A Comparative Study of Transfer Learning based Models for Lung Cancer Histopathology Classification

Mo Chen *

School of Biotechnology, East China University of Science and Technology, Shanghai, 200237, China

* Corresponding author's email: y21210017@mail.ecust.edu.cn

Abstract. Lung cancer is the deadliest form of cancer, which attracted a lot of attention in the past. Transfer learning is a very popular approach in deep learning as it can apply the knowledge obtained from a previous task to improve the performance in another. In this research, four transfer learning models with different complexity are utilized to detect lung cancer, which are AlexNet, VGG16, ResNet50 and Inception-v3. Since the early detection and histopathological diagnosis can considerably decrease the likelihood of mortality, the lung cancer histopathological images dataset is considered. This dataset contains histopathological images of 3 classes, all of them are considered in this study. Firstly, the four models are trained on the histopathological database from random initialization for 10 epochs. Next, the four models are first pre-trained on ImageNet and then trained on the histopathological dataset for 10 epochs. For each epoch, the testing accuracy is recorded so as to find the optimal number of epochs and determine whether transfer learning models are capable in lung cancer detection. Then, various evaluation metrics e.g., accuracy and precision are used to measure and compare the four models' performance. The study's finding shows that AlexNet, VGG16, ResNet50 and Inception-v3 pre-trained on ImageNet are adequate in lung cancer detection. There corresponding accuracy rates are 99.367%, 99.800%, 100% and 100% respectively, which are much higher than that trained from random initialization. Among the four transfer learning models, ResNet50 and Inception-v3 perform the best on lung cancer classification.

Keywords: Lung Cancer; ImageNet; AlexNet; VGG16; ResNet50 and Inception-v3.

1. Introduction

Cancer is a leading cause of mortality globally that accounts for nearly 10 million deaths in 2020, which means almost one in six of the world's population would die of cancer [1]. It is reported that lung cancer, accounting for 18% of the total cancer deaths worldwide in 2020, is by far the most common cause of cancer death [2, 3]. Although lung cancer is perilous, successful treatment is possible when the tumor is small and before it has spread. That is, an early detection of lung cancer as well as the accurate diagnosis of the types are significant for the timely and successful treatment.

Lung adenocarcinoma (LUAC) is the most prevalent type of lung cancer followed by lung squamous cell carcinoma (LUSC). LUAC, which develops from the mucosal glands and makes up about 40% of all lung cancers, is a major cause of cancer death in the United States [4]. It usually grows in the bronchi that branch off of the main airways in the center of the chest.

When it comes to identifying different cancer types, the traditional way is to carry out chest X-ray first, and then implement a Magnetic Resonance Imaging (MRI) or Computed Tomography (CT) scan to better view the lungs and look for a tumor. Since various subtypes consist of different histopathological features, the doctor may take a biopsy of the tumor and classify it under microscope [5]. However, all these approaches require experienced doctors. In order to save manpower, edge contour detection and segmentation on the cancer image has been invented to cancer classification in recent years. But its popularization is greatly limited by its low accuracy. In recent years, deep learning has made a difference in cancer detection with its overwhelmed efficiency and low error-rate. For instance, Talukder et al. built a feature extraction model to identify lung cancer with accuracy rate of 99.05% [6].

It shall be noticed that transfer learning (TL), a technique that enables algorithms to learn new tasks by using pre-trained models, boast three major benefits: 1) TL offers a better initial performance;

2) TL increases learning speed; 3) TL generally results in a more accurate and effective model [7]. However, no one has used TL techniques with Convolutional Neural Network (CNN) in lung cancer detection yet.

In this regard, different TL-based models are evaluated for lung cancer classification according to histopathological images in this study. In order to find out whether CNN models pre-trained on ImageNet dataset is capable for lung cancer classification and how model complexity affects the accuracy of diagnosis, AlexNet, VGG16, ResNet50 and Inception-v3 are trained on the same dataset consisting of 12,000 total images of LUAC, LUSC and benign lung tissue (N). Firstly, the capacity of the four models that are not pre-trained on ImageNet are measured, and then the pre-trained models. The performance analysis shows that pre-trained TL models perform much better than that trained from random initialization. In a total of 10 epochs, pre-trained AlexNet, VGG16, ResNet50 and Inception-v3 achieve their best testing accuracy at epoch 9, 9, 4 and 8 respectively. Specifically, the highest accuracy rates are 99.367%, 99.800%, 100% and 100% respectively. To compare the four models' performance further, a variety of evaluation metrics e.g., precision and f1-score are calculated at the epoch with the highest accuracy. The result is that ResNet50 and Inception-v3 still gives the highest score among the four TL models. This research demonstrates that it is reasonable to apply TL to lung cancer detection, and ResNet50 and Inception-v3 outperform the other two in feature extraction.

2. Methodology

2.1 Dataset Description and Preprocessing

In this project, the lung cancer histopathological images dataset from Kaggle was considered [8]. This dataset contains 15,000 histopathological images with 3 classes, namely LUAC, N and LUSC. Each class includes 5,000 images and all of them are JPEG images of size 768×768 pixels. In addition, all of them are RGB images with three channels.

There are three steps involved in image pre-processing. Firstly, for each class, these histopathological pictures were divided into two parts – 4,000 pictures for training and 1,000 pictures for testing. Therefore, the numbers of samples in the training and testing set were 12,000 and 3,000 respectively. Secondly, each image in the dataset was resized to 299×299. Finally, both the training and the testing datasets were shuffled and separated into batches with a size of 50 images for each batch. Samples of the images are depicted in Figure 1.

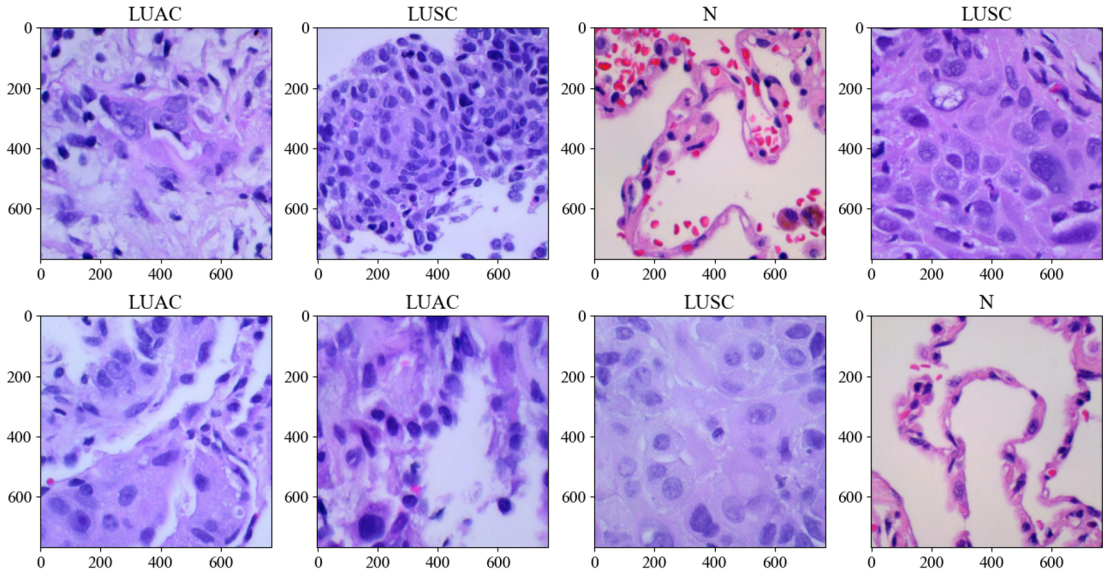

Figure 1. Samples of images from the lung cancer histopathological images dataset.

2.2 Employed CNN Models

For feature extraction, four kinds of pre-trained CNN models including AlexNet, VGG16, ResNet50 and Inception-v3 were used. For the reason that these models are pre-trained on ImageNet Dataset that has 1000 classes, the output features of the final layer were modified from 1000 to 3 in order to satisfy the requirement for training. The details of each model are introduced in the following part of this section.

2.2.1 AlexNet.

AlexNet was designed by Krizhevsky et al. [9], which contains five convolutional layers, three max pooling layers, two normalization layers, two fully connected layers and one softmax layer. Each convolutional layer consists of convolutional filters with a nonlinear activation function – ReLU. The structure of AlexNet is shown in Figure 2.

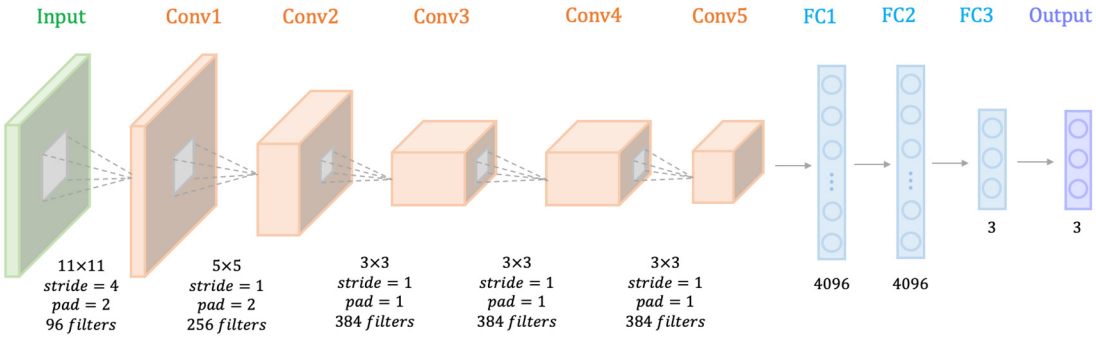

Figure 2. AlexNet architecture.

2.2.2 VGG16.

VGG16 is a variant of VGG model with 16 convolutional layers. Compared to AlexNet, VGG16 use a very small 3×3 receptive field throughout the entire network with the stride of 1 pixel, which makes the network very simple, elegant and easy to work with [10]. The architecture is illustrated in Figure 3.

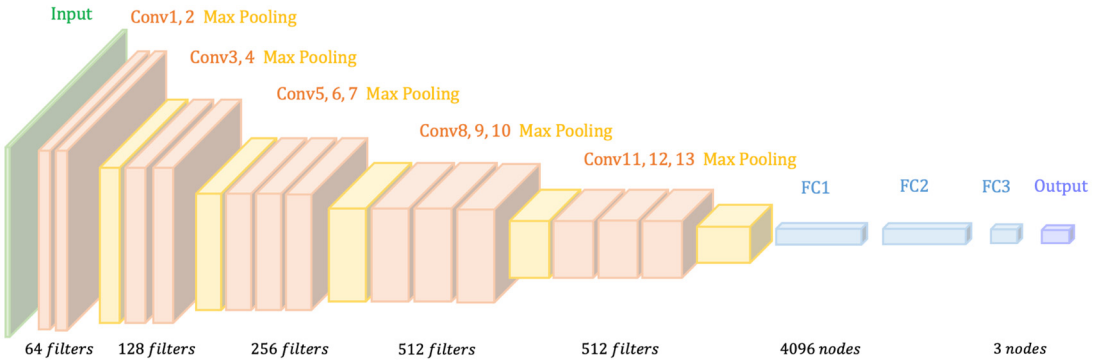

Figure 3. VGG16 architecture.

2.2.3 ResNet50.

ResNet 50 is made up of 48 convolution layers along with a max pooling layer and an average pooling layer. ResNet50 is featured by the deep residual learning framework, which is used to address the problem of degradation. The core idea of Residual Networks is the introduction of shortcut connections that simply perform identity mappings. In this way, no extra parameters are added to the model and the computational time is controlled [11]. The details of the architecture of ResNet50 is depicted in Figure 4.

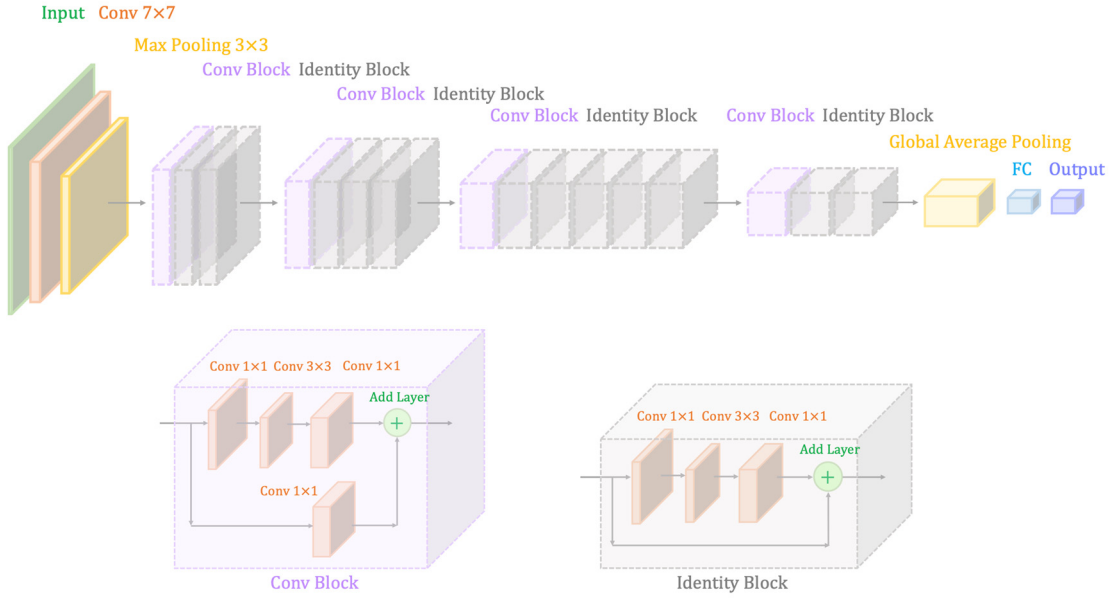

Figure 4. ResNet50 architecture.

2.2.4 Inception-v3

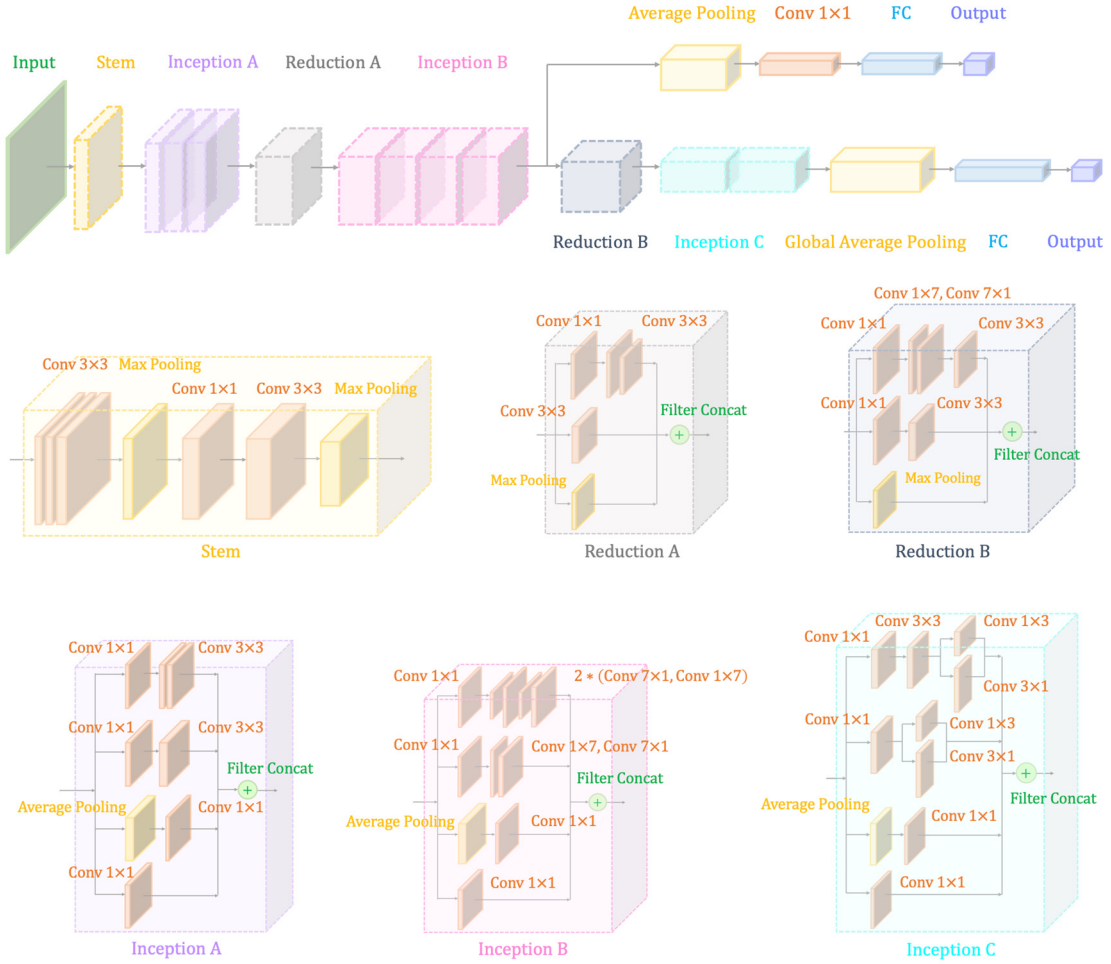

Figure 5. Inception-v3 architecture.

Inception-v3 with a total of 42 layers is an advanced version of the basic model Inception-v1 (i.e. GoogLeNet), which was released in 2014. Compared to AlexNet that has approximately 60 million parameters and VGGNet that has more than three times more parameters than AlexNet, Inception-

v3 boasts much fewer parameters, which is about 7 million. There are four main modifications done on Inception-v3 compared to its predecessors: 1) factorizing the larger convolutions into smaller ones; 2) spatially factorizing symmetric convolutions into asymmetric ones; 3) using auxiliary classifiers to combat the vanishing gradient problem in deep networks; 4) increasing the network filters' activation dimension to effectively lower the grid size. With these improvements, Inception-v3 boasts higher efficiency, deeper network and lower error rate. Its structure is shown in Figure 5.

2.3 Implementation Details

The four TL models mentioned above were trained on the lung cancer dataset for 10 epochs. To improve the models' performance with each epoch, this study used the cross-entropy as the cost function (equation (1)) and Stochastic Gradient Descent (SGD) as the optimizer (equation (2)). The details of the implement are listed in Table 1.

$$J = -\frac{1}{n}\sum_x [y\ln(a) + (1-y)\ln(1-a)] \tag{1}$$

J: cost
n: the number of samples
x: input
y: target
a: prediction

$$\theta_j' = \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta) \quad (j = 0, 1, 2, \cdots n) \tag{2}$$

θ_j: parameters
θ_j': updated parameters
α: learning rate
$J(\theta)$: cost function

Table 1. Implementation details of the training based on AlexNet, VGG16, ResNet50 and Inception-v3

Device	GPU, NVIDIA Tesla T4
CUDA Version	11.2
Deep Learning Framework	PyTorch
Image Size	299×299
Batch Size	50
Cost Function	Cross-Entropy
Optimizer	SGD
Learning Rate	0.01
Epochs	10

To appraise the performance of the four TL models, evaluation metrics such as confusion matrix, accuracy, precision, recall and f1-score are formulated in the following contents.

Confusion matrix displays the counts of TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative) of a model (Table 2). TP indicates a correctly forecasted positive value; FP indicates an incorrectly predicted positive value; TN denotes a correctly forecasted negative value; and FN denotes an incorrectly predicted negative value.

Table 2. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Accuracy is defined as the total number of predictions a model gets right. It is defined as the number of valid forecasts divided by the total number of forecasts (equation (3)).

$$Accuracy = \frac{TN+TP}{TP+FP+TN+FN} \tag{3}$$

Accuracy is the ratio of true positives to total predicted positive values. It evaluates the degree to which a model is precise in the prediction of positive labels. The formula for precision is down below:

$$Precision = \frac{TP}{FP+TP} \tag{4}$$

Recall measures the sample fraction of a class that is properly predicted by the model (equation (5)).

$$Recall = \frac{TP}{FN+TP} \tag{5}$$

F1-score is the harmonic mean for recall and precision, which is depicted as:

$$F1-score = 2 \times \frac{Recall \times Precision}{Recall+Precision} \tag{6}$$

3. Results and Discussion

Table 3. The testing accuracy of models trained from random initialization

	Accuracy (%)			
Epoch	AlexNet	VGG16	ResNet50	Inception-v3
1	33.333	34.833	33.333	54.500
2	64.633	78.233	33.333	52.033
3	65.133	75.700	33.333	53.167
4	73.367	79.267	33.333	53.600
5	71.800	81.633	33.333	53.033
6	75.567	87.533	33.333	51.900
7	77.533	83.333	78.467	52.500
8	84.700	83.333	75.967	53.833
9	85.200	90.333	72.067	53.533
10	78.167	89.700	84.167	52.367

Table 4. The testing accuracy of pre-trained TL models

	Accuracy (%)			
Epoch	AlexNet	VGG16	ResNet50	Inception-v3
1	95.700	97.200	99.133	99.233
2	93.767	97.133	93.633	99.733
3	94.233	99.067	99.867	99.833
4	97.433	99.000	100	99.933
5	98.833	93.333	100	99.967
6	97.833	70.767	99.900	99.933
7	98.900	99.433	96.833	99.967
8	83.100	97.600	99.967	100
9	99.367	99.800	100	100
10	98.300	99.767	99.967	100

Table 3 shows the testing accuracy of AlexNet, VGG16, ResNet50 and Inception-v3 that are not pre-trained on ImageNet and Table 4 elaborates that of pre-trained models in a total of 10 epochs. It can be observed that pre-trained models perform much better than models trained from random initialization. Specifically, the best accuracy rate for the four pre-trained models is 99.367%, 99.800%, 100% and 100% while the highest accuracy of modes trained from scratch are only 85.200%, 90.333%, 84.167% and 54.500% respectively. For each model, the accuracy curves of the pre-trained version and the default version (modes trained from scratch) are depicted in Figure 6. It is clear that pre-trained TL models can achieve much higher accuracy in less time. As is shown in Figure 7, the

accuracy curves of the four pre-trained models are drawn on one graph. And the result shows that in each epoch, ResNet50 and Inception-v3 give higher accuracy rates than others.

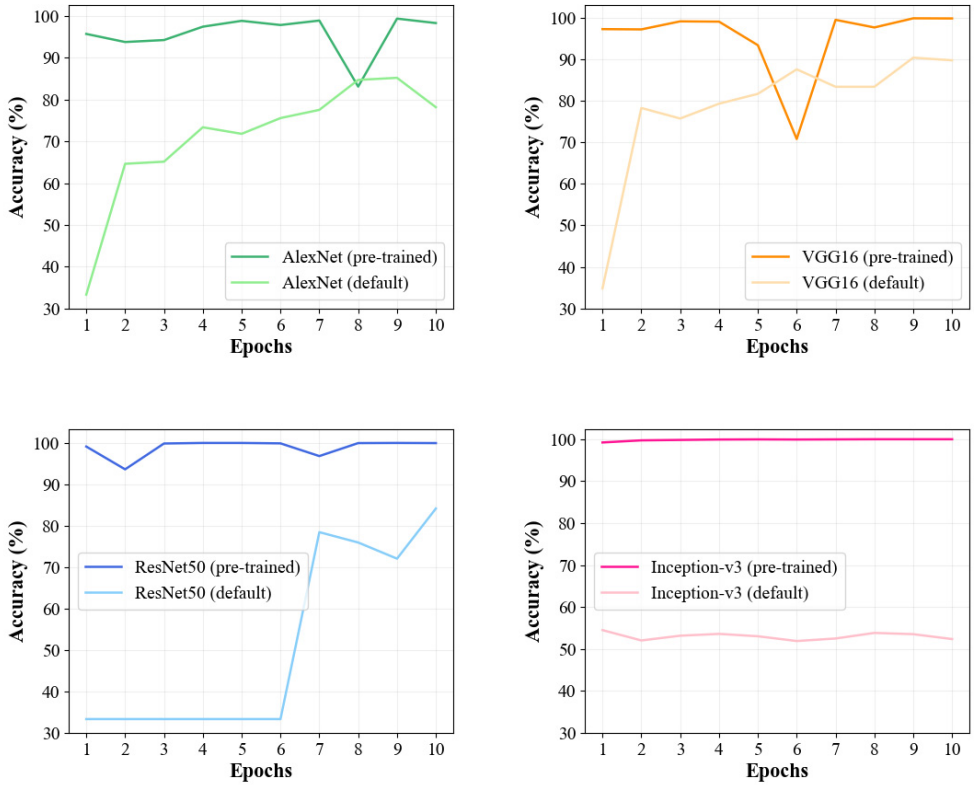

Figure 6. Comparison of the testing accuracy for default and pre-trained TL models.

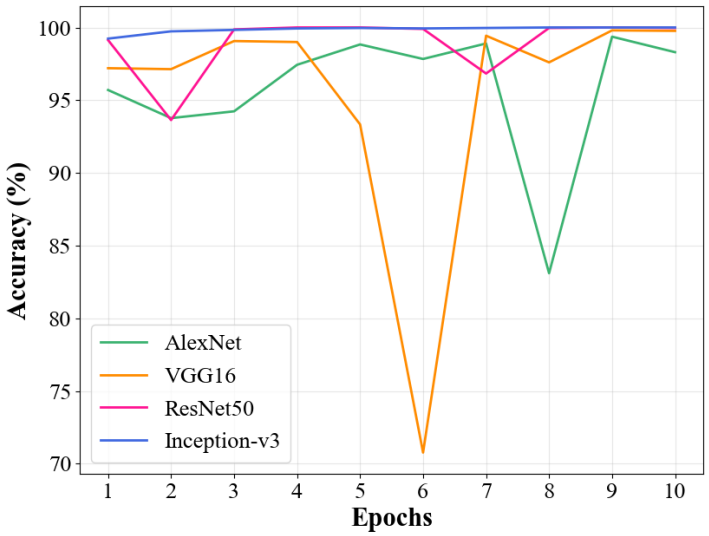

Figure 7. Illustration of the testing accuracy curves for the four pre-trained TL models.

For further assessment, other performance indicators of the pre-trained models at their best-accuracy-epoch are demonstrated in Table 5. Corresponding confusion matrixes are illustrated in Figure 8. As is shown in the performance analysis, AlexNet achieves a precision of 99.369%, recall 99.296%, f1-score 99.367%. VGG16 owns an improved precision of 99.801%, recall 99.800%, f1-score 99.800%. ResNet50 and Inception-v3 provide the best accuracy rate, precision rate, recall score and f1-score among the four TL models, which are all 100%. According to the information provided by the confusion matrix: AlexNet correctly predicts 99.4% LUAC images, 99.8% N images and 98.9%

LUSC images; VGG16 correctly predicts 99.9% LUAC images, 100% N images and 99.5% LUSC images; ResNet50 and Inception-v3 correctly predicts all of the 3,000 images.

Table 5. Evaluation of the performance of the four pre-trained TL models

	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
AlexNet	99.367	99.369	99.367	99.367
VGG16	99.800	99.801	99.800	99.800
ResNet50	100	100	100	100
Inception-v3	100	100	100	100

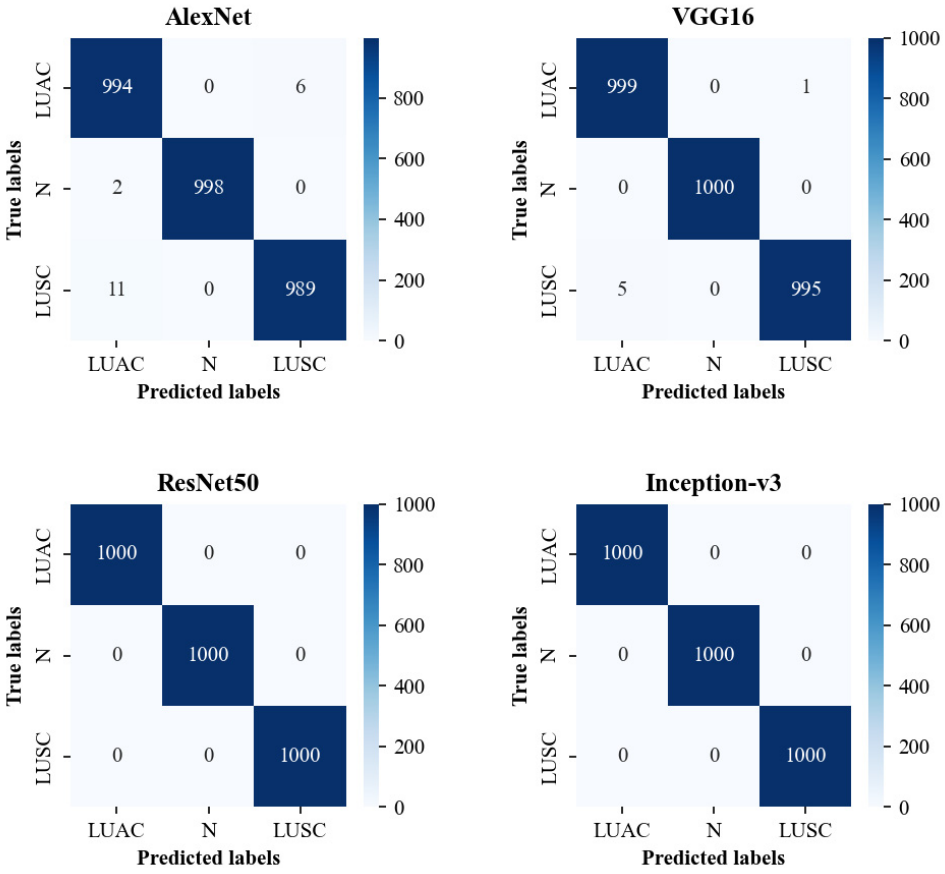

Figure 8. Confusion Matrixes of the four pre-trained TL models.

The results demonstrate that AlexNet, VGG16, ResNet50 and Inception-v3 are all capable in distinguishing histopathological images of LUAC and LUSC from N images. That is, all of the four models achieve an accuracy over 99% in 10 epochs. Besides, the performance comparison between the four models indicates that deeper neural networks generally perform better than shallower networks. To be specific, the shallowest model – AlexNet, also one of the first deep CNN models, has the worst overall performance. VGG16, with an increased number of layers and decreased kernel size in comparison with AlexNet, delivers better results. Likewise, the deeper model – ResNet50 boasts better performance than VGG16. Inception-v3 has similar complexity with ResNet50 gives the same values of accuracy, recall, precision and f1-score. Therefore, it can be observed that the depth and complexity of the network do have a profound effect on the model performance.

4. Conclusion

In this paper, four typical TL models including AlexNet, VGG16, ResNet and Inception-v3 were used for lung cancer classification. These models were trained on the lung histopathological dataset

for a total of 10 epochs and for each epoch, the testing accuracy was recorded to find the optimal epoch for each epoch. The result showed that although these models were pre-trained on ImageNet database, they provided much higher accuracy, better starting performance and faster convergence than models trained from random initialization. Specifically, AlexNet, VGG16, ResNet and Inception-v3 achieved their best accuracy at epoch 9, 9, 4 and 8 respectively, and the corresponding accuracy rates are 99.367%, 99.800%, 100% and 100% respectively. Since accuracy is not enough for evaluating CNN models' performance, precision, recall, f1-score and confusion matrix were also computed at their optimal epoch. The result found that ResNet50 and Inception-v3 still performed the best in terms of these evaluation indicators. Therefore, four conclusions can be summarized as follows:

- TL is adequate to classify histopathological images of LUAC, LUSC and N from each other;
- Pre-trained TL models have better overall performance that models trained from scratch;
- ResNet50 and Inception-v3 are best suited for detecting lung cancer among the four TL models – AlexNet, VGG16, ResNet50 and Inception-v3;
- An increase in model complexity generally can improve the performance in lung cancer classification.

In the future, more kinds of TL models such as SqueezeNet, DenseNet and other variants of VGG, ResNet and Inception will be employed in lung cancer detection. And also, their performance will be evaluated and compared to find the most effective model.

References

[1] Ferlay J et al. 2021 Cancer statistics for the year 2020: An overview Int. J. Cancer 149 778-789.

[2] Zheng Q W et al. 2022 PPDPF promotes lung adenocarcinoma progression via inhibiting apoptosis and NK cell-mediated cytotoxicity through STAT3 Oncogene 1-13.

[3] Sung H et al. 2021 Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries CA Cancer J. Clin. 71 209-249.

[4] Li C et al. 2018 Adenosquamous carcinoma of the lung Onco Targets Ther. 11, 4829-4835.

[5] Fass L 2008 Imaging and cancer: a review Mol. Oncol. 2 115-152.

[6] Talukder A et al. 2022 Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning Expert. Syst. Appl. 205.

[7] Azevedo V et al. 2022 Quantum transfer learning for breast cancer detection Quant. Mach. Intell. 4 5.

[8] Borkowski A A et al. 2019 Lung and Colon Cancer Histopathological Image Dataset (LC25000) arXiv: 1912. 12142v1.

[9] Krizhevsky A et al. 2017 ImageNet classification with deep convolutional neural networks Commun. ACM. 60 84-90.

[10] Simonyan K et al. 2014 Very deep convolutional networks for large-scale image recognition CoRR abs/ 1409.1556.

[11] He K M et al. 2015 Deep residual learning for image recognition CoRR abs/1512.03385.