STCovidNet: Automatic Detection Model of Novel Coronavirus Pneumonia Based on Swin Transformer

Boyuan Wang
Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau
Du Zhang (duzhang@must.edu.mo)
Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau
Zonggui Tian
Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau

Research Article

Keywords: COVID-19 detection, Swin Transfomer, Vision Transformer, Convolutional Neural Networks, Feature Visualization

Posted Date: March 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1401026/v1

Abstract

The novel coronavirus disease 2019 (COVID-19) has emerged as an enormous challenge facing China today. Preventive Medicine physicians and Artificial Intelligence (AI) researchers try to improve the ability to early automatic warning of coronavirus infections, promote epidemic prevention, and reduce medical costs using deep learning methods. In this work, we build an extensive database of chest computed tomography (CT) scans with image data from domestic and international open-source medical datasets. Swin Transformer is chosen as the backbone network to establish a model (STCovidNet) for the prediction of COVID-19. We then compare the performance of our technique against that of Vision Transformer (ViT) and Convolutional Neural Network (CNN). Next, to visualize our model's high-dimensional outputs in 2-dimensional space, we apply t-distributed stochastic neighbor embedding (t-SNE) as the dimension-reduction strategy. Finally, we employ gradient-weighted class activation mapping (Grad-CAM) to present a class activation map. The results indicate that STCovidNet's performance surpasses ViT and CNN with a 0.9811 AUC and 0.9858 accuracy score. Our network outperforms previous techniques to reduce intra-class variability and generate well-separated feature embedding. The CAM figure illustrates that the decision region corresponds to radiologists' detecting spots. The suggested method can be an effective way of catching COVID-19 instances.

Introduction

COVID-19 is now the most prevalent respiratory infectious disease in the 21st century [1]. By February 24, 2022, it had infected more than 430,220,905 individuals and caused the deaths of 5,936,914 patients worldwide [2]. Because of its rapid mutation and the many powerful immune escape variants that have been produced, such as Delta and Omicron, testing them is a daunting task for today's public health workers. Nevertheless, the real-time reverse transcription-polymerase chain reaction (RT-PCR) has a roughly 1/3 false-negative rate, necessitating repetitive testing to decrease incorrect diagnoses [3], [4]. Another important detection technique is the Chest CT scan, which improves sensitivity in diagnosing COVID-19 instances [6], [7]. The main manifestations of chest CT are ground-glass shadows, pulmonary consolidation, and leaving stone signs after SARS-CoV-2 virus infection, which are the most frequent radiological manifestation and critical diagnostic criteria for COVID-19 cases [5]. Combined with RT-PCR results, clinical symptoms, and epidemiological history, these findings are the sole basis for diagnosing or excluding COVID-19 pneumonia.

Therefore, to achieve automatic early warning of COVID-19, some studies have attempted to develop models to automatically identify patients with COVID-19 by learning lesion characteristics through artificial intelligence technology. Most of these studies mainly used convolutional neural networks (CNN) to automatically find COVID-19 patients on chest CT images [6]. Though CNN has demonstrated its ability to solve a variety of classification issues, it is not the best option for problems that require high-level categorization, where global features such as patterns, multiplicity, and distribution must be taken into account [7]. Recently, some studies based on the Vision Transformer (ViT) [8] architecture have been published to solve the problem of the receptive field through the attention mechanism, obtaining better

classification results than CNN in the field of CT image classification [9]. Motivated by this, we conduct this study with the following significant contributions:

1. We establish a dataset named MUST-COVID-19, consisting of 7930 chest CT images, which were preprocessed with image cropping and scaling. We applied various data enhancement methods to reduce the chance of model overfitting.

2. The Swin Transformer [10] was then chosen as the backbone network to establish a novel model (STCovidNet) for COVID-19 case detection.

3. An evaluation of the effectiveness of STCovidNet against the state-of-the-art (SOTA) CNN and ViT models is performed on the MUST-COVID-19 dataset. Our experimental findings demonstrate the efficacy of our technique for achieving the best performance on the CT image datasets under consideration.

To our knowledge, this is the first publication that evaluates and compares the Swin Transformer's classification performance on the COVID-19 pneumonia dataset with other models, providing a framework for medical experts to choose an excellent COVID-19 detection model and filling a research need.

The remainder of this article is structured in the following manner. Section 2 presents the details of related studies. The sources and construction methods for training, verifying, and testing data sets are provided in Section 3. Section 4 describes our proposed approach, STCovidNet architecture, cornerstones of t-SNE and Grad-CAM to visualize the model's high-dimensional outputs and class activation map, and performance evaluation metrics. Model parameter settings for the experimental study are discussed in Section 5. The experimental results, comparison with related models, and benefits of the proposed model are detailed in Section 6. Finally, Section 7 provides a conclusion with comments on future work.

Related Works

This section extensively reviews the primary research methods used for the current COVID-19 cases detection. CNN is the most often used approach to solve the challenge of automated COVID-19 cases diagnosis [11], [12]. The deep learning frameworks in the previous studies are primarily based on the pre-trained networks, including variants of Very Deep Convolutional Networks (VGGNet) [13], Deep Residual Neural Networks (ResNet) [14], Dense Convolutional Network (DenseNet) [15], Inception [16], Xception [17], MobileNet [18] and EfficientNet [19]. These models adapt to the new task of COVID-19 patients' detection and classification by modifying or adding custom layers and making use of the knowledge gained from previous experience based on Transfer Learning. For example, Brunese et al. [20] proposed two models using the VGG-16 network as a backbone model based on Transfer Learning. The first network is used to identify whether the target is healthy or getting pneumonia. Once the first network has a positive prediction result, the second network is used to find COVID-19. The VGG-16 network attained 98% accuracy for the three-class classification. ResNet is another common architecture in CNN that prevents gradient disappearance problems compared to earlier architectures such as VGG. Using the

residual network, Narin et al. [21] classified COVID-19 cases and healthy with ResNet-50, achieving the highest accuracy (98%) for binary classification. Other studies have used more efficient architectures such as DenseNet and EfficientNet. Wang et al. [22] developed a COVID-19 pneumonia classification pipeline using DenseNet-121. The proposed approach achieved an AUC with an overall performance of 0.88–0.99 in different datasets. Shamila et al. [23] applied the EffectiveNet architecture to establish a classification model with 95% accuracy and 93% F1-score on the test set.

While CNN is suited for image classification in deep learning, they have some conceptual limitations. In CNN, information about the location of entities is lost during the maximum pooling operation. In addition, CNN does not consider some spatial relationships between simple objects. They need a vast receptive field to capture long-range dependencies, which means developing large kernels or highly massive networks, resulting in an extremely complex model that is challenging to train. [9]. To overcome these drawbacks of CNN, some researchers have used other architectures, such as Capsule Neural Networks (Capsnets) [24] and ViT [8], for COVID-19 classification, which differs from the traditional CNN networks. Sabour et al. proposed Capsnets [24], which are the new architecture in neural networks to resolve the disadvantage of CNN in not using location and orientation information to perform recognition of objects [25]. Toraman et al. [26] suggested a five-convolutional-layer Capsnets model. The 4 layers contain 16, 32, 64, and 128 kernels, respectively, and the 5th layer includes 32 capsules. After 10-fold cross-validation and 50 epochs of training, the model's results are evaluated to reach 84.22% accuracy for the multi-classification.

The latest research is based on Transformer [27]. Dosovitskiy et al. [8] applied the standard Transformer architecture to image recognition. They proposed ViT based on the Self-attention to approach or exceed the SOTA model in several image recognition benchmarks. Some new COVID-19 detection algorithm based on the ViT architecture has been proposed in a few research projects. Shome et al. [28] built a dataset of 30,000 images and trained the ViT model on it. The trained model performed better than CNN, such as EfficientNet-B0, Inception-V3, and ResNet-50 in a multi-classification challenge, with 92% accuracy and 98% AUC. Mondal et al. [9] suggested a network based on the ViT-B/16 architecture and achieved the highest 98.1% accuracy, exceeding most existing methods.

Data description

Must-covid-19 Dataset

We establish a chest CT scan dataset named MUST-COVID-19 for this research, consisting of 7930 chest CT images. To make the results representative, our data are randomly sampled from a dataset consisting of eight open-source chest CT image sets, namely, (1) CNCB 2019 Novel Coronavirus Resource AI Diagnosis Dataset, which comes from a total of 2778 patients in the dataset of CC-CCII [29], including 917 COVID-19 pneumonia cases, 878 normal, and 983 none-COVID-19 pneumonia cases in the training set; (2) iCTCF [30], which comes from a total of 1521 patients in two hospitals of Huazhong University of Science and Technology, China, including 894 COVID-19 pneumonia cases (including mild, severe, and critical cases), 328 novel coronavirus-negative patients (control group), and 299 patients with suspected

COVID-19; (3) COVID-CTSet [31], which comes from the dataset of Negin Medical Center in Sari, Iran, including 377 patients with confirmed COVID-19, 95 novel coronavirus-negative patients, and 282 other pneumonia patients; and the remaining were collected from (4) TCIA [32], (5) COVID-19 Infection Segmentation Dataset [33], (6) LIDC-IDRI [34], (7) Radiopaedia [35], (8) and MosMedData [36]. In Table 1, MUST-COVID-19 contains images of about three classes, with 80 percent of images employed for training and verification and 20 percent for model testing.

Table 1
Training, validation, and test sets of MUST-COVID-19

Class	Dataset			
	Training	Validation	Testing	Total
COVID-19	1931	277	552	2760
non-COVID-19 Pneumonia	1762	252	504	2518
Normal	1856	265	531	2652
Total	5549	794	1587	7930

Image Pre-processing

The pixels outside the red bounding box have no value for diagnosing COVID-19 pneumonia [37], as illustrated in Fig. 1 (a). Therefore, to remove the irrelevant parts, we crop the images of MUST-COVID-19 to the body region using these bounding boxes. Figure 2 (b) shows the example images after image cropping and scaling.

Methods

Stcovidnet Architecture

Figure 2 (a) illustrates the architecture of STCovidNet. The backbone of STCovidNet is the Swin Transformer, and it is based on the transfer learning principle [38], [39]. Swin Transformer [10] is a subcategory of the ViT, which is created to be suitable for detection by introducing the idea of a hierarchical feature map and shifted windows to ViT. A detailed architecture for connecting two Swin Transformer blocks is shown in Fig. 2(b). A LayerNorm (LN) layer, a typical windows-based with multi-head and self-attention (W-MSA) module, a shifting window-based with multi-head and self-attention (SW-MSA) module, and Multi-Layer Perceptron (MLP) layers are used for each Swin Transformer block. S(W)-MSA has a LayerNorm (LN) in front and behind, and the final MPL has two GELU non-linearities. As a result, the Swin Transformer block has been deployed in groups of two [40]. The connections of the Swin Transformer blocks can be represented using Equations (1) to (4):

$$\hat{a}^{k} = W - MSA\left(LN\left(a^{k-1}\right)\right) + a^{k-1},$$

1

$$a^{\mathrm{k}} = MLP\left(LN\left(\hat{a}^{\mathrm{k}}\right)\right) + \hat{a}^{\mathrm{k}},$$

2

$$\hat{a}^{\mathrm{k}+1} = SW-MSA\left(LN\left(a^{\mathrm{k}-1}\right)\right) + a^{\mathrm{k}},$$

3

$$a^{\mathrm{k}+1} = MLP\left(LN\left(\hat{a}^{\mathrm{k}+1}\right)\right) + \hat{a}^{\mathrm{k}+1},$$

4

where $\hat{a}^{\mathrm{k}}$ is (S)W-MSA's result, a^{k} is MLP's result, and k denotes the Swin Transformer block position.

The Swin Transformer provides four versions of the model, which, from small to large, are the Swin-Tiny (Swin-T), Swin-Small, Swin-Base (Swin-B), Swin-Large [10]. This research presents Swin-T as a backbone network, considering the performance and computational complexity, and it has been pre-trained on ImageNet-21k [41].

In Fig. 2 (a), the initial of the STCovidNet framework is the data augmentation. Augmentation approaches including random rotation, random horizontal flip, random crop, random blur, random salt pepper noise, and random Gaussian noise are used to improve data representativeness, reduce over-fitting, and develop a more generic model. To further boost the randomization of the operations, the random order command is used to disrupt the order of all the previous transform operations.

After passing through the augmentation, the input CT image with a size of 244 × 244 passes through the patch partition layer. It is segmented into patches with a 4 × 4 size to generate patch tokens having a shape of ($\frac{224}{4}$, $\frac{224}{4}$, 4 × 4 × channel). The generated patch token is linearly embedded in the first stage, and the patch token of ($\frac{224}{4}$, $\frac{224}{4}$, 48) is projected to the dimension of C (C = 96) to generate tokens of ($\frac{224}{4}$, $\frac{224}{4}$, C). They are then input into several Swin Transformer blocks. The first two Swin Transformer blocks keep the shape of input and output tokens constant at ($\frac{224}{4}$, $\frac{224}{4}$, C) and are designated as Stage 1 together with the linear embedding layer. Stages 2, 3, and 4 consist of patch merging and Swin Transformer blocks, respectively. As the network deepens, the shape of tokens is gradually reduced by patch merging. In patch merging, adjacent 2 × 2 patches are merged into one patch, and the tokens are down-sampled to 1/2, whereas C is doubled. In stages 2, 3, and 4, the output shape of tokens is ($\frac{W}{8}$, $\frac{H}{8}$, 2C) and ($\frac{W}{16}$, $\frac{H}{16}$, 4C), ($\frac{W}{32}$, $\frac{H}{32}$, 8C), respectively. After stage 4, we end up with $\frac{224}{32}$×$\frac{224}{32}$, i.e., 7×7 tokens each of an embedding size of 768 dimensions.

The last layer of STCovidNet is an average pooling followed by a Norm layer. The CT image has been successfully converted into one representation with 768 embeddings. A new classification head for the target domain MUST-COVID-19 is attached to convert these 768 embeddings into the 3 dimensional to finally obtain the predicted results.

SOTA ViT and CNN Models

As a comparison, we used the following SOTA model:

ViT [8] was implemented in computer vision applications by employing multi-head self-attention [8], [10], [42], [43] as an image extraction approach, following the recent breakthrough of Transformers [27] in handling natural language processing tasks [44]. ViT mainly consists of the following parts: Linear Projection of Flattened Patches (Embedding layer), Transformer Encoder, and MLP Head. Figure 3 illustrates the ViT model.

The principle of ViT is first to divide the input into patches and then reshape each patch into a vector to get a flattened patch. The input image is the H×W×C size, and then ViT gets N patches by dividing the picture with a P×P patch. The shape of each patch is P×P×C, which is converted into vectors to get P2×C-dimensional vectors. These vectors are eventually integrated to obtain a two-dimensional matrix of N × (P2 × C), similar to the Word Vectors in Natural Language Processing. The formula of input sequence z_0 of ViT is shown in Eq. (6).

$$z_0 = \left[\mathrm{x}_{class}; \mathrm{x}_p^1\mathrm{E}; \mathrm{x}_p^2\mathrm{E}; \ldots; \mathrm{x}_p^N\mathrm{E}\right] + \mathrm{E}_{pos}$$

$$\mathrm{E} \in \mathrm{R}^{\left(p^2C\right) \times D}, \mathrm{E}_{pos} \in \mathrm{R}^{(N+1) \times D}$$

6

where x represents an image block, the formula for ViT is shown in (7) to (9).

$$\hat{a}^{\mathrm{i}} = MSA\left(LN\left(a^{\mathrm{k}-1}\right)\right) + a^{\mathrm{k}-1}$$

7

$$a^{\mathrm{i}} = MLP\left(LN\left(\hat{a}^{\mathrm{k}}\right)\right) + \hat{a}^{\mathrm{k}}$$

8

$$\mathrm{z} = LN\left(\hat{a}^{\mathrm{k}}\right)$$

9

where z is the output value of ViT. ViT consists primarily of multiple self-attention (MSA) and MLP, with LayerNorm (LN) and residual connections added before MSA and MLP in Fig. 4.

CNN is a feed-forward neural network capable of representing learning from images automatically. [45]. It extracts features with diagnostic values from medical images to realize the disease's automatic classification and detection [46], [47]. The layer transition occurs when CNN connects the i^{th} layer's result as input to the $(i+1)^{th}$ layer [48], [49], as shown in (10):

$$z_i = T_i\left(z_{i-1}\right)$$

10

where i is the network layer's index, $T_i(\bullet)$ is the nonlinear function that includes convolutional computation, pooling and batch normalization, etc., z_i is the i^{th} layer's result. The general framework of CNN is shown in Fig. 5.

After the breakthrough of AlexNet [48], a variant of CNN, researchers constructed various models such as Deep residual networks (ResNet) [14], DenseNet [49], Xception [17], EfficientNet [19], etc., which were trained on the ImageNet database, and obtained better classification results.

ResNet [14] is a series of extremely deep convolutional networks which includes a skip link that uses an identity function to eliminate exploding or vanishing gradients problems, as shown in (11):

$$z_i = T_i\left(z_{i-1}\right) + z_{i-1}$$

11

The activation of a layer is directly applied to the activation of other layers further in the network by employing skip connections, allowing gradients to flow straight from the later layers to the earlier ones. This aids in the generation of deeper CNNs while retaining accuracy. In this study, we employ ResNet-50 [14], a typical ResNet variation, and its design in Fig. 6.

DenseNet, developed by Huang et al. [49], is designed to achieve more excellent anti-fitting properties. DenseNet extends ResNet's shortcut connections by connecting all levels; each layer z_i receives all preceding layers, $z_0, \ldots, z_{i-1}$, as its new input to guarantee that the most inter-layer information is conveyed, as shown in (12):

$$z_i = T_i\left([z_0, \ldots, z_{i-1}]\right)$$

12

The use of dense connections helps reduce the problem of overfitting in networks with limited datasets [50]. Figure 7 depicts a three-layer dense block in which each layer executes batch normalization, ReLU activation, and convolution processes.

EfficientNet [19] is a series of models from B0 to B7 obtained by Google following a multi-objective neural network search (NAS) approach. Based on the recombination coefficient, EfficientNet scales the three dimensions with the formula shown in Eq. (13).

$$\textbf{depth}\ \mathrm{d} = \rho^{\varphi}$$

$$\textbf{width}: \mathrm{w} = \sigma^{\varphi}$$

$$\textbf{resolution}: \mathrm{r} = \tau^{\varphi}$$

$$\text{s. t. } \rho \bullet \sigma^2 \bullet \tau^2 \approx 2$$

$$\rho \geq 1,\ \sigma \geq 1,\ \tau \geq 1$$

13

where ϕ is a composite coefficient, ρ is the scaling factor for depth, σ is the scaling factor for width, and τ is the scaling factor for resolution. Three scaling coefficients are determined by the grid search method, upon which the B0 model is scaled to generate a series of required models. The framework of EfficientNet mainly consists of mobile inverted bottleneck convolution (MBConv) [18], [19]. The structure of EfficientNet-B0 is shown in Fig. 8(a), with a 224×224 resolution of the input images. The 1st part is a convolutional operator, and the 2nd part completes feature extraction by 16 MBConv operators. The third part consists of the convolution, global average pooling, and classification layer.

t-SNE

t-SNE [53] is an effective method of scaling down high-dimensional data to explore the distribution of features generated by models [54]:

Suppose X is a vector containing all samples, and Y is a target vector of a low-dimensional representation of X. $P_{j|i}$ is a conditional probability in the original high-dimensional space to describe the similarity of data point x_j to data point x_i[53], as shown in Eq. (14):

$$P_{j|i} = \frac{exp\left(\frac{-||x_i - x_j||^2}{2\sigma_i^2}\right)}{\sum_{r \neq s} exp\left(\frac{-||x_r - x_s||^2}{2\sigma_i^2}\right)}$$

14

As a result, in the original space, the probabilities may be stated as follows:

$$P_{j|i} = \frac{(P_{j|i} + P_{i|j})}{2n}$$

15

The size of the data collection is denoted by the number *n*. The probability at low dimension Q_{ij} is calculated using this distribution [55], as illustrated in the equation below.

$$Q_{ij} = \frac{\left(1 + ||\mathrm{y}_i - \mathrm{y}_j||^2\right)^{-1}}{\sum_{r \neq s} \left(1 + ||\mathrm{y}_r - \mathrm{y}_s||^2\right)^{-1}}$$

16

Using the Kullback–Leibler divergence [56] as a loss function and a gradient-based algorithm, t-SNE then determines the projections of x_i in lower dimension as y_i:

$$KL(P||Q) = \sum_i \sum_j P_{ij} log \frac{P_{ij}}{Q_{ij}}$$

17

Grad-CAM

Grad-CAM [57] generates a class-discriminative localization map based on the gradient value, emphasizing critical regions in images and offering an interpretable perspective of models. The estimation equation for the class-discriminative localization map $\mathrm{L}^{\mathrm{c}}_{\mathrm{Grad-CAM}}$ is shown in Eq. (18):

$$\mathrm{L}^{\mathrm{c}}_{\mathrm{Grad-CAM}}(\mathrm{x}, \mathrm{y}) = ReLU\left(\sum_{\mathrm{k}} \alpha^{\mathrm{c}}_{\mathrm{k}} \mathrm{A}^{\mathrm{k}}(\mathrm{x}, \mathrm{y})\right)$$

18

where A is the feature map activations operator, c is the target class of the model, $\alpha^{\mathrm{c}}_{\mathrm{k}}$ is the network's partial linearization downstream of A [57]. The calculation equation for $\alpha^{\mathrm{c}}_{\mathrm{k}}$ is shown in Eq. (19).

$$\alpha^{\mathrm{c}}_{\mathrm{k}} = \frac{1}{\mathrm{Z}} \sum_{\mathrm{i}} \sum_{\mathrm{j}} \frac{\partial \mathrm{Y}^{\mathrm{c}}}{\partial \mathrm{A}^{\mathrm{k}}_{\mathrm{nm}}}$$

19

where A^k is the feature map of the k^{th} layer and n, m are the locations in the map. Y^c is the gradient of the score for class c, and $\frac{\partial Y^c}{\partial A^k}$ is the gradient of Y^c.

Performance evaluation metrics

In this research, we use multiple evaluation metrics to assess the model:

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

20

$$Precision = \frac{TP}{TP + FP}$$

21

$$Sensitivity = \frac{TP}{TP + FN}$$

22

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN}$$

23

The average performance was calculated by the macro average and weighted average [58].

Experimental Setup

The suggested STCovidNet model, as well as the following SOTA models, are used in this research: (1) ViT-B/32 (base size model), (2) ViT-L/32 (large size model), (3) ResNet-50, (4) DenseNet-201, and (5) Efficientnet-B4, which are pre-trained on ImageNet-21k. Each model has been trained in a maximum of 30 epochs using the Adam optimizer with train batch size: 16, test batch size: 8, and initial learning rate: 3e-5. We used Python as our programming language, and all experimentation was conducted with NVIDIA CUDA GPU 11.0 using a Tesla P100-16GB. We also use the PyTorch 1.9.1 deep learning library (https://pytorch.org/) and the PyTorch image model library (https://fastai.github.io/timmdocs/) to carry out experiments on preformed deep learning models.

Table 2
STCovidNet vs. SOTA techniques on MUST-COVID-19.

Model	Class Label	Evaluation Metrics				
		Precision	**Sensitivity**	**F1-score**	**Accuracy**	**AUC**
DenseNet-201	COVID-19	0.8883	0.8641	0.8760	0.9055	0.9290
	Normal	0.9247	0.9718	0.9477		
	Pneumonia	0.9024	0.8810	0.8916		
	Macro average	0.9055	0.9055	0.9055		
	Weighted average	0.9051	0.9056	0.9051		
Efficientnet-B4	COVID-19	0.7715	0.8442	0.8062	0.8381	0.8778
	Normal	0.8998	0.8456	0.8718		
	Pneumonia	0.8574	0.8234	0.8401		
	Macro average	0.8381	0.8381	0.8381		
	Weighted average	0.8429	0.8377	0.8394		
ResNet-50	COVID-19	0.8812	0.6449	0.7448	0.8311	0.8743
	Normal	0.8612	0.9699	0.9123		
	Pneumonia	0.7658	0.8889	0.8228		
	Macro average	0.8311	0.8311	0.8311		
	Weighted average	0.8361	0.8346	0.8266		
ViT-L/32	COVID-19	0.9685	0.9475	0.9579	0.9666	0.9749
	Normal	0.9631	0.9831	0.9730		
	Pneumonia	0.9683	0.9702	0.9693		
	Macro average	0.9666	0.9666	0.9666		
	Weighted average	0.9666	0.9669	0.9667		
ViT-B/32	COVID-19	0.9700	0.9384	0.9540	0.9609	0.9707
	Normal	0.9495	0.9906	0.9696		
	Pneumonia	0.9639	0.9544	0.9591		

	Macro average	0.9609	0.9609	0.9609		
	Weighted average	0.9611	0.9611	0.9609		
STCovidNet (Proposed)	COVID-19	0.9766	0.9837	0.9801	**0.9811**	**0.9858**
	Normal	0.9812	0.9849	0.9831		
	Pneumonia	0.9859	0.9742	0.9800		
	Macro average	0.9811	0.9811	0.9811		
	Weighted average	0.9813	0.9809	0.9811		

Results And Discussions

Table 2 and Fig. 9 illustrate the experimental findings obtained by STCovidNet and other models. As can be seen, STCovidNet outperformed the ViT and CNN techniques on MUST-COVID-19, with 0.9811 AUC and 0.9858 accuracy score. We then analyze the model's accuracy, sensitivity (recall), and F1-scores and explain their importance in assessing the model's classification quality. It is worth mentioning that under the “dynamic zero-case” policy of COVID-19 adhered in China, sensitivity is considered the critical indicator of the COVID-19 auto-detection model, as any missed positive case can pose a severe risk to the communities.

According to Table 2, STCovidNet has the highest sensitivity value of 0.9837, revealing that a tiny number of pneumonia patients caused by COVID-19 pneumonia are wrongly categorized, which is a highly desired feature in a COVID-19 early warning model. Furthermore, our proposed approach received the most outstanding F1 scores in all categories, demonstrating that it is the best-balanced model among the baseline models regarding accuracy and sensitivity.

We can further understand the prowess of the proposed model by examining the confusion matrix (Fig. 9). Notably, among the 556-novel coronavirus-positive patients predicted by STCovidNet, 543 are confirmed to be novel coronavirus-infected patients. The actual labels of the other 8 and 5 are normal and non-COVID-19 pneumonia cases, respectively. Additionally, 523 of the 533 normal patients predicted by STCovidNet are consistent with their actual labels, and 491 of the 498 patients with other types of pneumonia predicted by STCovidNet are consistent with their actual labels. Overall, STCovidNet performed the best in identifying COVID-19, healthy, and other types of pneumonia patients.

The t-SNE Visualization

The results of t-SNE are depicted in Fig. 10. Figure 10 (a) appears to plot best in a compact space. It displays a clear difference between COVID-19 pneumonia instances and healthy and none-COVID-19 pneumonia patients, demonstrating that STCovidNet can decrease intra-class variations and generate well-separated feature embedding better than other approaches.

The Grad-CAM Visualization

We use the Grad-CAM visualization to generate a localization map that pinpoints the image's essential parts for prediction. The highlights of the activation region of the CT images are shown in Fig. 11, and it has significantly different activation regions for COVID-19 patients in Fig. 11 (a) and healthy people in Fig. 11 (b). Figure 11 (a) identifies the apparent feature region of infection in the patient's lung, which is the main activation area for the used model to classify the image as a COVID-19 instance. In Fig. 11(b), the healthy regions of the normal CT images are uniformly localized. Figure 11(a) and 11(c) show that it effectively distinguishes between COVID-19 and other types of pneumonia patients by locating different focal areas, which is close to the ability of advanced radiologists. All results visualized by the Grad-CAM method show that the used model learns valid representations before making classification decisions and is interpretable.

Conclusion

This research offers a COVID-19 detection model (STCovidNet) using the Swin Transformer blocks, trained and evaluated on MUST-COVID-19. The suggested approach yields the best results and adheres to medical judgment guidelines. Our findings suggest that STCovidNet can be considered a promising architecture to detect COVID-19 cases.

In future experiments, we intend to create further the proposed method for different types of pneumonia, to investigate whether the model can distinguish between multiple different cases of pneumonia. Furthermore, we will study the use of STCovidNet to COVID-19 detection on chest X-ray (CXR) to assess its efficacy.

Declarations

Acknowledgements

The authors express their sincere thanks to Prof. Wang Wenmin, International Institute for Next Generation Internet (IINGI), Macau University of Science and Technology, Taipa, Macau

Authors' contributions

Boyuan Wang completed the entire experiment and wrote the main manuscript text. Du Zhang directed the whole experiment and revised the manuscript. Zonggui Tian contributed to the experimental method design, assisted in writing Section II.

Funding

This work was supported in part by the Science and Technology Development Fund, Macao SAR, under Macao funding scheme for key R&D projects (0025/2019/AKP).

Availability of data and materials

The datasets used during the current study are available at kaggle, https://www.kaggle.com/mustai/covid19-ct-7930

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

[1] C. Huang et al., “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,” Lancet, vol. 395, no. 10223, pp. 497–506, 2020.

[2] “COVID Live - Coronavirus Statistics - Worldometer.” https://www.worldometers.info/coronavirus/ (accessed Feb. 24, 2022).

[3] Y. Fang et al., “Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR,” Radiology, p. 200432, Feb. 2020.

[4] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, “Chest CT for Typical 2019-nCoV Pneumonia: Relationship to Negative RT-PCR Testing,” Radiology, p. 200343, Feb. 2020.

[5] C. S. Guan et al., “Imaging Features of Coronavirus disease 2019 (COVID-19): Evaluation on Thin-Section CT,” Acad Radiol, vol. 27, no. 5, pp. 609–613, May 2020.

[6] A. Das, “Adaptive UNet-based Lung Segmentation and Ensemble Learning with CNN-based Deep Features for Automated COVID-19 Diagnosis,” Multimed Tools Appl, pp. 1–35, Dec. 2021.

[7] S. Park et al., “Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification,” Med Image Anal, vol. 75, p. 102299, Jan. 2022.

[8] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv:2010.11929 [cs], Jun. 2021.

[9] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, “xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography,” IEEE J Transl Eng Health Med, vol. 10, p. 1100110, Jan.

2022.

[10] Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” arXiv:2103.14030 [cs], Aug. 2021, Accessed: Feb. 04, 2022.

[11] V. Perumal, V. Narayanan, and S. J. S. Rajasekar, “Detection of COVID-19 using CXR and CT images using Transfer Learning and Haralick features,” Appl Intell, vol. 51, no. 1, pp. 341–358, Jan. 2021.

[12] M.-R. Lascu, “Deep Learning in Classification of Covid-19 Coronavirus, Pneumonia and Healthy Lungs on CXR and CT Images.,” J Med Biol Eng, pp. 1–9, Jun. 2021.

[13] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv 1409.1556, Sep. 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Jun. 2016, pp. 770–778.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” arXiv:1608.06993 [cs], Jan. 2018.

[16] C. Szegedy et al., “Going deeper with convolutions,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[17] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” Jul. 2017, pp. 1800–1807.

[18] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv:1704.04861 [cs], Apr. 2017.

[19] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proceedings of the 36th International Conference on Machine Learning, May 2019, pp. 6105–6114.

[20] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, “Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays,” Comput Methods Programs Biomed, vol. 196, p. 105608, Nov. 2020.

[21] A. Narin, C. Kaya, and Z. Pamuk, “Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks,” Pattern Anal Applic, vol. 24, no. 3, pp. 1207–1220, Aug. 2021.

[22] G. Wang et al., “A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images,” Nature Biomedical Engineering, vol. 5, no. 6, pp. 509–521, 2021.

[23] A. Shamila Ebenezer, S. Deepa Kanmani, M. Sivakumar, and S. Jeba Priya, “Effect of image transformation on EfficientNet model for COVID-19 CT image classification,” Mater Today Proc, Dec. 2021.

[24] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic Routing Between Capsules,” arXiv:1710.09829 [cs], Nov. 2017.

[25] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, “COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images,” Pattern Recognit Lett, vol. 138, pp. 638–643, Oct. 2020.

[26] S. Toraman, T. B. Alakus, and I. Turkoglu, “Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks,” Chaos Solitons Fractals, vol. 140, p. 110122, Nov. 2020.

[27] Y. Dong, J.-B. Cordonnier, and A. Loukas, “Attention is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth,” arXiv:2103.03404 [cs], Mar. 2021.

[28] D. Shome et al., “COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare,” Int J Environ Res Public Health, vol. 18, no. 21, p. 11086, Oct. 2021.

[29] K. Zhang et al., “Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography,” Cell, vol. 181, no. 6, pp. 1423-1433.e11, Jun. 2020.

[30] W. Ning et al., “Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning,” Nat Biomed Eng, vol. 4, no. 12, pp. 1197–1207, 2020.

[31] M. Rahimzadeh, A. Attar, and S. M. Sakhaei, “A fully automated deep learning-based network for detecting COVID-19 from a new and large lung CT scan dataset,” Biomedical Signal Processing and Control, vol. 68, p. 102588, Jul. 2021.

[32] K. Clark et al., “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository,” J Digit Imaging, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.

[33] M. Jun et al., “COVID-19 CT Lung and Infection Segmentation Dataset.” Zenodo, Apr. 20, 2020.

[34] A. I. G Samuel et al., “Data From LIDC-IDRI.” The Cancer Imaging Archive, 2015.

[35] “Radiopaedia.org, the wiki-based collaborative Radiology resource,” Radiopaedia. https://radiopaedia.org/ (accessed Feb. 06, 2022).

[36] S. P. Morozov et al., “MosMedData: Chest CT Scans With COVID-19 Related Findings Datase.” 2020. Accessed: Feb. 06, 2022.

[37] X. Li, Y. Zhou, P. Du, G. Lang, M. Xu, and W. Wu, “A deep learning system that generates quantitative CT reports for diagnosing pulmonary Tuberculosis,” Applied Intelligence, vol. 51, pp. 1–12, Jun. 2021.

[38] F. Zhuang et al., “A Comprehensive Survey on Transfer Learning,” Proceedings of the IEEE, vol. PP, pp. 1–34, Jul. 2020.

[39] Y. Brima, M. Atemkeng, S. Tankio Djiokap, J. Ebiele, and F. Tchakounté, “Transfer Learning for the Detection and Diagnosis of Types of Pneumonia including Pneumonia Induced by COVID-19 from Chest X-ray Images.,” Diagnostics (Basel), vol. 11, no. 8, Aug. 2021.

[40] Y. Gu, Z. Piao, and S. J. Yoo, “STHarDNet: Swin Transformer with HarDNet for MRI Segmentation,” Applied Sciences, vol. 12, no. 1, Art. no. 1, Jan. 2022.

[41] O. Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision, vol. 115, Sep. 2014.

[42] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, “Rethinking Spatial Dimensions of Vision Transformers,” arXiv:2103.16302 [cs], Aug. 2021, Accessed: Feb. 05, 2022.

[43] H. Fan et al., “Multiscale Vision Transformers,” arXiv:2104.11227 [cs], Apr. 2021, Accessed: Feb. 05, 2022.

[44] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li, and J. fu, “Pre-trained Language Models in Biomedical Domain: A Systematic Survey,” arXiv:2110.05006 [cs], Oct. 2021, Accessed: Feb. 04, 2022.

[45] H. Farhat, G. E. Sakr, and R. Kilany, “Deep learning applications in pulmonary medical imaging: recent updates and insights on COVID-19,” Mach Vis Appl, vol. 31, no. 6, p. 53, 2020.

[46] M. T. McCann, K. H. Jin, and M. Unser, “A Review of Convolutional Neural Networks for Inverse Problems in Imaging,” IEEE Signal Process. Mag., vol. 34, no. 6, pp. 85–95, Nov. 2017.

[47] E. Irmak, “COVID-19 disease severity assessment using CNN model,” IET Image Process, vol. 15, no. 8, pp. 1814–1824, Jun. 2021, doi: 10.1049/ipr2.12153.

[48] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Neural Information Processing Systems, vol. 25, Jan. 2012.

[49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[50] Z. Tao, H. Bingqiang, L. Huiling, Y. Zaoli, and S. Hongbin, “NSCR-Based DenseNet for Lung Tumor Recognition Using Chest CT Image,” Biomed Res Int, vol. 2020, p. 6636321, Dec. 2020.

[51] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-Excitation Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[52] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for Activation Functions,” arXiv:1710.05941 [cs], Oct. 2017, Accessed: Jan. 05, 2022.

[53] L. van der Maaten and G. Hinton, “Viualizing data using t-SNE,” Journal of Machine Learning Research, vol. 9, pp. 2579–2605, Nov. 2008.

[54] M. Shorfuzzaman, M. Masud, H. Alhumyani, D. Anand, and A. Singh, “Artificial Neural Network-Based Deep Learning Model for COVID-19 Patient Detection Using X-Ray Chest Images,” J Healthc Eng, vol. 2021, p. 5513679, Jun. 2021.

[55] B. Melit Devassy and S. George, “Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE,” Forensic Science International, vol. 311, p. 110194, Jun. 2020, doi: 10.1016/j.forsciint.2020.110194.

[56] S. Kullback, “Information Theory and Statistics,” 1959.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 618–626.

[58] G. Wang et al., “A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images,” Nat Biomed Eng, vol. 5, no. 6, pp. 509–521, Jun. 2021.

Figures

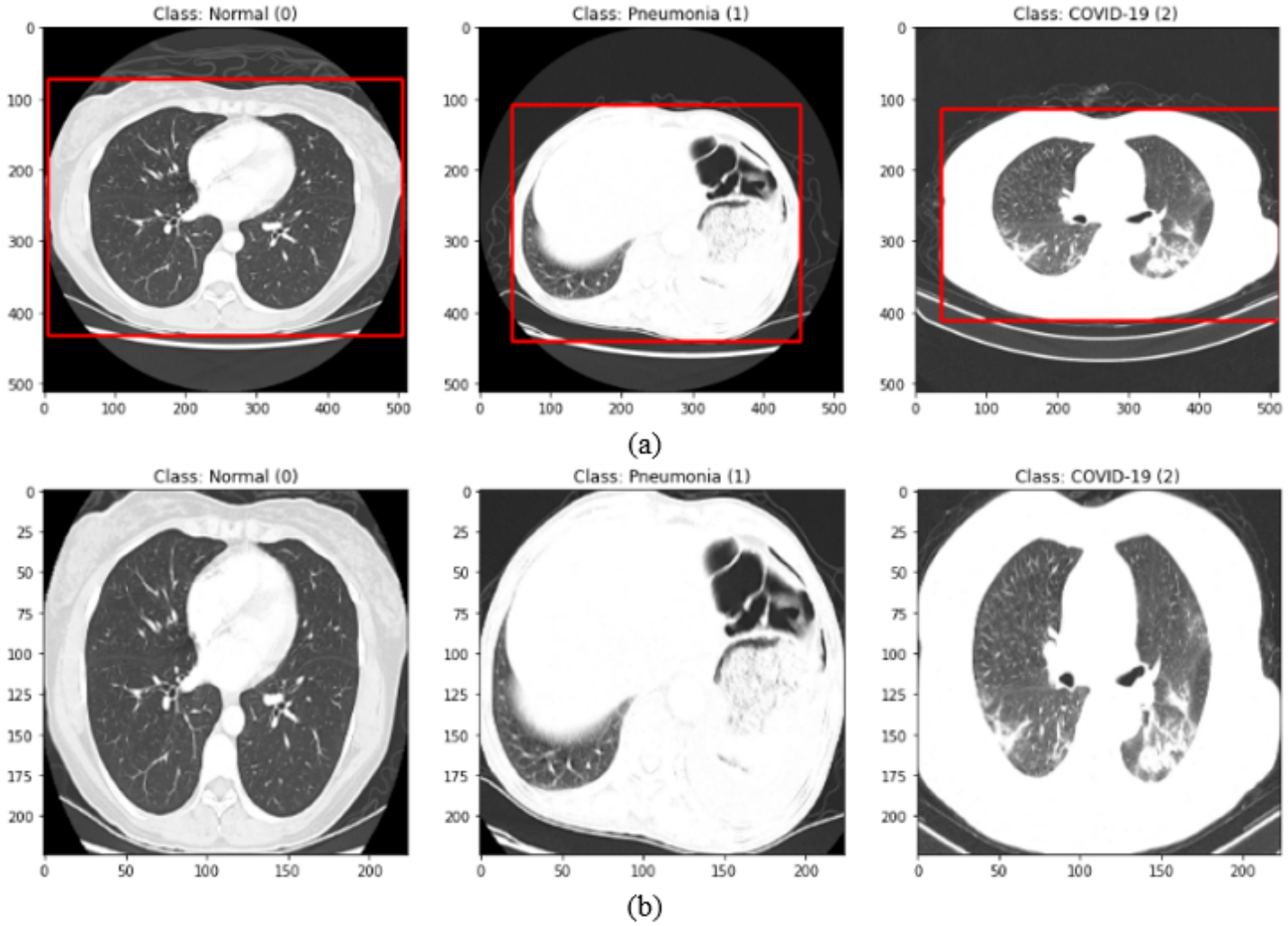

Figure 1

Examples of chest CT scans: (a) Original image; (b) Pre-processed image.

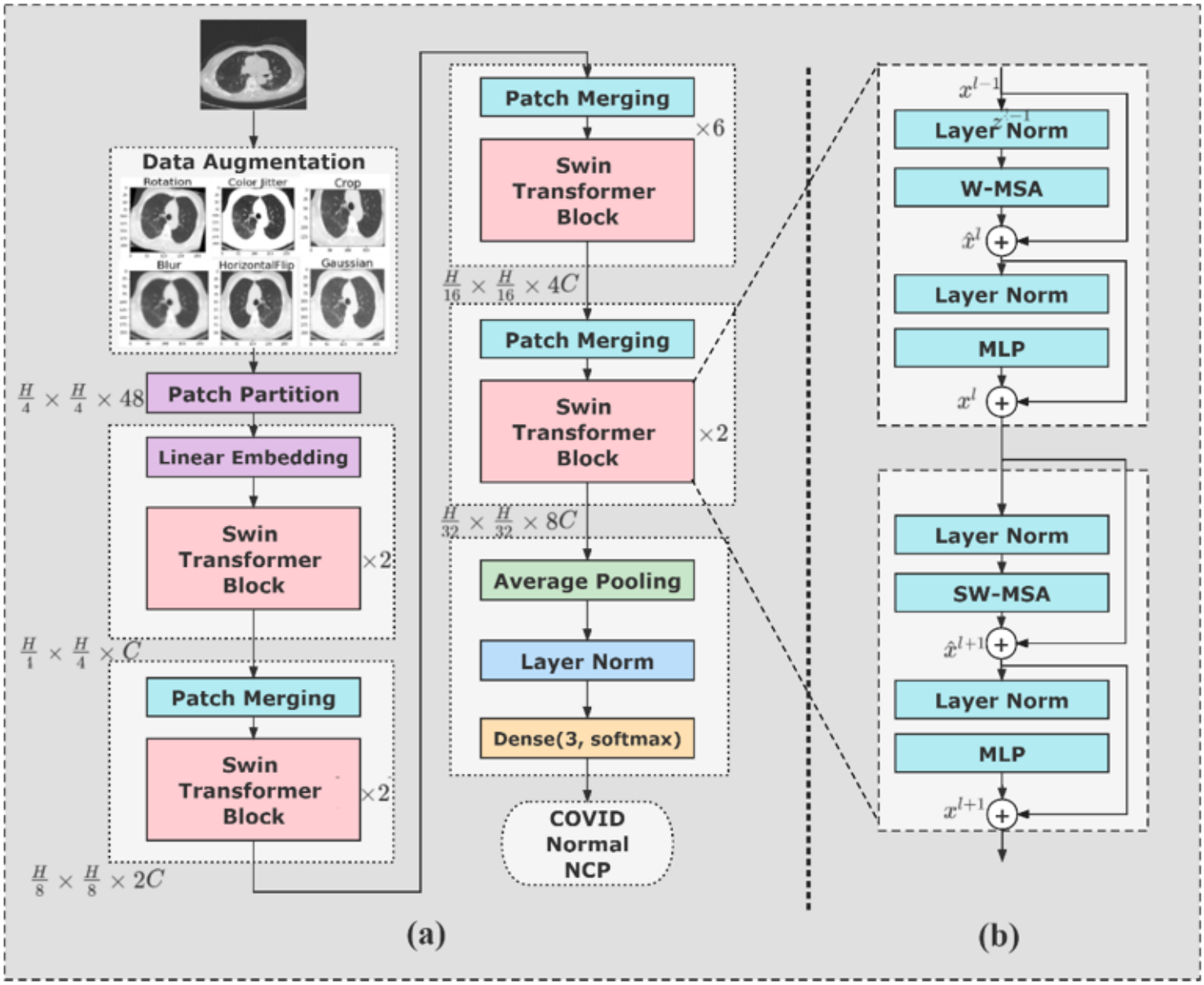

Figure 2

The proposed STCovidNet for COVID-19 case detection. (a) overall architecture; (b) Swin Transformer block

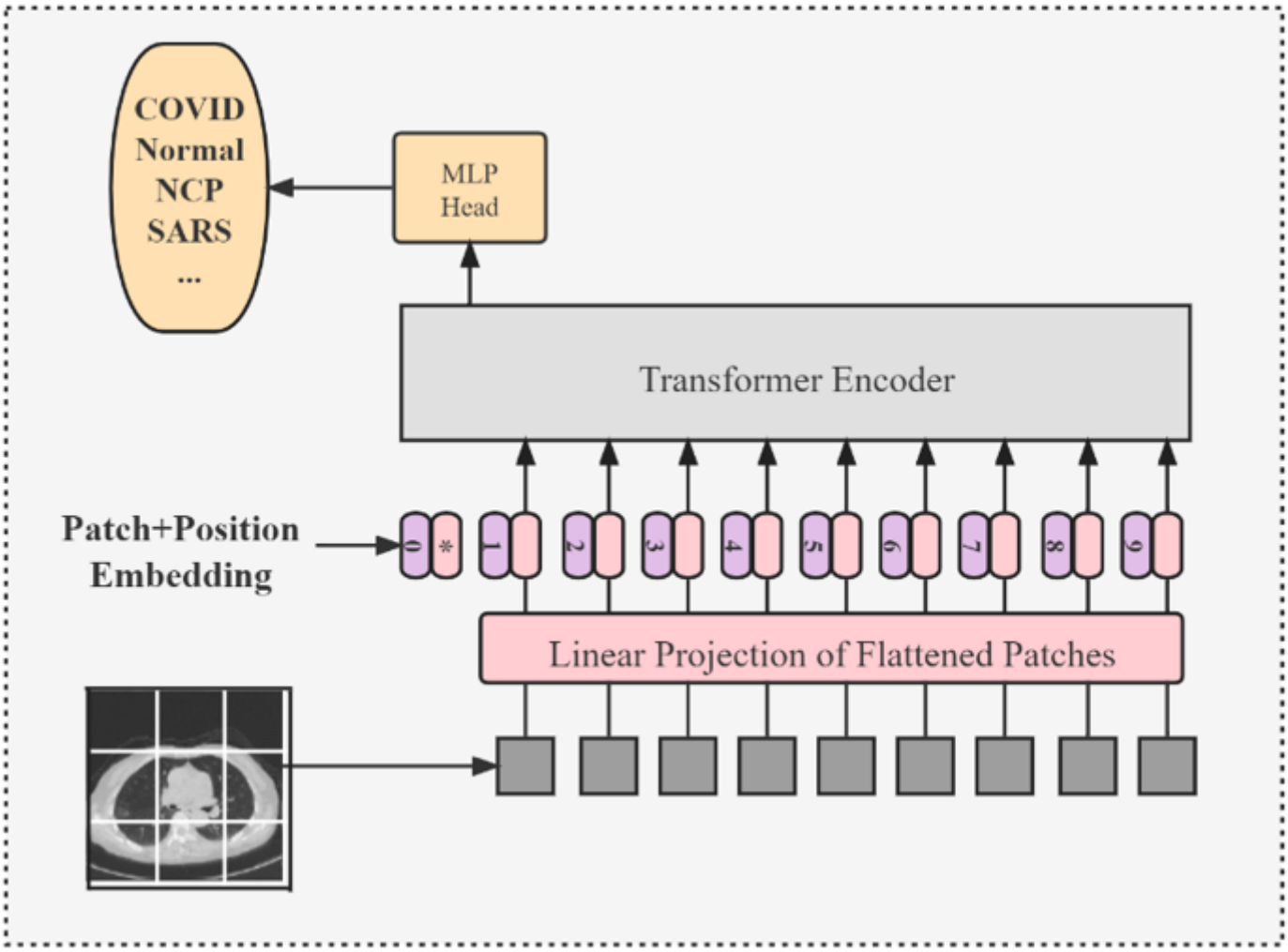

Figure 3

The architecture of ViT

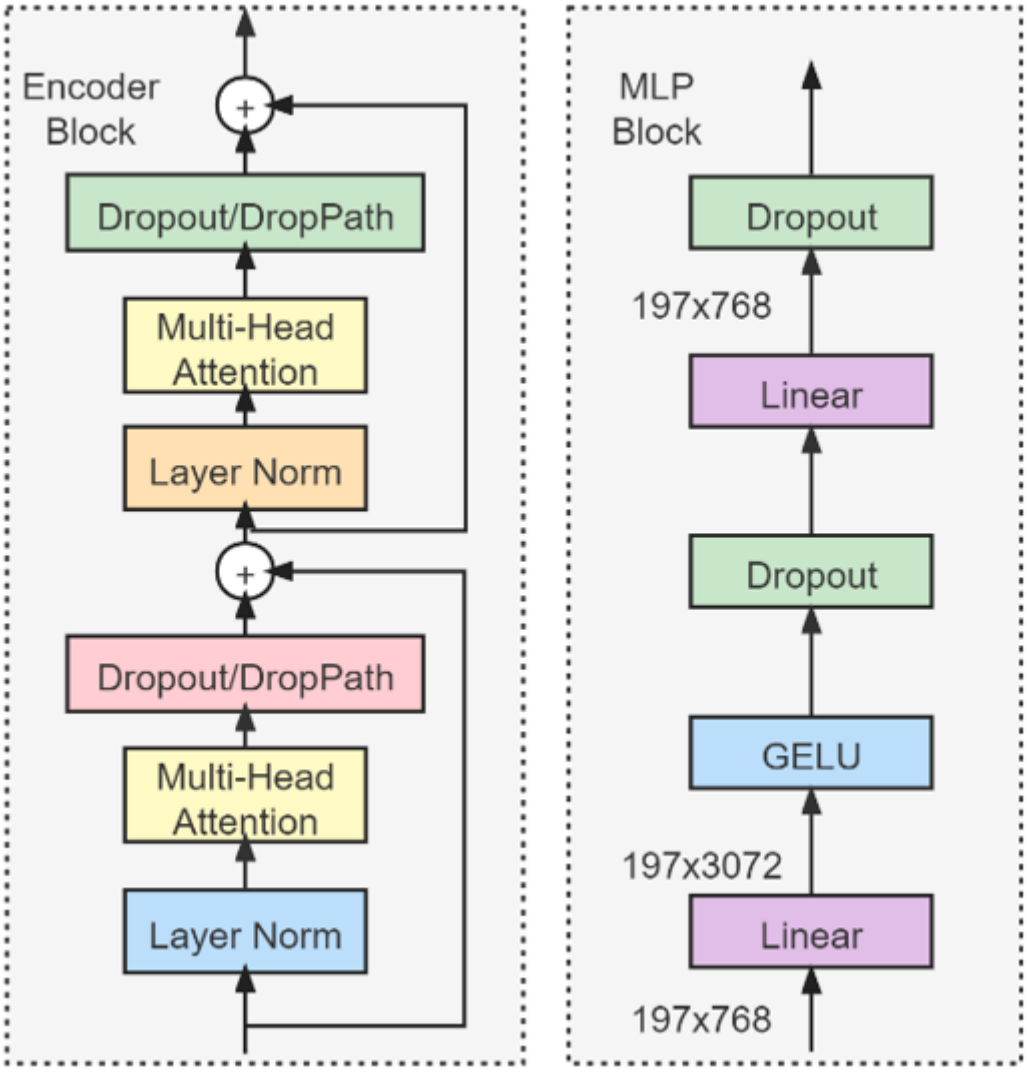

Figure 4

Detailed architecture of Encoder Block and MLP.

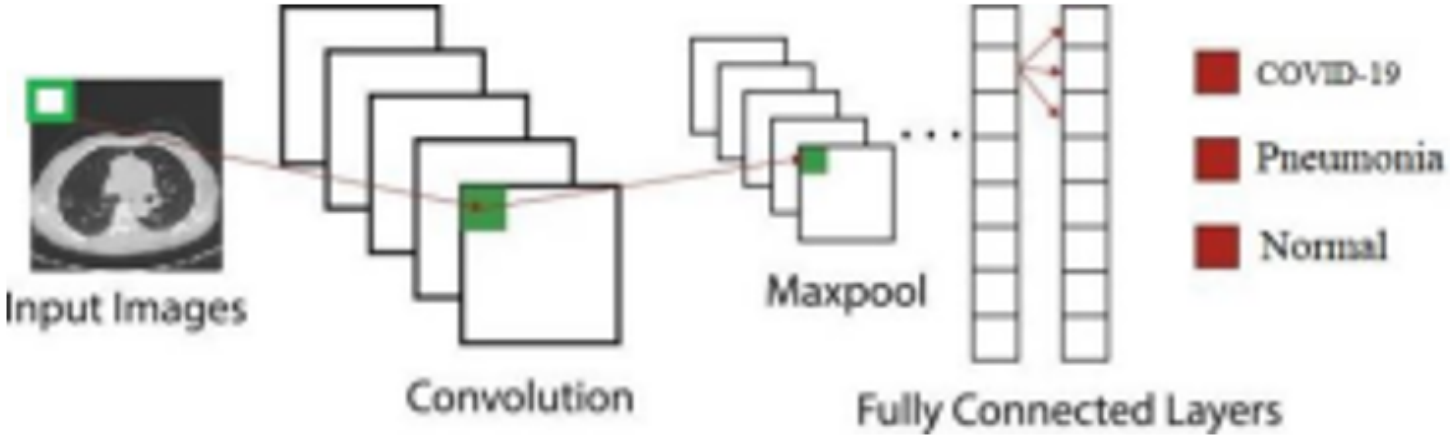

Figure 5

The general framework of CNN.

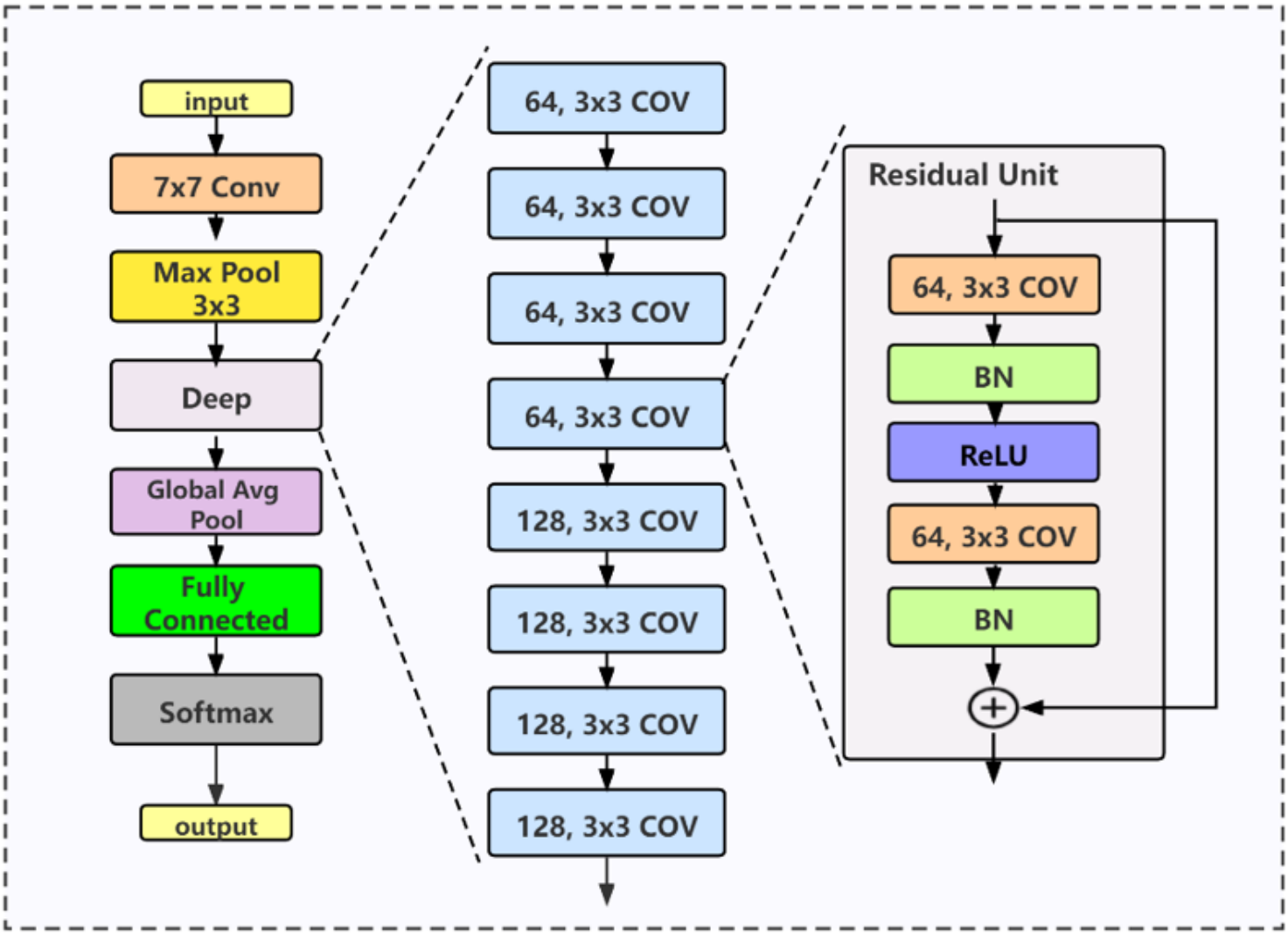

Figure 6

The architecture of ResNet-50 for COVID-19 classification.

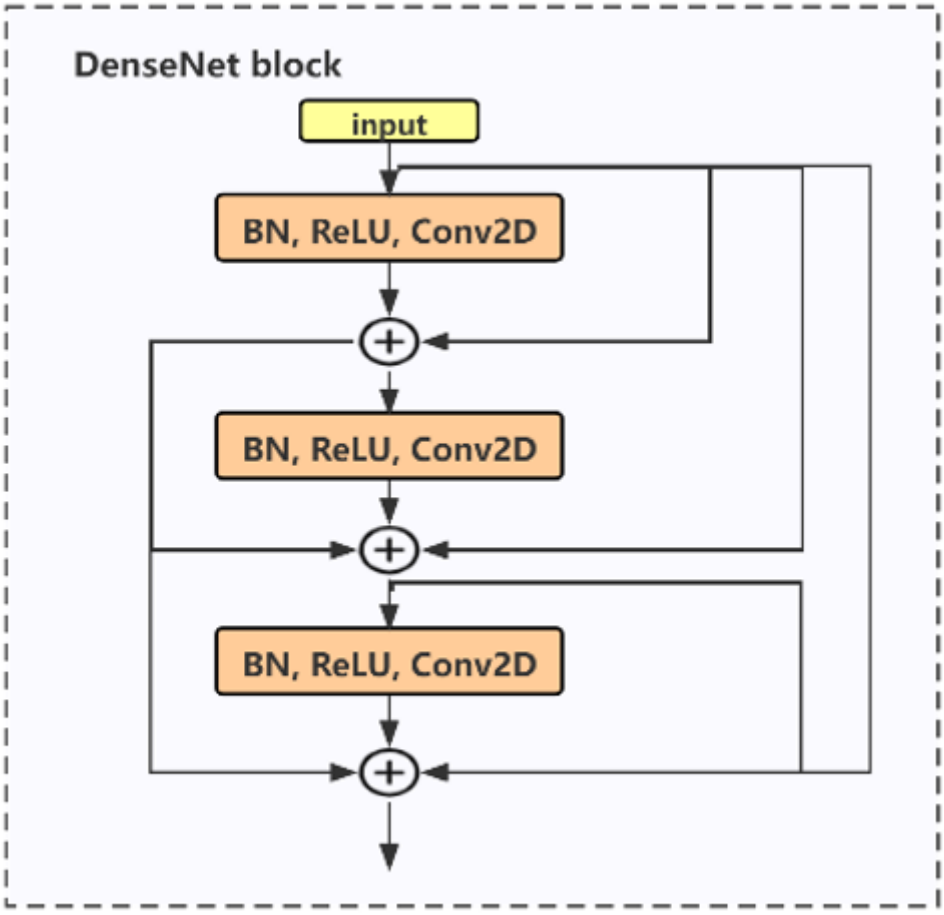

Figure 7

The detailed structure of a three-layer Dense block.

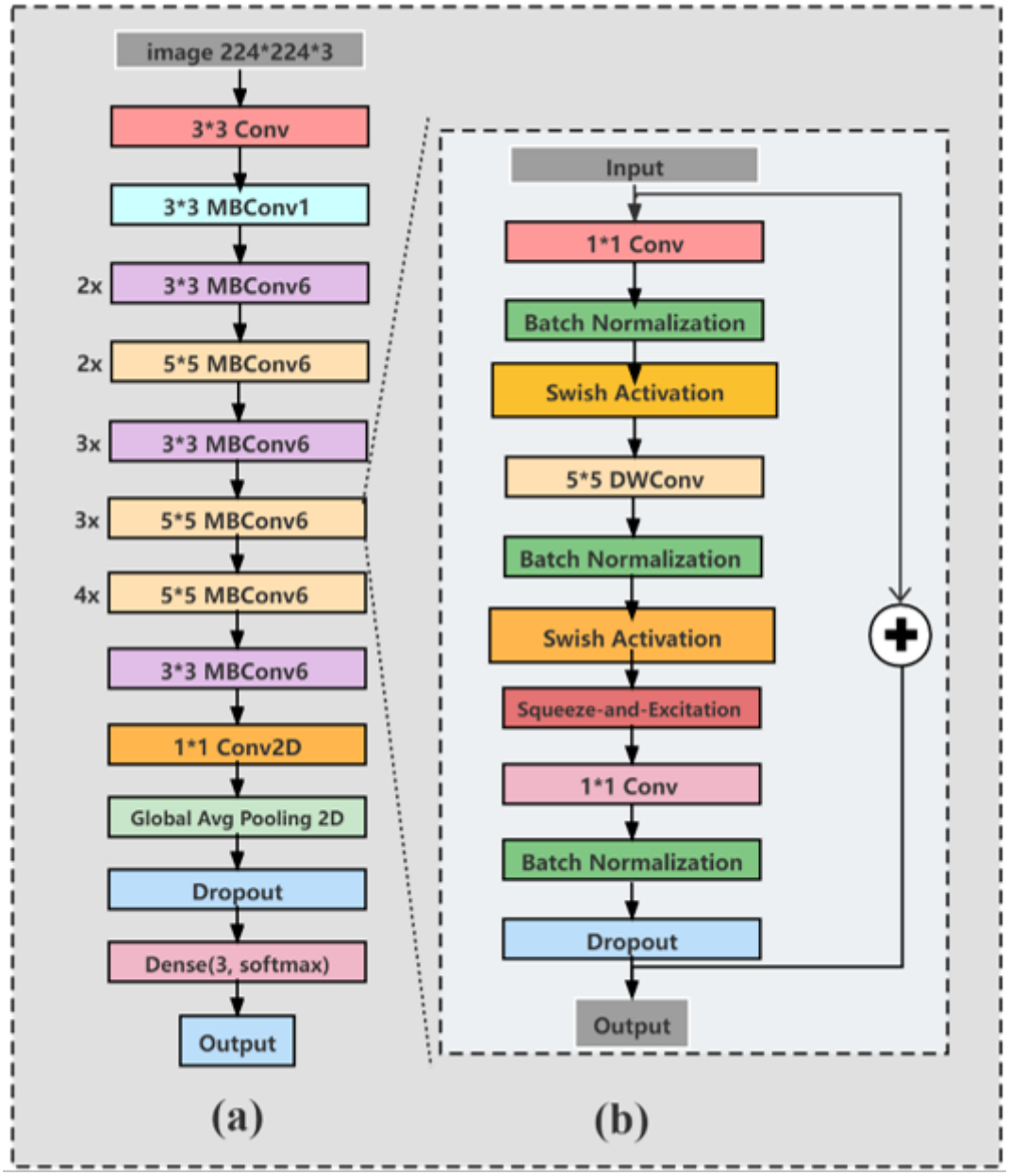

Figure 8

The architecture of EfficientNet. (a) B0 architecture; (b) Mobile Inverted Bottleneck Conv (MBconv) block

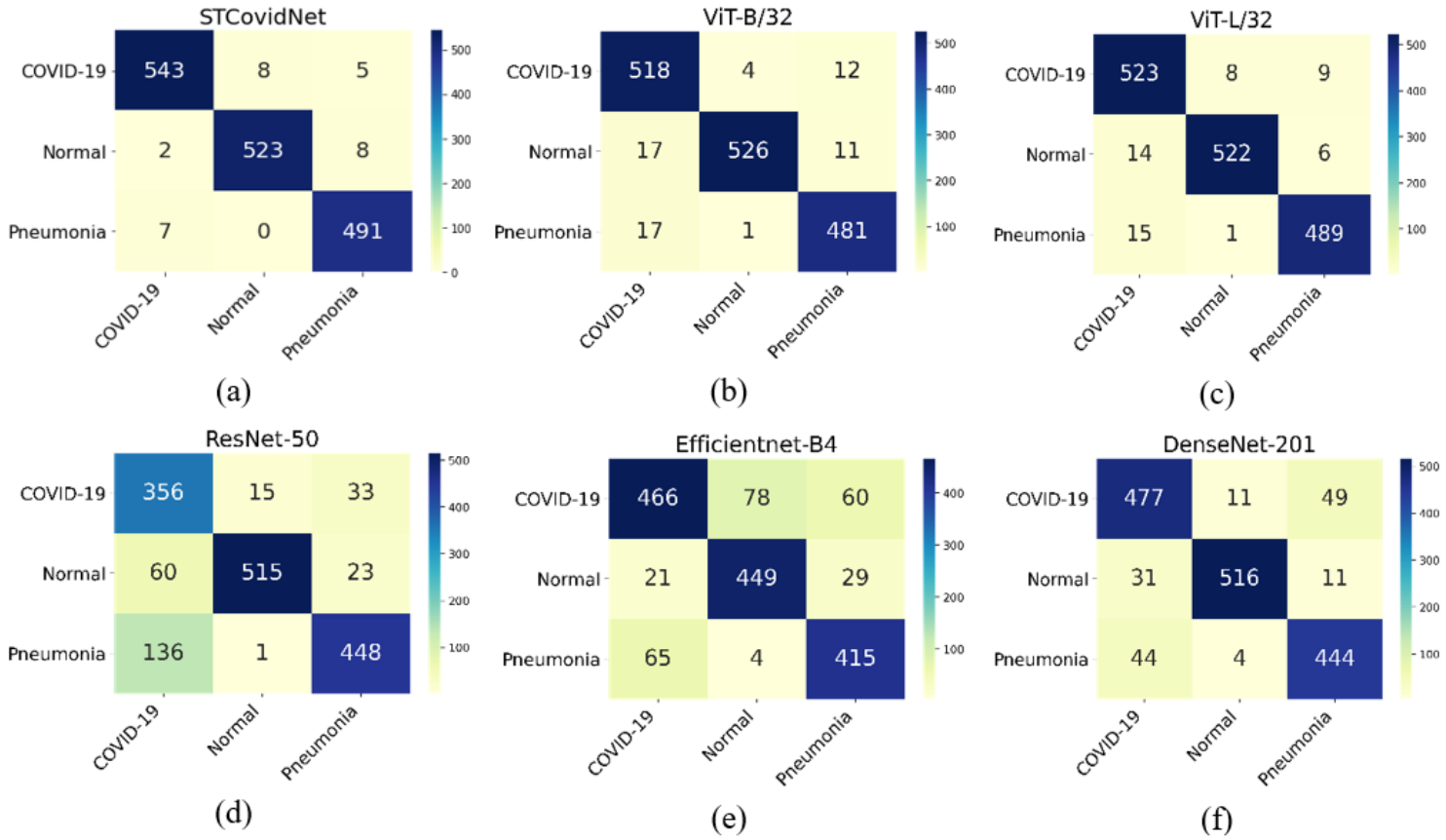

Figure 9

Confusion matrix of all architectures. (a) STCovidNet ; (b) ViT-B/32; (c) ViT-L/32; (d) ResNet-50; (d) Efficientnet-B4; (e) DenseNet-201

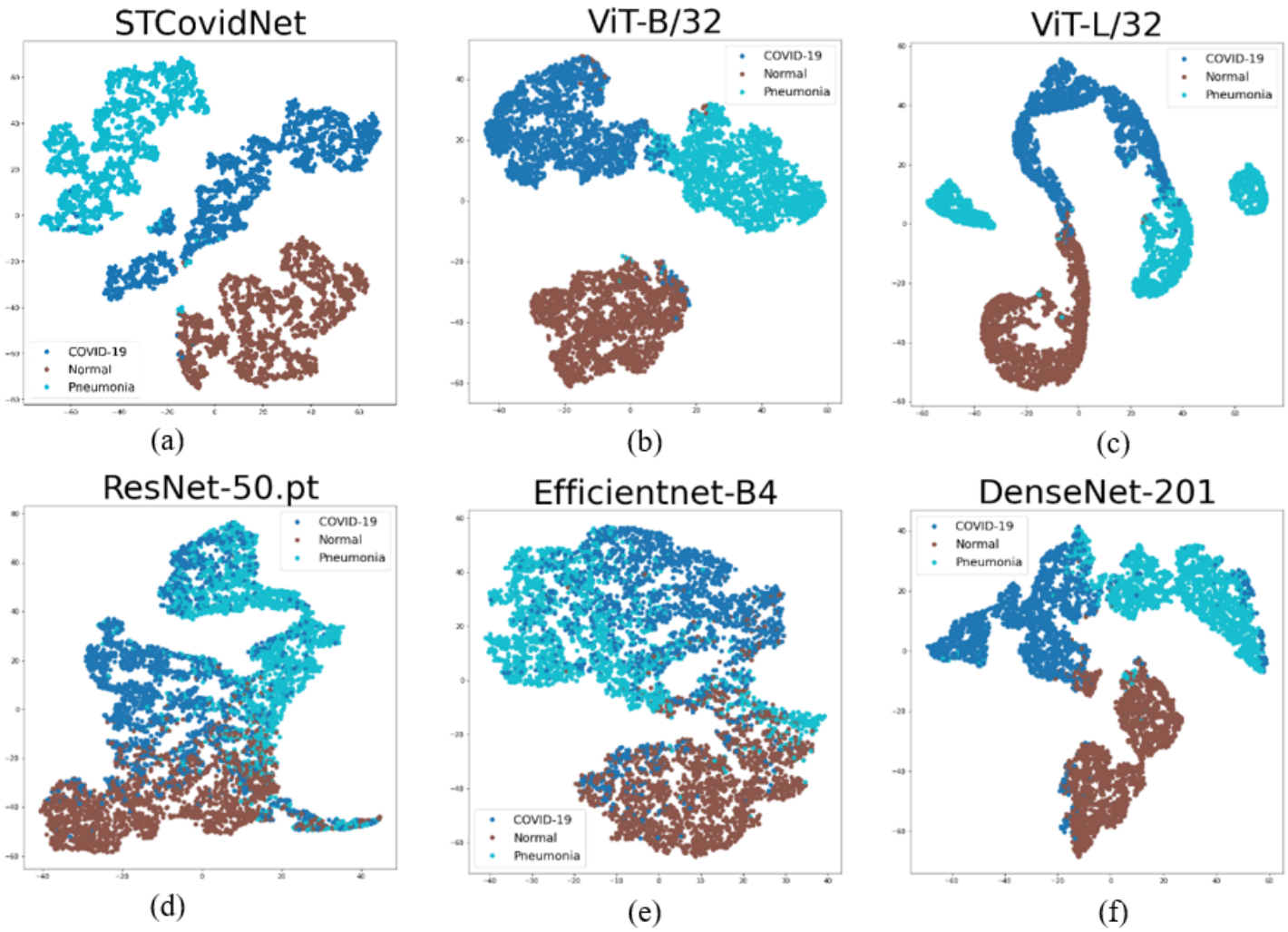

Figure 10

The t-SNE visualization. (a) STCovidNet ; (b) ViT-B/32; (c) ViT-L/32; (d) ResNet-50; (e) Efficientnet-B4; (f) DenseNet-201

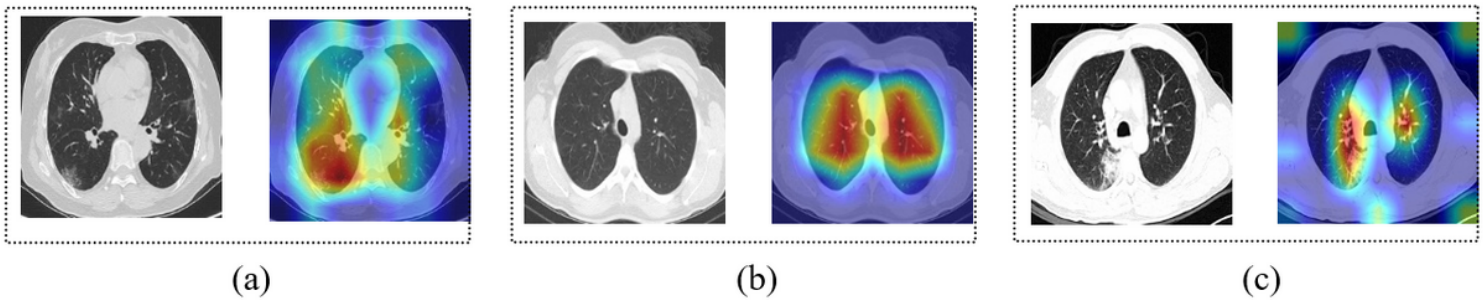

Figure 11

The Grad-CAM visualization. (a) COVID-19 pneumonia; (b) healthy; (c) non-COVID-19 pneumonia.