional and deep learning segmentation for fast X-ray CT porosity measurements of polymer laser sintered AM parts. *Polym. Test.* **2022**, *110*, 107540. [CrossRef]
9. Zhang, L.; Mueller, R. Large-scale recognition of natural landmarks with deep learning based on biomimetic sonar echoes. *Bioinspir. Biomimetics* **2022**, *17*, 026011. [CrossRef]
10. Le Dinh, T.; Lee, S.H.; Kwon, S.G.; Kwon, K.R. COVID-19 Chest X-ray Classification and Severity Assessment Using Convolutional and Transformer Neural Networks. *Appl. Sci.* **2022**, *12*, 4861. [CrossRef]
11. Sajun, A.R.; Zualkernan, I.; Sankalpa, D. Investigating the Performance of FixMatch for COVID-19 Detection in Chest X-rays. *Appl. Sci.* **2022**, *12*, 4694. [CrossRef]
12. Furtado, A.; Andrade, L.; Frias, D.; Maia, T.; Badaró, R.; Nascimento, E.G.S. Deep Learning Applied to Chest Radiograph Classification—A COVID-19 Pneumonia Experience. *Appl. Sci.* **2022**, *12*, 3712. [CrossRef]
13. Malhotra, P.; Gupta, S.; Koundal, D.; Zaguia, A.; Kaur, M.; Lee, H.N. Deep Learning-Based Computer-Aided Pneumothorax Detection Using Chest X-ray Images. *Sensors* **2022**, *22*, 2278. [CrossRef]
14. Abd Elaziz, M.; Mabrouk, A.; Dahou, A.; Chelloug, S.A. Medical Image Classification Utilizing Ensemble Learning and Levy Flight-Based Honey Badger Algorithm on 6G-Enabled Internet of Things. *Comput. Intell. Neurosci.* **2022**, *2022*, 5830766. [CrossRef]
15. Adel, H.; Dahou, A.; Mabrouk, A.; Abd Elaziz, M.; Kayed, M.; El-Henawy, I.M.; Alshathri, S.; Amin Ali, A. Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm. *Mathematics* **2022**, *10*, 447. [CrossRef]
16. Niu, S.; Liu, M.; Liu, Y.; Wang, J.; Song, H. Distant domain transfer learning for medical imaging. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3784–3793. [CrossRef]
17. Lee, H.C.; Aqil, A.F. Combination of Transfer Learning Methods for Kidney Glomeruli Image Classification. *Appl. Sci.* **2022**, *12*, 1040. [CrossRef]
18. Mabrouk, A.; Redondo, R.P.D.; Kayed, M. SEOpinion: Summarization and Exploration of Opinion from E-Commerce Websites. *Sensors* **2021**, *21*, 636. [CrossRef]
19. Chandrasekaran, G.; Antoanela, N.; Andrei, G.; Monica, C.; Hemanth, J. Visual Sentiment Analysis Using Deep Learning Models with Social Media Data. *Appl. Sci.* **2022**, *12*, 1030. [CrossRef]
20. Mabrouk, A.; Redondo, R.P.D.; Kayed, M. Deep learning-based sentiment classification: A comparative survey. *IEEE Access* **2020**, *8*, 85616–85638. [CrossRef]
21. 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.

22. Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv* **2014**, arXiv:1404.1869.

23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

24. Maselli, G.; Bertamino, E.; Capalbo, C.; Mancini, R.; Orsi, G.; Napoli, C.; Napoli, C. Hierarchical convolutional models for automatic pneu-monia diagnosis based on X-ray images: New strategies in public health. *Ann. IG* **2021**, *33*, 644–655.

25. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**. [CrossRef]

26. Ayan, E.; Karabulut, B.; Ünver, H.M. Diagnosis of Pediatric Pneumonia with Ensemble of Deep Convolutional Neural Networks in Chest X-ray Images. *Arab. J. Sci. Eng.* **2022**, *47*, 2123–2139. [CrossRef]

27. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv* **2017**, arXiv:1711.05225.

28. Stephen, O.; Sain, M.; Maduh, U.J.; Jeong, D.U. An efficient deep learning approach to pneumonia classification in healthcare. *J. Healthc. Eng.* **2019**, *2019*, 4180949. [CrossRef]

29. Wu, H.; Xie, P.; Zhang, H.; Li, D.; Cheng, M. Predict pneumonia with chest X-ray images based on convolutional deep neural learning networks. *J. Intell. Fuzzy Syst.* **2020**, *39*, 2893–2907. [CrossRef]

30. Cheplygina, V.; de Bruijne, M.; Pluim, J.P. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal.* **2019**, *54*, 280–296. [CrossRef]

31. Ayan, E.; Ünver, H.M. Diagnosis of pneumonia from chest X-ray images using deep learning. In Proceedings of the 2019 IEEE Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 24–26 April 2019; pp. 1–5.

32. Chouhan, V.; Singh, S.K.; Khamparia, A.; Gupta, D.; Tiwari, P.; Moreira, C.; Damaševičius, R.; De Albuquerque, V.H.C. A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Appl. Sci.* **2020**, *10*, 559. [CrossRef]

33. Rahman, T.; Chowdhury, M.E.; Khandakar, A.; Islam, K.R.; Islam, K.F.; Mahbub, Z.B.; Kadir, M.A.; Kashem, S. Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. *Appl. Sci.* **2020**, *10*, 3233. [CrossRef]

34. Toğaçar, M.; Ergen, B.; Cömert, Z.; Özyurt, F. A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models. *IRBM* **2020**, *41*, 212–222. [CrossRef]

35. Mittal, A.; Kumar, D.; Mittal, M.; Saba, T.; Abunadi, I.; Rehman, A.; Roy, S. Detecting pneumonia using convolutions and dynamic capsule routing for chest X-ray images. *Sensors* **2020**, *20*, 1068. [CrossRef]

36. Liang, G.; Zheng, L. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Comput. Methods Programs Biomed.* **2020**, *187*, 104964. [CrossRef]

37. Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; Feng, J. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3435–3444.

38. Li, P.; Yang, Y.; Grosu, R.; Wang, G.; Li, R.; Wu, Y.; Huang, Z. Driver Distraction Detection Using Octave-Like Convolutional Neural Network. *IEEE Trans. Intell. Transp. Syst.* **2021**. [CrossRef]

39. das Neves, R.B.; Verçosa, L.F.; Macêdo, D.; Bezerra, B.L.D.; Zanchettin, C. A fast fully octave convolutional neural network for document image segmentation. In Proceedings of the 2020 IEEE International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 18–23 June 2020; pp. 1–6.

40. Wang, B.; Yang, J.; Ai, J.; Luo, N.; An, L.; Feng, H.; Yang, B.; You, Z. Accurate tumor segmentation via octave convolution neural network. *Front. Med.* **2021**, *8*, 501. [CrossRef]

41. Mahmoudi, R.; Benameur, N.; Mabrouk, R.; Mohammed, M.A.; Garcia-Zapirain, B.; Bedoui, M.H. A Deep Learning-Based Diagnosis System for COVID-19 Detection and Pneumonia Screening Using CT Imaging. *Appl. Sci.* **2022**, *12*, 4825. [CrossRef]

42. Chen, P.Y.; Zhang, X.H.; Wu, J.X.; Pai, C.C.; Hsu, J.C.; Lin, C.H.; Pai, N.S. Automatic Breast Tumor Screening of Mammographic Images with Optimal Convolutional Neural Network. *Appl. Sci.* **2022**, *12*, 4079. [CrossRef]

43. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [CrossRef]

44. Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Appl. Sci.* **2018**, *8*, 1715. [CrossRef]

45. Ali, R.; Chuah, J.H.; Talip, M.S.A.; Mokhtar, N.; Shoaib, M.A. Structural crack detection using deep convolutional neural networks. *Autom. Constr.* **2022**, *133*, 103989. [CrossRef]

46. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

47. Rahman, Z.; Hossain, M.S.; Islam, M.R.; Hasan, M.M.; Hridhee, R.A. An approach for multiclass skin lesion classification based on ensemble learning. *Inform. Med. Unlocked* **2021**, *25*, 100659. [CrossRef]

48. Gupta, K.D.; Sharma, D.K.; Ahmed, S.; Gupta, H.; Gupta, D.; Hsu, C.H. A Novel Lightweight Deep Learning-Based Histopatho-logical Image Classification Model for IoMT. *Neural Process. Lett.* **2021**, 1–24. [CrossRef]

49. Kassani, S.H.; Kassani, P.H.; Wesolowski, M.J.; Schneider, K.A.; Deters, R. Classification of histopathological biopsy images using ensemble of deep learning networks. *arXiv* **2019**, arXiv:1909.11870.

50. Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; Van Gool, L. Localvit: Bringing locality to vision transformers. *arXiv* **2021**, arXiv:2104.05707.
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
52. Pérez, E.; Ventura, S. An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis. *Neural Comput. Appl.* **2021**, 1–20. [CrossRef]
53. Madani, A.; Moradi, M.; Karargyris, A.; Syeda-Mahmood, T. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In Proceedings of the Medical Imaging 2018: Image Processing, Houston, TX, USA, 11–13 February 2018; Volume 10574, p. 105741M.
54. Salehi, M.; Mohammadi, R.; Ghaffari, H.; Sadighi, N.; Reiazi, R. Automated detection of pneumonia cases using deep transfer learning with paediatric chest X-ray images. *Br. J. Radiol.* **2021**, *94*, 20201263. [CrossRef]