Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

BIOCYBERNETICS AND BIOMEDICAL ENGINEERING 42 (2022) 842–855

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

TL-med: A Two-stage transfer learning recognition model for medical images of COVID-19

Jiana Meng, Zhiyong Tan, Yuhai Yu, Pengjie Wang, Shuang Liu*

School of Computer Science and Engineering, Dalian Minzu University, Dalian, Liaoning, China

ARTICLE INFO

Article history:
Received 21 November 2021
Received in revised form
16 April 2022
Accepted 20 April 2022
Available online 29 April 2022

Keywords:
COVID-19
ViT
Pretrained Model
Transfer Learning

ABSTRACT

The recognition of medical images with deep learning techniques can assist physicians in clinical diagnosis, but the effectiveness of recognition models relies on massive amounts of labeled data. With the rampant development of the novel coronavirus (COVID-19) worldwide, rapid COVID-19 diagnosis has become an effective measure to combat the outbreak. However, labeled COVID-19 data are scarce. Therefore, we propose a two-stage transfer learning recognition model for medical images of COVID-19 (TL-Med) based on the concept of "generic domain-target-related domain-target domain". First, we use the Vision Transformer (ViT) pretraining model to obtain generic features from massive heterogeneous data and then learn medical features from large-scale homogeneous data. Two-stage transfer learning uses the learned primary features and the underlying information for COVID-19 image recognition to solve the problem by which data insufficiency leads to the inability of the model to learn underlying target dataset information. The experimental results obtained on a COVID-19 dataset using the TL-Med model produce a recognition accuracy of 93.24%, which shows that the proposed method is more effective in detecting COVID-19 images than other approaches and may greatly alleviate the problem of data scarcity in this field.

1. Introduction

Since the outbreak of the novel coronavirus (COVID-19) in late 2019, the virus has been ravaging the world to date, which has had a great impact on global politics, the economy, culture, etc., and has caused immeasurable economic and property losses. Due to its extremely high infection rate and mortality rate, the World Health Organization declared that the COVID-19 outbreak was a global health emergency in a very short period [1]. As of 6 September 2021, more than 220 million cumulative diagnoses and more than 85 million cumulative deaths have been confirmed worldwide. These numbers may be even higher due to asymptomatic cases and flawed tracking policies. Some researchers have modeled COVID-19 through the fractal approach of the epidemic curve to help the medical community understand the dynamics and evolution of the COVID-19 outbreak, and thus control the spread of the outbreak [2]. Other researchers have used mice to study

* Corresponding author at: School of Computer Science and Engineering, Dalian Minzu University, Dalian, Liaoning 116600, China.
E-mail address: yuyh@dlnu.edu.cn (Y. Yu).

https://doi.org/10.1016/j.bbe.2022.04.005

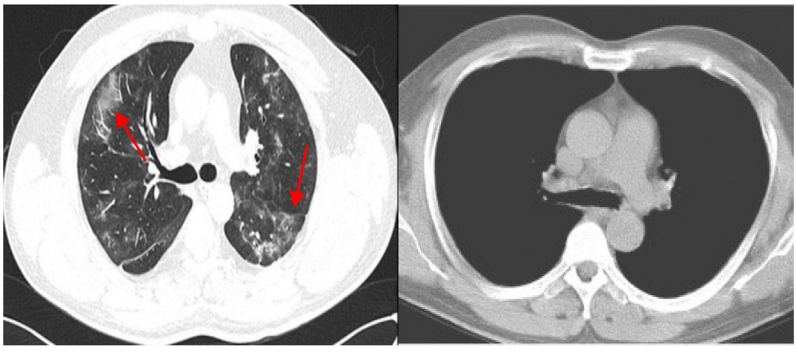
the pathogenesis of COVID-19 in order to evaluate the effectiveness of new treatments and vaccines and to find more effective ways to combat COVID-19 [3]. Despite global efforts to prevent rapid outbreaks of the disease, hundreds of thousands of COVID-19 cases continue to be confirmed worldwide every day, making rapid COVID-19 detection an effective measure with which governments can respond to COVID-19; this can help health departments and government authorities develop effective resource allocation strategies and break the chain of transmission.

The most typical symptoms of COVID-19 include fever, dry cough, myalgia, dyspnea, and headache, but in some cases, no symptoms can be seen (these are called asymptomatic cases), making this disease a greater public health threat than other diseases. Currently, reverse transcription-polymerase chain reaction (RT-PCR) is viewed as the gold standard for the diagnosis of COVID-19. However, the rapid and effective screening of suspected cases is limited by the lack of resources and stringent testing environment requirements. In addition, RT-PCR tests are time-consuming and have high false negative (FN) rates [4]. The spread of COVID-19 has led to the production of many variants of this strain, such as delta variant, which is hundreds of times more virulent and infectious than common strains; these mutations undoubtedly add to the pressure of COVID-19 detection. However, computed tomography (CT) scan images are considered a better method for detecting COVID-19 [5], with a sensitivity of 98 % (compared to 71 % for RT-PCR) [4]. In COVID-19 cases, CT scan images show some specific manifestations, including multilobular ground glass shadows (GGOs) distributed on both sides, in the periphery or in the rear [6], predominantly in the lower lobes and less frequently in the middle lobes. Diffuse distribution, vascular thickening, and fine reticular clouding are other common features reported for patients with neocoronavirus pneumonia. For example, Fig. 1 shows two CT scans: one for a patient with COVID-19 and one for a non-COVID-19 patient [7]; the CT scan of the lungs of a patient infected with COVID-19 is located on the left side of the figure, where some distinct GGOs are shown as red arrows, and a normal CT scan of the lungs is shown on the right side. The use of computers to classify medical images and thus to assist physicians in diagnosis is now a common and effective method [8,9]. The use of artificial intelligence technology to classify CT images has become a widespread concern in medical image analysis [10,11].

During the pandemic, hospitals generate thousands of CT scan images every day. For such a large number of CT scan images, it is a great challenge to rely on the naked eye of a professional physician for detection and identification, and the human eye tends to become fatigued and easily overlook some details, leading to misdiagnosis; the cost of misdiagnosis is unbearable. In contrast, machines do not become fatigued, and some details that are easily overlooked by humans can be detected. Therefore, the deep learning approach can effectively help us to quickly detect and identify COVID-19. Based on these facts, our approach introduces a Vision Transformer (ViT) into the COVID-19 detection and classification task.

The proposed model uses CT scan images to identify normal CT scan images and patient CT scan images. Since the generic domain dataset used for the ViT-pretrained model and the dataset used in this experiment are different in terms of their feature spaces and dimensions and the tuberculosis (TB) dataset is highly similar to the COVID-19 dataset with respect to their feature spaces and dimensions, in the first stage, we use heterogeneous transfer learning to learn the generic features of the images and use the ViT model (which is trained on a large proportion of the generic domain data) to detect the five types of TB. Then, in the second stage, the domain features of the images are learned by using homogeneous transfer learning. Based on the first stage, the model obtained from the TB dataset is used as the second stage pretrained model in the second stage, and this model then fine-tuned to detect and identify either COVID-19 or non-COVID-19 patients. The performance of the proposed technique is evaluated by comparing the results derived from the model with those of Residual Network 34 (ResNet34) [12], ResNet101 [12], and DenseNet169 [13], which are based on convolutional neural network (CNN) technology. The experimental results show that the developed model outperforms the existing models based on CNN techniques.

The remainder of the paper is organized as follows. Section 2 introduces the related works on COVID-19 detection. Section 3 describes the proposed method in detail. Section 4 gives the related experimental results, and Section 5 draws conclusions.



(a) (b)

Fig. 1 – A CT scan of the lungs of a patient infected with COVID-19 and a CT scan of a normal lung [7].

2. Related work

In this section, we discuss the research areas relevant to our work - research related to the detection of COVID-19.

Since the COVID-19 outbreak, rapid diagnostic testing has become one of the most effective methods for interrupting the spread of COVID-19. Because deep learning-based detection methods are more convenient and faster than traditional approaches, researchers have developed many effective models for detecting COVID-19 based on deep learning.

Heidarian et al. [14] developed a two-stage fully automated CT framework (COVID-FACT), which is mainly composed of a capsule network. COVID-FACT can capture spatial information without extensive data augmentation and large datasets and is performed in two stages: the first stage detects infected slices, and the second stage classifies the patient CT scan image of the patient. The supervision and annotation of COVID-FACT depend less on the input data relative to the same type of model.

Chaudhary et al. [15] developed a two-stage classification framework based on CNNs. The authors first used a pretrained DenseNet [13] for COVID-19 or community-acquired pneumonia (CAP) detection in the first stage and then used the EfficientNet [16] network for COVID-19, CAP and normal controls for triclassification. Its classification effectiveness was ranked first in the IEEE ICASSP 2021 Signal Processing Grand Challenge (SPGC) evaluation. Data are expensive resources; in particular, datasets in the medical field are extremely scarce. For this reason, He et al. [17] collected and collated hundreds of COVID-19 CT scan images and made them publicly available. A self-supervised transfer learning method was also proposed, reducing the bias generated on the source images and their class labels by conducting an auxiliary task on the CT images, and its performance was superior to that of several state-of-the-art methods. Polsinelli et al. [18] designed a lightweight CNN based on SqueezeNet [19], whose processing speed surpassed that of other models and whose processing time when running without GPU acceleration also surpassed that of many models that require GPU acceleration; however, its performance was not greatly improved over that of other networks, and its accuracy was not high. Sani et al. [20] constructed a new network structure for COVID-19 recognition by using a high-precision Hopfield neural network (HNN) to find symptoms and using a mathematical model to improve the accuracy of masking. Scarpiniti et al. [21] proposed a novel unsupervised method that used deep denoising convolutional autoencoders to provide compact and meaningful hidden representations. The experimental results show that the method has high reliability and low computational cost for the recognition of COVID-19.

Similarly, Pathak et al. [22] designed a transfer learning network based on ResNet-50 [23] that extracts the latent features of CT COVID-19 images through ResNet-50 and uses transfer learning to train the classification model. Finally, based on its training results, optimized hyperparameters are obtained by using a CNN. Jaiswal et al. [24] used DenseNet201 [13] to classify patients with COVID-19. They used transfer learning techniques to extract the image features learned by pretraining DenseNet201 and then fed these obtained features into a CNN for classification. Loey et al. [25] generated more images via classical data augmentation with conditional generative adversarial networks (CGANs) [26], and then they trained these networks for classification via deep transfer learning methods. Muhammet et al. [27] proposed two architectures to classify COVID-19. In this study, AlexNet is used as the backbone network for transfer learning. Compared with architecture 1, which directly used AlexNet for transfer learning, architecture 2 consists of AlexNet and BiLSTM, considering the time and order of the data.

There are also many researchers working on the detection of COVID-19 through the study of X-ray modal data. Xiao et al. [28] adopted a local phase-based image enhancement method to obtain a multi-feature CXR image, which was fed into the network together with the original image for fusion, which further improved the classification performance of the model. For the tuning problem in deep learning, Mohammad et al. [29] optimized the CNN structure in multiple stages by iteratively using heuristic optimization methods to finally evolve the best performing network with the smallest number of convolutional layers. Govardhan et al. [30] utilized two CNN models (ResNet50 and ResNet101) for a two-stage detection task. In this study, the first-stage ResNet50 distinguished bacterial pneumonia, viral pneumonia, and X-rays of normal healthy people, and the detected viral pneumonia samples are used as the input data of the second-stage ResNet101 network to distinguish COVID-19 and other viral pneumonia patients. Bejoy et al. [31] used multiple pre-trained CNN models for feature extraction, before selecting significant features through correlation-based feature selection techniques and subset size forward selection, and finally classifying them through a Bayesnet classifier. Ruochi et al. [32] proposed the COVID19XrayNet model, which was designed using two-step transfer learning. In this study, the first step used pre-trained ResNet34 for fine-tuning on the common pneumonia dataset, transfering the model weights trained in the first step to the corresponding network modules in the second step, and trained on the COVID-19 dataset. Shervin et al. [33] fine-tuned four pre-trained network models (ResNet18, ResNet50, SqueezeNet, DenseNet-169) with a small number of COVID-19 datasets, and the model performed well. Shayan et al. [34] directly utilized standard CNN to classify lung images, and the model performed well. Gupta et al. [35] proposed the COVID-WideNet capsule network. Compared with other CNN models, its parameters are greatly reduced, and it can detect COVID-19 quickly and efficiently. Goel et al. [36] proposed the Multi-COVID-Net model, which was an ensemble network composed of InceptionV3 and ResNet50 pre-trained networks, and optimized hyperparameters through the Multi-Objective Grasshopper Optimization Algorithm (MOGOA). Since fine-tuning the architecture and hyperparameters of deep learning models is a complex and time-consuming process, Jalali et al. [37] proposed a novel deep neuroevolutionary algorithm, which was mainly achieved by modifying the competitive swarm optimizer algorithm and adjusting the volume Hyperparameters and Architecture of Productive Neural Networks.

Compared with the CNN structure, ViT can pay attention to more global information at the lower level. And more data for training can help ViT to further learn local information, thereby improving the performance of the model [38]. Therefore, many researchers use the ViT architecture to detect COVID-19. In order to accurateely classify and quantify the severity of COVID-19, Sangjoon et al. [39] proposed a multi-task Vision Transformer model.The method obtained its low-level corpus features by pre-training on a large and general CXR dataset, and then used the acquired corpus features are input as a general Transformer for classification and severity quantification of COVID-19. To address the problem that the Vision Transformer required large-scale data for training, in order to obtain a Vision Transformer with good performance, Li et al. [40] used ResNet as a teacher network through a knowledge distillation method to extract its knowledge learned in COVID-19 CT images into the student network Vision Transformer. Debaditya et al. [41] proposed the COVID-Transformer model, which fine-tuned the Vision Transformer on a large dataset collected and merged, and performed well compared to other standard CNN models. Sangjoon et al. [42] proposed the FESTA framework. By dividing the Vision Transformer into three parts: head, tail and Transformer for training, it reduced the consumption of client resources and bandwidth by joint training, and improved the performance of the model for single task processing. By fine-tuning the Vision Transformer, Koushik et al. [43] performed better than other standard CNN models. Sangjoon et al. [44] used DenseNet to extract abnormal lung features on a large low-level dataset, and then used it as a corpus for Vision Transformer training to obtain a better feature embedding and improve the performance of the model. ARNAB et al. [45] proposed the xViTCOS framework, which employed Vision Transformer as the backbone network through a multi-stage transfer learning approach, pre-training and fine-tuning on multiple scales of ImageNet datasets, followed by fine-tuning training on the COVID-19 CXR dataset and CT dataset, respectively, to allow the model to them for classification.Sangjoon et al. [46] trained a backbone network with a large number of carefully curated public records to generate generic CXR results, thereby maximizing the performance of the Transformer model using a low-level CXR corpus from the backbone network.

3. Proposed method

In this paper, we propose a two-stage transfer learning approach, which is based on the ViT model, to classify COVID-19 by using different transfer learning methods.

The overall model structure is shown in Fig. 3. In the figure, ① indicates that the ImageNet dataset is trained through the ViT model, ② indicates that the trained model in ① is used as a pretraining model to complete heterogeneous transfer learning on the TB dataset, ③ represents the transfer of the relevant medical feature knowledge learned in ② (as shown by the dotted line in the figure) and the completion of homogeneous transfer learning on the COVID-19 dataset. The main process of the model is as follows. First, a multi-modal image is input $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ and undergoes a series of image preprocessing steps to convert the image to a uniform resolution size (224, 224); the image is then passed through a convolution layer and flattened to $\mathbf{x}_p \in \mathbb{R}^{N\times\left(P^2\cdot C\right)}$, where p is the resolution size of each image patch and $N = HW/P^2$ is the number of image patches, which is then spread and projected using a trainable linear $E \in \mathbb{R}^{\left(P^2\cdot C\right)\times D}$ onto a D-dimensional vector and then concatenated with a trainable class token for classification, i.e., $\mathbf{x}_{class} \in \mathbb{R}^{1\times\left(P^2\cdot C\right)}$. Thus, its dimensionality becomes $\mathbf{x}_p \in \mathbb{R}^{(N+1)\times D}$, after which a positional embedding with the same dimensions is added: $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$. The obtained result is fed into the transformer encoder module for computation, and finally, the class token with dimensions $\mathbf{x}_{class} \in \mathbb{R}^{1\times\left(P^2\cdot C\right)}$ is extracted and sent to the multilayer perceptron (MLP) module for classification.

In this work, several main parts of the model include a self-attention (SA) mechanism, a multiheaded self-attention (MSA) mechanism, and an MLP.

3.1. *Attention mechanisms*

Since the transformer [47] was proposed in 2017, it has developed in the field of natural language processing (NLP) at an astonishing rate, and its attention model soon received great attention from NLP researchers. In the following years, transformer-based models have emerged, and some of the better models, such as the bidirectional encoder representations from transformers (BERT) model [48], have occupied the main position in the field of NLP in recent years. Inspired by the success of transformers in NLP, researchers have introduced them into the computer vision (CV) field and conducted some experiments. The experimental results show that transformers have great potential to surpass models with pure CNN architectures in some areas. Some researchers have used combinations of CNNs and transformers [49,50], and other researchers have directly used a pure transformer architecture instead of a CNN architecture [51,52]. The biggest reason for the success of transformers is their attention mechanism. For translation tasks, the transformer model proposed by Google replaced long short-term memory (LSTM) with an attention mechanism and has been a great success. The principle of an attention mechanism is that different features are contained in each layer of a network; these features can be in different channels and different locations and have different levels of importance, and the later layers should pay attention to the most important information and suppress the less important information. In other words, an attention mechanism should increase the presence of the areas that need attention and give them extra attention, while it should reduce the presence of less important areas and then reduce their influence on the overall situation.

The ViT model proposed in [51] has greatly stimulated interest in transformer research. Similar to the sequence processing approach in the field of NLP, ViT uses a pure transformer structure, where the input images are split into fixed-size patches, and then the embedded sequences of these image patches are used as inputs for the transformer. Through a series of experiments, it has been shown that ViT has great potential for image processing, especially for large-scale image processing.

Nonlocal neural networks [53] are viewed as early works on attention mechanisms in computer vision and early attempts by researchers to use transformers in the CV field. The use of self-attention mechanism involves building remote dependencies and then directly capturing these remote dependencies for determining the interactions between any two locations without being restricted to adjacent points; this is equivalent to constructing a convolution kernel as large as the size of the feature map and thus can maintain more information. The drawbacks of this network model are that its time complexity and space complexity are both large and that expensive GPU devices are required to train the model on massive data. Bahdanau et al. [54] first introduced an attention mechanism into the field of NLP. Previously, the main architecture for neural machine translation models was the sequence-to-sequence (Seq2Seq) framework, and the main bottleneck of this model framework was its intermediate transformation of the fixed dimensional sizes of vectors. To solve this problem, a Seq2Seq + attention model was used for machine translation, achieving excellent results. Vaswani et al. [47] completely discarded common network structures such as recurrent neural networks (RNNs) and CNNs for machine translation tasks, used the attention mechanism only for machine translation tasks, and achieved excellent results. Since then, the attention mechanism has become a hot research issue.

3.2. *Transfer learning*

With the rapid development of machine learning, an increasing number of machine learning application scenarios have emerged. The better performance of supervised learning is driven by a large amount of data, but in some domains, such data are often hard to obtain or too small to support the training of a good model; thus, transfer learning was born. Transfer learning models are mostly pretrained on large-scale datasets (e.g., ImageNet [55]) and then fine-tuned for downstream tasks. This idea has been successfully applied to many scenarios, such as dense pose recognition [56], image classification [57], and language understanding [48]. Transfer learning also has many applications in the medical field, such as tuberculosis detection [58], chest X-ray pneumonia classification [28], and breast cancer classification [59]. The main transfer learning methods that have emerged include sample-based transfer learning [60], feature-based transfer learning [61], model-based transfer learning [62], homogeneous transfer learning [63], heterogeneous transfer learning [64], and adversarial transfer learning [65]. Among them, homogeneous transfer learning and heterogeneous transfer learning are the main methods used in this paper for the classification of COVID-19.

Domains and tasks are two common basic concepts in transfer learning. A domain D in transfer learning contains two parts, i.e., a feature space X and a marginal probability distribution $P(X)$; the domain is given by:

$$D = \{X, P(X)\} \tag{1}$$

A task T in transfer learning consists of two parts, i.e., a label space Y and a target prediction function $f(X)$, which can also be viewed as the conditional probability $P(Y \mid X)$; this is given by:

$$T = \{Y, P(Y \mid X)\} \tag{2}$$

Among the source domain and the target domain in transfer learning, the former is the domain used to train the model or the tasks, and the latter is the machine learning domain used to predict, classify, and cluster the data by using the former model.

A generalized definition of nonuniform transfer learning, schematically shown in Fig. 2, is as follows:

- Condition: Given a source domain Ds and a learning task on the source domain Ts, a target domain Dt and a learning task Tt on the target domain.
- Goal: Use the knowledge of Ds and Ts to improve the learning of the prediction function on the target domain $f(\cdot)$.
- Restrictions: $Ds \neq Dt, Ts \neq Tt$.

3.2.1. *Homogeneous transfer learning*

If the source and target domain data have the same or similar representation structures but obey different probability distributions, i.e., the feature spaces of the source and target domains are the same, ($\mathcal{X}^s = \mathcal{X}^t$) and have the same dimensionality ($d^s = d^t$), homogeneous transfer learning is applicable. Both data- and model-based transfer learning belong to homogeneous transfer learning. The shortcoming of homogeneous transfer learning lies in its ability to improve the generalization of the target domain with only the help of the source domain of the homogeneous representation space.

3.2.2. *Heterogeneous transfer learning*

In this case, the source and target domains have different feature spaces or different feature representations, i.e., $\mathcal{X}^s \neq \mathcal{X}^t$. For example, if the source domain is a dataset from the generic domain and the target domain is a proprietary dataset, the feature spaces have different dimensions, i.e., $d^s \neq d^t$. For the scenario of medical image classification, suppose that the source domain is an ImageNet dataset about the aspects of our daily life and that the target domain is a CT scan image regarding tuberculosis. In general, these domains contain different features and possess different feature dimensions. In heterogeneous transfer learning, the source domain generally possesses richer labeling samples, while the target domain is unlabeled or has a small number of labeled samples, and this approach overcomes the shortcomings of homogeneous transfer learning by allowing the domains to be different.

3.3. *SA and MSA*

Attention mechanism is essentially derived from the visual attention mechanism of humans. Humans do not always focus their attention on the whole of something when viewing it, but rather on particular parts that interest them. Moreover, when people find that a scene often shows a certain part of something they want to observe, they perform learning to focus their attention on that part when a similar scene occurs again in the future. SA is a class of attention mechanism that goes through different parts of the same sample to obtain the part that should be focused on. SA has various forms, and the

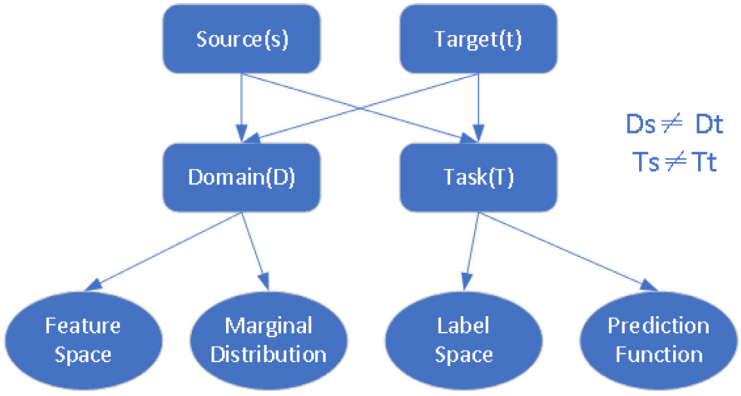


Fig. 2 – Schematic diagram of transfer learning.

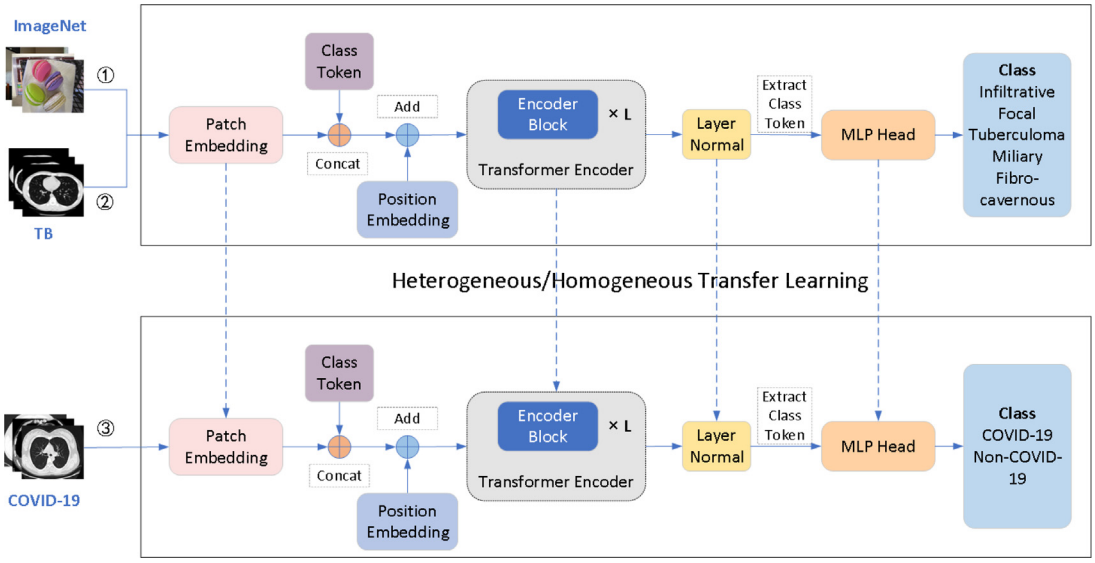


Fig. 3 – Overall structure of the model.

generic converter relies on the scaled dot product form shown in Fig. 4. In the SA layer, the input vector is first converted into three different vectors, i.e., a query matrix Q, a key matrix K and a value matrix V. Then, the weights of these vectors are obtained by the dot product of Q and each K. After normalizing these weights by using a softmax function, finally, the weights and their corresponding key values V are weighted and summed to obtain the final attention value. The function is calculated as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{3}$$

where d_k is the dimensionality of the vector and $\sqrt{d_k}$ normalizes the vector.

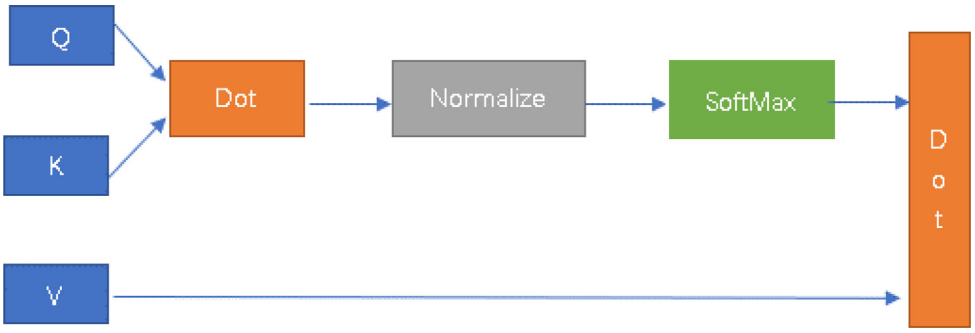


Fig. 4 – SA structure.

MSA is the core component of the transformer. Its structure is shown in Fig. 5, which differs from that of SA in that the input of MSA is divided into many small parts. Then, the scaled dot product of each input is calculated in parallel, and all the attention outputs are concatenated to obtain the final result. The MSA equation is expressed as follows:

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (4)$$

$$\text{MSA}(Q,K,V) = \text{Concat}\left(\text{head}_i, \cdots, head_i\right)W^O \quad (5)$$

where W_i^Q, W_i^K, W_i^V, and W^O are trainable parameter matrices.

3.4. *MLP and embedding layers*

In [17], the MLP head immediately follows the MSA and is composed of a fully connected layer (Linear), and the final class token passes through its output prediction class; moreover, an MLP block is also present in the transformer encoder. The transformer encoder consists of a stack of encoder blocks, and the transformer encoder is also alternatively composed of MSA and MLP blocks. In the transformer encoder, layer normalization (LN) is applied before each block, and residual concatenation is used after each block. One of the MLP blocks is composed of a nonlinear activation function (a Gaussian error linear unit (GELU)), a dropout layer, and two fully connected layers. The structural composition is shown in Fig. 6.

The input image is split into fixed-size patches and then convolved, and the resulting vector is flattened and mapped to the corresponding size dimension by using a trainable linear projection.

4. Experiments

4.1. *Datasets*

4.1.1. *COVID-19 dataset*

The COVID-19 dataset required for the experiments in this paper is obtained from [6], with a total of 746 CT scans. In the dataset, 349 of these COVID-19-positive CT images were collected from papers on COVID-19 in medRxiv and bioRxiv, and the other 397 COVID-19-negative CT images were obtained from PubMedCentral (a search engine) and MedPix (a publicly available online medical image database containing CT scans of various diseases). The data distributions for these two categories are shown in Table 1.

4.1.2. *TB dataset*

The TB dataset is obtained from the ImageCLEF2021 challenge, which was used to classify TB cases into five main types: (1) infiltrative; (2) focal; (3) tuberculoma; (4) miliary; and (5) fibro-cavernous. We use 917 3D CT scans, which are stored in the NIFTI file format. Each slice has an image size of 512 × 512 pixels, and the number of slices is approximately 100; the dataset distribution is shown in Table 2 below.

4.2. *Preprocessing*

The images are first matched one by one with the category labels, and then each image is resampled to (224, 224). Thus, the final 746 CT scans are all processed to sizes of (3, 224, 224). The data are enhanced using horizontal flipping, vertical flipping, rotation, brightening, and darkening operations. The distributions of the data before and after enhancement are shown in Table 3 and the brackets indicate the data after augmentation.

4.3. *Experimental setup and evaluation criteria*

In our experiments, we randomly divide the dataset into a training set (598) and a test set (148) based on a ratio of 8:2, where data augmentation is used on the training set. The distributions of the dataset before and after augmentation are shown in Table 3. The training set is then used to optimize the parameters. We set the optimizer (stochastic gradient descent (SGD)) learning rate to 0.01, the momentum to 0.9, and the weight decay to $5x10^{-5}$. Training is conducted for a total of 30 rounds.

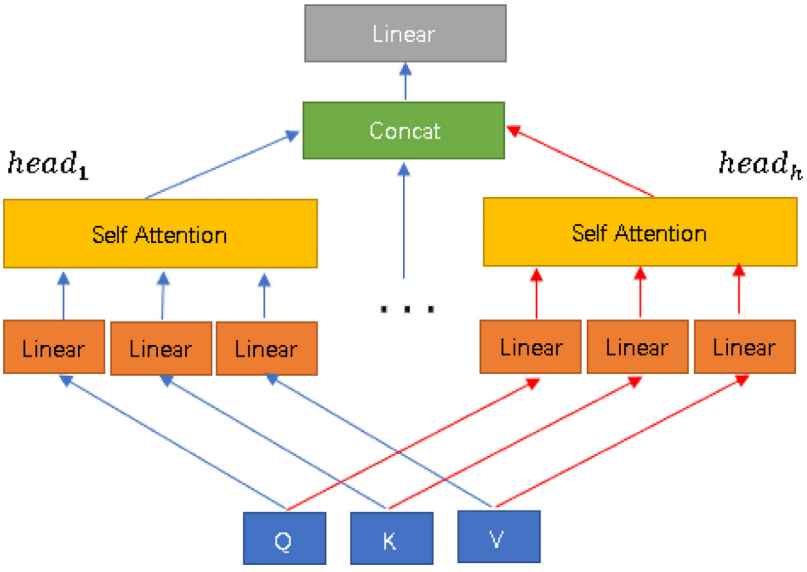


Fig. 5 – Multiheaded SA structure.

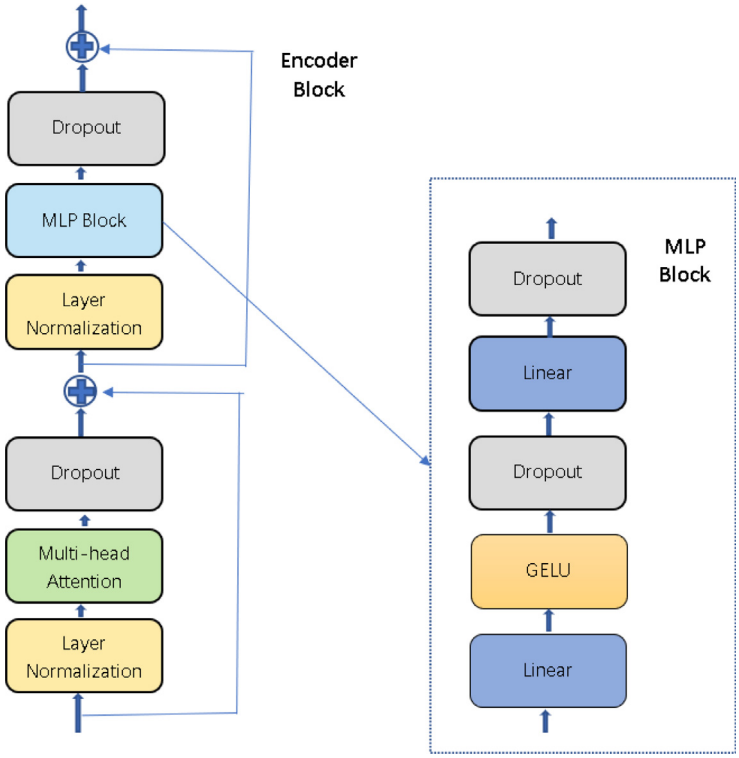


Fig. 6 – Transformer block structure.

Table 1 – COVID-19 dataset table.

categories	quantities
COVID-19	349
NonCOVID-19	397

Table 2 – TB dataset.

Types	Amounts
Infiltrative	420
Focal	226
Tuberculoma	101
Miliary	100
Fibro-cavernous	70

Table 3 – Distributions before and after data enhancement.

Type	Categories		Total
	COVID-19	NonCOVID-19	
Train	280(2520)	318(2862)	598(5382)
Test	69	79	148
Total	349(2589)	397(2941)	746(5530)

The main objective of this paper is to classify COVID-19. Classification results can be positive (infected with COVID-19) or negative (not infected with COVID-19). The predicted outcome of each classification match may or may not align with the actual category. In this setup, it is assumed that a true positive (TP) indicates an actual COVID-19 case and that the CT image can be correctly classified as COVID-19. A false positive (FP) indicates that an actual nonCOVID-19 case is incorrectly classified as COVID-19. A true negative (TN) indicates an actual nonCOVID-19 case whose CT images can be correctly classified as nonCOVID-19. Finally, an FN indicates that the actual COVID-19 case is incorrectly classified as nonCOVID-19.

The performance of the model is evaluated with several commonly used evaluation criteria, i.e., the accuracy (Acc), precision (Precision), recall (Recall), and F1 value. These metrics are defined below:

$$\text{Acc} = \frac{TP + TN}{TP + FP + +TN + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{9}$$

In medical research, especially for major infectious diseases such as COVID-19, it is very important to reduce the numbers of FP and FN results in the modeling process. In particular, FNs should be minimized to avoid classifying any

COVID-19 patients as nonCOVID-19 pneumonia patients, as coronavirus may cause more significant damage. Of course, it is also important to minimize the FP rate to avoid the unnecessary waste of manpower and resources.

4.4. Classification model comparison

We evaluate DenseNet169 [12], ResNet101 [22] and ResNet34 [22] on the COVID-19 dataset and compare them with the ViT model. And they all have pretrained on ImageNet dataset. The experimental results are shown in Table 4. The number of model training rounds is set to 30. From the experimental results, ViT achieves accuracy improvements of 3.37 %, 4.05 %, and 5.4 % over those of DenseNet169, ResNet101, and ResNet34, respectively. Moreover, compared to these models, the ViT converges more quickly, and the area under the receiver operating characteristic curve (Auc) = 0.9545, which means that the COVID-19 classification model based on the ViT performs better. Through these comparative experiments, the ViT model is still more effective than other classification models in this area of image classification.

4.5. Comparison of transfer models

We first use the TB dataset, whose construction is shown in Table 2. Since the pretrained model is obtained via training on the ImageNet dataset [27], which belongs to the general-purpose domain, and the dataset used in this experiment consists of the COVID-19 CT images in the specialized domain, this scenario involves heterogeneous transfer learning. Since these CT images are stored in the NIFTI file format, we first slice them, train the obtained 110,000 CT images by using the pretrained model, and save the best-performing parametric model. Several types of slices are shown in Fig. 7.

We train the best-performing model obtained through heterogeneous transfer learning as a pretrained model on the COVID-19 dataset, which is distributed as shown in Table 1. The comparison results obtained in our experiments are shown in Table 5. Among the tested methods, ViT-HTL indicates heterogeneous transfer learning, and TL-Med indicates further homogeneous transfer learning on top of heterogeneous transfer learning. And ResNet34-TL, ResNet101-TL and DenseNet169-TL indicate that they are pre-trained on the TB dataset, and then fine-tuned on the COVID-19 dataset. It is shown that homogeneous transfer learning yields better results. The main reason for this is that the structures of the TB and COVID-19 datasets are similar, i.e., the feature spaces of the source and target domains are similar, and the features learned by the models are also similar.

4.6. Ablation experiments

In this subsection, for the overall TL-Med model, we explore the impact of changing the corresponding settings on the model performance; in other words, we explore the impacts of whether the pretrained model is frozen, whether the prelogit module (PL) which consists of a linear layer and a nonlinear activation function(Tanh) is increased, and a before-and-after comparison of the enhancing data on the resulting model.

From Table 6, concerning the freezing of the pretraining model, we can see that the pretraining models without freezing are generally better than those that freeze. The accuracy is improved by 6.08 %, 4.06 %, and 7.43 % under the same experimental settings by only changing the freeze setting of the pretrained model, which shows that changes in the pretrained model settings have great impacts on the model performance. It is not difficult to understand that the freezing operation involves freezing all the previous layers and training only the MLP head module, i.e., train the classification output layer, whose model has fewer adjustable parameters and therefore results in limited model performance improvement. The operation without freezing uses the pretraining weights as the initial weights, and all of the weights participate in the training of the model, so the overall effect is better.

To achieve greater learning and fitting capabilities for the network model and to strengthen the representation capability of the network, we add the PL module before the last layer (linear layer), where the PL and the linear layer constitute the MLP head; this is denoted as the linear + tanh activation function. That is, the MLP head consists of two linear layers and a nonlinear activation function. Compared with Method Ⅰ and Method III, this approach achieves improved experimental results by adding the PL module, with a 2.02 % improvement in accuracy. Compared with Method Ⅱ and Method Ⅳ, this approach does not achieve obviously improved accuracy, but its recall is improved by 1.26 %, and the F1 value is increased by 0.11 %. The improvement in recall, which indicates a reduction in FNs, allows us to minimize the risk of the model diagnosing a COVID-19 patient as a nonCOVID-19 patient during the test identification process for infectious diseases, especially for a major infectious disease such as COVID-19.

The distribution of the data after data augmentation is shown in Table 3. Pretraining is viewed as a major modeling technique in CV, in which pretraining is always applied to one dataset for use with another dataset. From Table 6, we can see that when the pretraining weights are frozen, the performance of the model decreases by 1.35 % when utilizing data augmentation. The reason for this phenomenon may be that after the pretraining model of the network is frozen,

Table 4 – Comparison among the pretrained models.

Model	Acc.	Precision	Recall	F1	Auc	Time (min)
DenseNet169	0.8649	0.9538	0.7848	0.8611	0.9274	8
ResNet101	0.8581	0.9028	0.8228	0.8609	0.9176	15
ResNet34	0.8446	0.8500	0.8608	0.8553	0.9439	10
ViT	0.8986	0.9103	0.8987	0.9045	0.9545	9.4

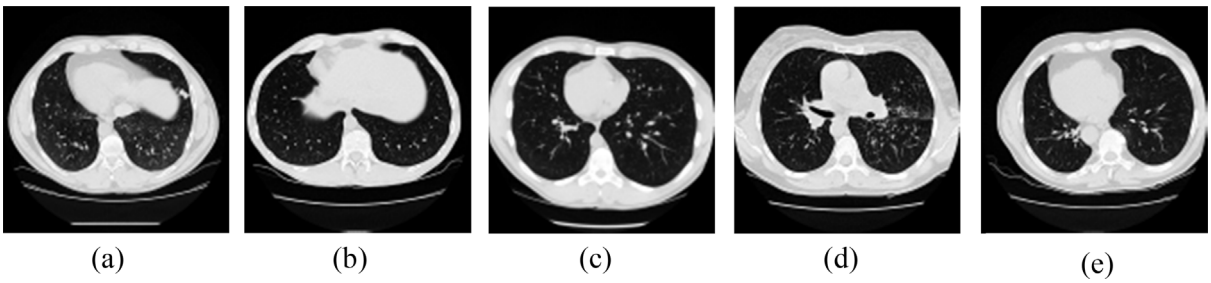


Fig. 7 – Five types of TB slices.

Table 5 – Transfer learning comparison.

Model	Acc.	Precision	Recall	F1	Auc	Time
ResNet34-TL	0.777	0.8833	0.6709	0.7626	0.8556	9
ResNet101-TL	0.7635	0.7683	0.7975	0.7826	0.7666	10
DenseNet169-TL	0.8919	0.8795	0.9241	0.9012	0.9532	10
ViT-HTL	0.8986	0.9103	0.8987	0.9045	0.9545	9.4
TL-Med	0.9122	0.9459	0.8861	0.9150	0.9606	9.3

Table 6 – Ablation experiments.

Model	Acc.	Precision	Recall	F1	Auc	Time
TL-Med + freeze	0.8514	0.8800	0.8354	0.8571	0.9066	7.2
TL-Med + no-freeze	0.9122	0.9459	0.8861	0.9150	0.9606	9.3
TL-Med + freeze + PL	0.8716	0.9054	0.8481	0.8758	0.9197	7.1
TL-Med + no-freeze + PL	0.9122	0.9342	0.8987	0.9161	0.9576	9.3
TL-Med + freeze + PL + dataAug	0.8581	0.8919	0.8354	0.8627	0.9145	337
TL-Med + no-freeze + PL + dataAug	**0.9324**	**0.9600**	**0.9114**	**0.9351**	**0.9686**	**364.9**

the use of the data augmentation technique adds more labeled data to the training process, which has a bad effect on pretraining and then reduces the effect of pretraining. In Fig. 8, we can see that the accuracy achieved on the test set before data augmentation increases more; however, it increases slowly after data augmentation or even becomes gradual, and the overall increase is less than that before data augmentation. Therefore, we can notice that the use of the data augmentation technique adds more labeled data but has some negative effects on pretraining, which directly results in a decrease in model accuracy. That is, data augmentation and the use of more labeled data in the pretraining mode do not necessarily improve the performance of the model. Zoph et al. [66] studied the problems associated with pretraining in detection and segmentation tasks and suggested that using more powerful data augmentation techniques and more labeled data reduced the role of pretraining. However, by comparing Method IV and Method VI, the method improves the performance of the model by increasing the amount of data through data augmentation techniques. Therefore, with the pre-trained model frozen, the limited number of trainable parameters limits the ability of the model to learn more data-relevant features through the data augmentation technique. It can be seen that the number of trainable parameters has an impact on the data augmentation technique. After employing the data augmentation techniques without freezing the model, we can see that the accuracy, precision, recall, F1 value and Auc of the model are substantially improved, which implies that increasing the number of trainable parameters and combining them with more labeled data may effectively contribute to improving the model performance.

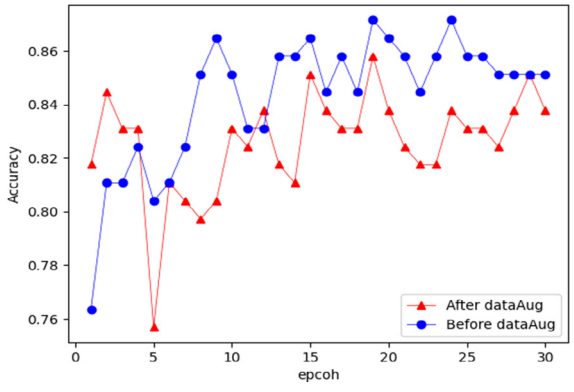


Fig. 8 – Test set accuracies before and after data enhancement experiments.

5. Discussion

In the past period of time, many researchers have studied the detection of COVID-19, and developed some COVID-19 detection models based on deep learning. Table 7 compares our proposed method with the existing literature.Due to the different datasets, validation methods, and some corresponding evaluation metrics, a fair comparison with the model results proposed in the literature cannot be made. However, it is worth noting that our proposed method achieves relatively good performance on a dataset size of 746 CT images. The number of images we use is less than that used by other methods. The method by Xiao et al.[28] achieved high accuracy on the two datasets respectively. The method by Narendra et al.[67] achieved 99.12 % accuracy, 99 % recall and 99 % f-score in a balanced dataset of 400 COVID-19 and 400 normal images. The method of Nayeeb et al.[68] achieves 99.39 % precision, 99.39 % recall and 99.19 % f-score using a limited dataset. Mahesh et al.[69] achieved an accuracy of 98.30 %, a recall of 98.31 and an f-score of 98 % on a relatively large dataset. Bejoy et al.[31] method was able to achieve an accuracy of 91.15 % on the first dataset and 97.43 % on the second dataset containing only 71 COVID-19 images and 7nonCOVID images. Shome et al.[41] achieved 93.2 % precision and 96.09 % recall. The training of the model is a process of continuous adjustment and optimisation. Ruochi et al.[32] achieved 91.08 % accuracy on a small datasets.Mohammad et al.[29] achieved an accuracy of 99.11 % by continuously optimising the model to obtain the best combination of model hyperparameters. However, Panwar et al.[70] achieved an accuracy of 88.1 % on a balanced dataset with experimental results obtained on a relatively small dataset. And from the table, we can also conclude that the size of the data has a certain impact on the performance improvement of the model.The studies by [67 69,70] were based on popular convolutional neural network architectures such as VGG16, ResNet50, Xception. [30,68] used a two-stage training scheme, which differed from the training scheme used in this paper in that they both separate network architectures, and both perform transfer learning on relevant datasets closely related to COVID-19.

The model proposed in this paper is developed with the aim of being used in clinical conditions for detecting COVID-19 patients from their chest CT images.Based on this, the model can be used to assist specialist practitioners in rapid diagnostic screening during an outbreak. Therefore, the purpose of this model is to rapidly screen COVID-19 from other diseases. From the experimental results, our model can be used for initial rapid screening of suspicious people and can provide effective assistance to front-line medical staff to further improve the efficiency of detection of COVID-19

Table 7 – Comparison of our proposed method with the existing literature.

Method	dataset	Acc.	Precision	Recall	F1	Auc
Xiao et al.[28]	Test data 1: 2567COVID-19, 2567normal, 2567 Pneumonia	0.9557	0.99	0.99	0.99	–
	Test data 2: 756COVID-19, 6284normal, 3478 Pneumonia	0.9444	0.95	0.95	0.95	–
Narendra et al.[67]	400 COVID-19, 400 normal	0.9912	0.99	0.99	0.99	–
Nayeeb et al.[68]	408COVID-19, 816Non-COVID	0.9939	0.9919	0.9939	0.9919	–
Mahesh et al.[69]	2249COVID-19, 2396no-Findings	0.983	–	0.9831	0.98	0.999
Bejoy et al.[31]	Test data 1: 453COVID-19, 497 Non-COVID	0.9115	0.853	0.985	0.914	0.963
	Test data 2: 71COVID-19, 7non-COVID	0.9743	0.986	0.986	0.986	0.911
Govardhan et al.[30]	250COVID-19, 965 other	0.9777	0.9714	0.9714	–	–
Shome et al.[41]	10819COVID-19, 10,314 normal	0.9320	–	0.9609	–	–
Ruochi et al.[32]	189COVID-19, 63 other, 235 normal	0.9108	–	–	–	–
Mohammad et al.[29]	184COVID-19, 5000 normal	0.9911	–	–	–	–
Panwar et al.[70]	142COVID-19, 142Non-COVID	0.881	–	–	–	0.881
Proposed Method	349COVID-19, 397Non-COVID	0.9324	0.96	0.9114	0.9351	0.9686

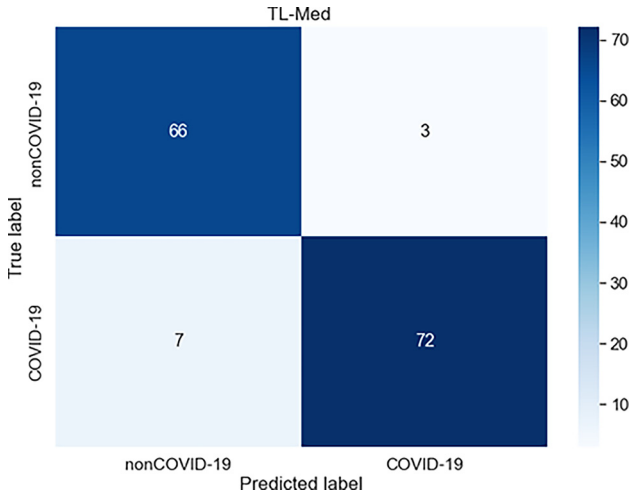


Fig. 9 – The confusion matrix of the proposed TL-Med framework.

and stop the spread of COVID-19. And as shown in the confusion matrix in Fig. 9, the proposed model in this paper produces more false negatives, which will be a direction for our future improvement. As to whether the results of the deep learning method constitute a reliable diagnosis, due to the rigorous nature of medicine, we will conduct experiments on additional COVID-19 datasets in the future to further demonstrate the performance of our model.

6. Conclusion

In the past few decades, machine learning has developed rapidly and has been applied in many industries and fields. Since medical image recognition is the most basic and difficult problem in the medical field, the use of computers to assist doctors in identifying and detecting cases is a common application of machine learning. In this paper, a two-stage transfer learning model (TL-Med) based on the ViT is proposed to detect and identify CT data. To detect COVID-19 effectively, we first perform pretraining on the TB dataset, where the aim is to obtain medical features and use the best obtained results as a pretrained model, and then we train and test the resulting models on the COVID-19 dataset. To overcome the problem of data scarcity, we use a data augmentation technique. The data augmentation effectively improves the training of TL-Med by enabling it to learn more classification features and learning parameters from the rich training dataset. The classification model can effectively aid clinicians in the detection and identification of COVID-19.

In the medical field, data scarcity is a common phenomenon. In the future, we plan to use TL-Med to explore other medical image detection and classification scenarios, such as breast cancer, brain tumors, and diabetic retinopathy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No.61876031), Natural Science Foundation of Liaoning Province, China (20180550921, 2019-ZD-0175) and Scientific Research Fund Project of the Education Department of Liaoning Province, China (LJYT201906).

REFERENCES

[1] Sohrabi C, Alsafi Z, O'neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Internat J Surg 2019;76(2020):71–6.
[2] Păcurar C-M, Necula B-R. An analysis of COVID-19 spread based on fractal interpolation and fractal dimension. Chaos, Solitons Fractals 2020;139 110073.
[3] Sun J, Zhuang Z, Zheng J, Li K, Wong RL-Y, Liu D, Huang J, He J, Zhu A, Zhao J. Generation of a broadly useful model for COVID-19 pathogenesis, vaccination, and treatment. Cell 2020;182. 734–743. e735.
[4] Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 2020;296:E115–7.
[5] Chua F, Armstrong-James D, Desai SR, Barnett J, Kouranos V, Kon OM, et al. The role of CT in case ascertainment and management of COVID-19 pneumonia in the UK: insights from high-incidence regions. Lancet Respir Med 2020;8:438–40.
[6] Hu Q, Guan H, Sun Z, Huang L, Chen C, Ai T, et al. Early CT features and temporal lung changes in COVID-19 pneumonia in Wuhan, China. Eur J Radiol 2020;128 109017.
[7] J. Zhao, Y. Zhang, X. He, P. Xie, Covid-ct-dataset: a ct scan dataset about covid-19, arXiv preprint arXiv:2003.13865, 490 (2020). https://github.com/UCSD-AI4H/COVID-CT.
[8] Li J, Wu Y, Shen N, Zhang J, Chen E, Sun J, et al. A fully automatic computer-aided diagnosis system for hepatocellular carcinoma using convolutional neural networks. Biocyber Biomed Eng 2020;40:238–48.
[9] Kurzyński M, Majak M, Żołnierek A. Multiclassifier systems applied to the computer-aided sequential medical diagnosis. Biocyber Biomed Eng 2016;36:619–25.
[10] Munusamy H, Karthikeyan J, Shriram G, Revathi ST, Aravindkumar S. FractalCovNet architecture for COVID-19 Chest X-ray image classification and CT-scan image segmentation,. Biocyber Biomed Eng 2021;41:1025–38.
[11] Kassania SH, Kassanib PH, Wesolowskic MJ, Schneidera KA, Detersa R. Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach. Biocyber Biomed Eng 2021;41:867–79.
[12] 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. p. 770–8.
[13] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 4700–8.
[14] Heidarian S, Afshar P, Enshaei N, Naderkhani F, Rafiee MJ, Fard FB, et al. Covid-fact: A fully-automated capsule network-based framework for identification of covid-19 cases from chest ct scans. Front Artif Intell 2021;4.
[15] Chaudhary S, Sadbhawna S, Jakhetiya V, Subudhi BN, Baid U, Guntuku SC. Detecting covid-19 and community acquired pneumonia using chest CT scan images with deep learning.

In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 8583–7.
[16] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, PMLR. p. 6105–14.
[17] He X, Yang X, Zhang S, Zhao J, Zhang Y, Xing E, et al. Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. IEEE 2020.
[18] Polsinelli M, Cinque L, Placidi G. A light CNN for detecting COVID-19 from CT scans of the chest. Pattern Recogn Lett 2020;140:95–100.
[19] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size, arXiv preprint arXiv:1602.07360, (2016).
[20] Sani S, Shermeh HE. A novel algorithm for detection of COVID-19 by analysis of chest CT images using Hopfield neural network. Expert Syst Appl 2022;197 116740.
[21] Scarpiniti M, Ahrabi SS, Baccarelli E, Piazzo L, Momenzadeh A. A novel unsupervised approach based on the hidden features of Deep Denoising Autoencoders for COVID-19 disease detection. Expert Syst Appl 2022;192 116366.
[22] Pathak Y, Shukla PK, Tiwari A, Stalin S, Singh S. Deep transfer learning based classification model for COVID-19 disease. Irbm 2020.
[23] Rezende E, Ruppert G, Carvalho T, Ramos F, De Geus P. Malicious software classification using transfer learning of resnet-50 deep neural network. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE; 2017. p. 1011–4.
[24] Jaiswal A, Gianchandani N, Singh D, Kumar V, Kaur M. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. J Biomol Struct Dyn 2020:1–8.
[25] Loey M, Manogaran G, Khalifa NEM. A deep transfer learning model with classical data augmentation and cgan to detect covid-19 from chest ct radiography digital images. Neural Comput Appl 2020:1–13.
[26] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784, (2014).
[27] Aslan MF, Unlersen MF, Sabanci K, Durdu A. CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection. Appl Soft Comput 2021;98 106912.
[28] Qi X, Brown LG, Foran DJ, Nosher J, Hacihaliloglu I. Chest X-ray image phase features for improved diagnosis of COVID-19 using convolutional neural network. Int J Comput Assist Radiol Surg 2021;16:197–206.
[29] Khishe M, Caraffini F, Kuhn S. Evolving deep learning convolutional neural networks for early COVID-19 detection in chest X-ray images. Mathematics 2021;9:1002.
[30] Jain G, Mittal D, Thakur D, Mittal MK. A deep learning approach to detect Covid-19 coronavirus with X-Ray images. Biocyber Biomed Eng 2020;40:1391–405.
[31] Abraham B, Nair MS. Computer-aided detection of COVID-19 from X-ray images using multi-CNN and Bayesnet classifier. Biocyber Biomed Eng 2020;40:1436–45.
[32] Zhang R, Guo Z, Sun Y, Lu Q, Xu Z, Yao Z, et al. COVID19XrayNet: a two-step transfer learning model for the COVID-19 detecting problem based on a limited number of chest X-ray images, Interdisciplinary Sciences: Computational. Life Sci 2020;12:555–65.
[33] Minaee S, Kafieh R, Sonka M, Yazdani S, Soufi GJ. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. Med Image Anal 2020;65 101794.
[34] Hassantabar S, Ahmadi M, Sharifi A. Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches. Chaos Solitons Fractals 2020;140 110170.
[35] Gupta P, Siddiqui MK, Huang X, Morales-Menendez R, Pawar H, Terashima-Marin H, et al. COVID-WideNet—A capsule network for COVID-19 detection. Appl Soft Comput 2022;108780.
[36] Goel T, Murugan R, Mirjalili S, Chakrabartty DK. Multi-COVID-Net: Multi-objective optimized network for COVID-19 diagnosis from chest X-ray images. Appl Soft Comput 2022;115 108250.
[37] Jalali SMJ, Ahmadian M, Ahmadian S, Hedjam R, Khosravi A, Nahavandi S. X-ray image based COVID-19 detection using evolutionary deep learning approach. Expert Syst Appl 2022;116942.
[38] Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? Adv Neural Inform Process Syst 2021;34.
[39] Park S, Kim G, Oh Y, Seo JB, Lee SM, Kim JH, et al. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. Med Image Anal 2022;75 102299.
[40] Li J, Yang Z, Yu Y. A medical AI diagnosis platform based on vision transformer for coronavirus. In: 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI). IEEE; 2021. p. 246–52.
[41] Shome D, Kar T, Mohanty SN, Tiwari P, Muhammad K, AlTameem A, et al. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. Int J Environ Res Public Health 2021;18:11086.
[42] S. Park, G. Kim, J. Kim, B. Kim, J.C. Ye, Federated Split Vision Transformer for COVID-19CXR Diagnosis using Task-Agnostic Training, arXiv preprint arXiv:2111.01338, (2021).
[43] Krishnan KS, Krishnan KS. Vision transformer based COVID-19 detection using chest X-rays. In: 2021 6th international conference on signal processing, computing and control (ISPCC). IEEE; 2021. p. 644–648..
[44] S. Park, G. Kim, Y. Oh, J.B. Seo, S.M. Lee, J.H. Kim, S. Moon, J.-K. Lim, J.C. Ye, Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus, arXiv preprint arXiv:2103.07055, (2021).
[45] Mondal AK, Bhattacharjee A, Singla P, Prathosh A. xViTCOS: explainable vision transformer based COVID-19 screening using radiography. IEEE J Transl Eng Health Med 2021;10:1–10.
[46] S. Park, G. Kim, Y. Oh, J.B. Seo, S.M. Lee, J.H. Kim, S. Moon, J.-K. Lim, J.C. Ye, Vision Transformer using Low-level Chest X-ray Feature Corpus for COVID-19 Diagnosis and Severity Quantification, arXiv preprint arXiv:2104.07235, (2021).
[47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inform Process Syst 2017:5998–6008.
[48] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, (2018).
[49] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European conference on computer vision. Springer; 2020. p. 213–29.
[50] Z. Liu, S. Luo, W. Li, J. Lu, Y. Wu, C. Li, L. Yang, Convtransformer: A convolutional transformer network for video frame synthesis, arXiv preprint arXiv:2011.10185, (2020).
[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An Image is Worth 16x16 words: transformers for image recognition at scale, 2020, pp. arXiv:2010.11929.

[52] Chen M, Radford A, Child R, Wu J, Jun H, Luan D, et al. Generative pretraining from pixels. In: International conference on machine learning, PMLR. p. 1691–703.
[53] Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 7794–803.
[54] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473, (2014).
[55] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L, et al. IEEE conference on computer vision and pattern recognition. IEEE 2009;2009:248–55.
[56] Sanakoyeu A, Khalidov V, McCarthy MS, Vedaldi A, Neverova N. Transferring dense pose to proximal animal classes. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. p. 5233–42.
[57] Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 1717–24.
[58] Duong LT, Le NH, Tran TB, Ngo VM, Nguyen PT. Detection of tuberculosis from chest X-ray images: boosting the performance with vision transformer and transfer learning. Expert Syst Appl 2021;184 115519.
[59] Beevi KS, Nair MS, Bindu G. Automatic mitosis detection in breast histopathology images using Convolutional Neural Network based deep transfer learning, Biocybernetics and Biomedical. Engineering 2019;39:214–23.
[60] Zhao W. Research on the deep learning of the small sample data based on transfer learning. AIP conference proceedings. AIP Publishing LLC; 2017.
[61] Cheng B, Liu M, Zhang D, Munsell BC, Shen D. Domain transfer learning for MCI conversion prediction. IEEE Trans Biomed Eng 2015;62:1805–17.
[62] Zhang W, Li R, Zeng T, Sun Q, Kumar S, Ye J, et al. Deep model based transfer and multi-task learning for biological image analysis. IEEE Trans Big Data 2016;6:322–33.
[63] Lu H, Zhang L, Cao Z, Wei W, Xian K, Shen C, et al. van den Hengel, When unsupervised domain adaptation meets tensor representations. In: Proceedings of the IEEE international conference on computer vision. p. 599–608.
[64] Shen C, Guo Y. Unsupervised heterogeneous domain adaptation with sparse feature transformation. In: Asian conference on machine learning, PMLR. p. 375–90.
[65] Saito K, Watanabe K, Ushiku Y, Harada T. Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 3723–32.
[66] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E.D. Cubuk, Q.V. Le, Rethinking pre-training and self-training, arXiv preprint arXiv:2006.06882, (2020).
[67] Mishra NK, Singh P, Joshi SD. Automated detection of COVID-19 from CT scan using convolutional neural network. Biocyber Biomed Eng 2021;41:572–88.
[68] Rashid N, Hossain MAF, Ali M, Sukanya MI, Mahmud T, Fattah SA. AutoCovNet: Unsupervised feature learning using autoencoder and feature merging for detection of COVID-19 from chest X-ray images. Biocyber Biomed Eng 2021;41:1685–701.
[69] Gour M, Jain S. Automated COVID-19 detection from X-ray and CT images with stacked ensemble convolutional neural network. Biocyber Biomed Eng 2022;42:27–41.
[70] Panwar H, Gupta P, Siddiqui MK, Morales-Menendez R, Singh V. Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. Chaos Solitons Fractals 2020;138 109944.