Biomedical Signal Processing and Control 55 (2020) 101566

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Cervical precancerous lesions classification using pre-trained densely connected convolutional networks with colposcopy images

Tao Zhang[a,b], Yan-min Luo[a,b,*], Ping Li[c], Pei-zhong Liu[d], Yong-zhao Du[d], Pengming Sun[e], BinHua Dong[e], Huifeng Xue[e]

[a] College of Computer Science and Technology, Huaqiao University, Xiamen 361021, PR China
[b] Key Laboratory for Computer Vision and Pattern Recognition of Xiamen City, Huaqiao University, Xiamen 361021, PR China
[c] Department of Gynecology and Obstetrics, the First Hospital of Quanzhou, Fujian Medical University, Quanzhou 362000, Fujian, PR China
[d] College of Engineering, Huaqiao University, Quanzhou 362021, PR China
[e] Laboratory of Gynecologic Oncology, Fujian Provincial Maternity and Children's Hospital, affiliated hospital of Fujian Medical University, Fuzhou 350001, PR China

ARTICLE INFO

Article history:
Received 31 December 2018
Received in revised form 17 April 2019
Accepted 10 May 2019
Available online 31 August 2019

Keywords:
Colposcopy
Cervical precancerous
Computer-aided diagnosis
Densely connected convolutional networks
Transfer learning

ABSTRACT

Colposcopy is currently a common medical technique for preventing cervical cancer. However, with the increase of the workload, screening by artificial vision has the problems of misdiagnosis and low diagnostic efficiency. Based on transfer learning, pre-trained densely connected convolutional networks are used to propose a computer-aided-diagnosis (CAD) method for automatic classification of cervical precancerous. The proposed method is applied to determine CIN2 or higher-level lesions in cervical images. In the present work, image data are initially prepossessing with ROI extraction and data augmentation. Then, parameters of all layers are fine-tuning with pre-trained DenseNet convolutional neural networks from two datasets (ImageNet and Kaggle). The impact of different training strategies on the model performance with limited training data is analyzed, including random initialization (RI) training from scratch, fine-tuning (FT) pre-trained model, different size of training data and K-fold cross-validation. Experimental results show that our method (FT) achieves an accuracy of 73.08% (AUC ≈ 0.75) in 600 test images. Compared with previous related work and clinicians, the performance of our approach can effective diagnosis CIN2+ and comparable with a senior physician, which proves the feasibility and promising of the proposed computer-aided diagnostic method.

1. Introduction

According to the latest World Cancer Survey statistics, cervical cancer is one of the most common female cancers worldwide and ranks fourth for both incidence (6.6%) and mortality (7.5%) [1]. Cervical cancer is mainly caused by the deterioration of precancerous lesions. It can be prevented and treated by the detection of cervical intraepithelial neoplasia (CIN). According to the World Health Organization (WHO) classification of cervical lesions, precancerous lesions cervical intraepithelial neoplasia is mainly divided into CIN1, CIN2, CIN3 [2]. CIN1 indicates mild atypical hyperplasia, which can generally be treated according to the cervical inflammation. CIN2 and CIN3 (including carcinoma in situ) are moderate and severe lesions, respectively that require timely treatments to prevent cervical cancer. In the clinical diagnosis of the CIN level, one of the crucial goals is to distinguish between normal/CIN1 and CIN2/CIN3, which is referred to herein as the diagnosis of CIN2+ [10]. Fig. 1 shows cervical data samples, we divide the data into two categories according to different lesion levels. Our goal is to identify lesions above CIN2.

Current early screening methods for cervical precancerous lesions mainly include Pap smear, HPV test, and colposcopy. Among them, Pap smear is popular in developed countries and has been effective reduce the mortality caused by cervical cancer. However, it has the problem of high cost and low sensitivity, as well as the HPV test [3]. On the other hand, colposcopy is widely used in developing countries, because of its simple operation and low cost. It is expected to solve the shortcomings of Pap smear and the HPV test. Colposcopy is to visualize the cervical region under a high light source for visual inspection. It is usually necessary to observe the cervical epidermal changes under the action of acetic acid to diag-

* Corresponding author at: College of Computer Science and Technology, Huaqiao University, Xiamen 361021, PR China.
E-mail address: lym@hqu.edu.cn (Y.-m. Luo).

https://doi.org/10.1016/j.bspc.2019.101566
1746-8094/

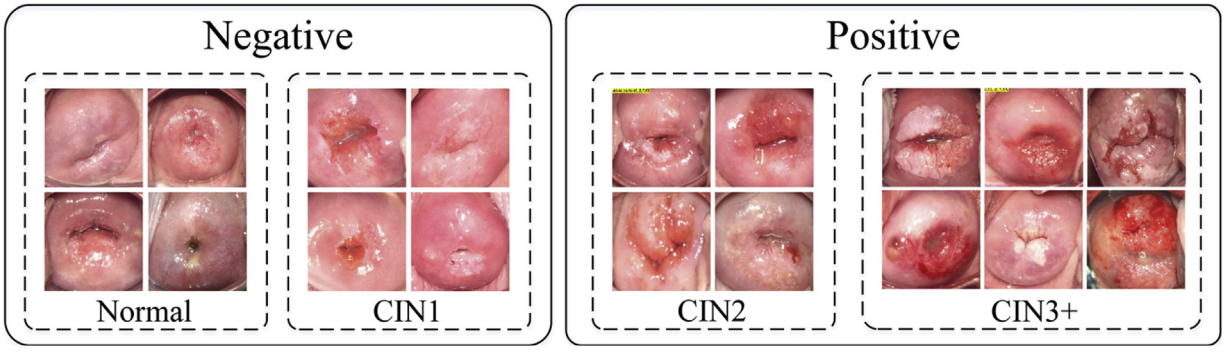

Fig. 1. Data samples. Normal and CIN1 are classified as Negative samples. CIN2 and CIN3+ (including cancer) are classified as Positive samples.

nose the grade of the lesion. However, due to limitations of the hand-craft feature selection, colposcopy has problems that highly depend on the diagnostic experience of clinical experts. Therefore, colposcopy may have less efficiency in the cancer diagnosis. Moreover, as the workload increases, it may simply lead to missed diagnosis and misdiagnosis.

Extensive computer-aided diagnosis (CAD) methods [4–8] do relevant diagnostic by processing visual information of colposcopy images. Previous works mainly focused on the shallow lesion features of cervical images such as texture and color. For example, Ji et al. [5] analyzed the vascular texture patterns of different stages of cervical lesions and introduced a generalized texture analysis technique. Experiment study with real images demonstrated the feasibility and promising of the proposed approach. Li et al. [9] proposed an automatic detection method that detects abnormal vascular characteristics of mosaic and punctuation in cervical lesion CIN based on morphological, which provided a theoretical basis for the development of the CAD system for cervical cancer screening and diagnosis. Park et al. [8] used the conditional random field model to solve problem of previous related work neglecting the spatial relationships between diagnostic features. Subsequently, data-driven methods based on machine learning increasingly being applied to colposcopy image analysis. Kim et al. [10] used thousands of colposcopy cervical images and extracted the color and texture features as a feature training set. Then, they trained the classic Support Vector Machine (SVM) [11] classifier for lesions classification. On this basis, Song et al. [12] used a multimodal aggregation scheme for cervical precancerous classification by combining multimodal data such as clinical diagnostic results (Pap, HPV, pH) and image hand-craft features. Xu et al. [13,14] used three types of complementary image features for training, including Pyramid histogram in L*A*B* color space (PLAB), Pyramid Histogram of Oriented Gradients (PHOG), and Pyramid histogram of Local Binary Patterns (PLBP). The optimal Random forest [15] classification results have certain advantages compared with the Pap test and the HPV test. These methods based on conventional image processing and machine learning have achieved great results in the application of colposcopy images. However, the conventional shallow feature learning mechanism and hand-craft feature extraction are highly depend on the characteristics of manual selection, it cannot meet the modeling requirements of complex functions in practical applications [16].

Deep learning [17] is widely used in the medical image processing and made great achievements, including classification, detection, segmentation and registration. Most of the processed image data are medical images such as MRI, CT and ultrasound images [18]. The multi-layer neural network perception mechanism of deep learning can learn more abstract features in images and is expected to solve the problems of traditional medical CAD systems. However, the deep learning needs to be supported by large-scale data to achieve desired results. Different from ImageNet million-level image dataset [19], medical images, especially positive samples, are often not readily available. To solve the problem of limited datasets, many studies have applied a variety of solutions. Currently, it is more popular to use the transfer learning [20–25] and data augmentation [17] to avoid the problem of over-fitting during training CNN.

The present study applies the deep learning technology to a new colposcopy cervical image dataset and proposes a cervical precancerous lesion diagnosis method based on pre-trained densely convolutional networks. Since our dataset is limited, the pre-trained DenseNet [26] model from the ImageNet dataset and Kaggle dataset are fine-tuned. The main purpose of experiments is CIN2+ diagnosis to determine whether a given colposcopy cervical image has a moderate or higher lesion.

2. Related works

Transfer learning (TL), also known as fine-tuning pre-trained networks, has been widely used in the medical image analysis with deep learning. It should be indicated that the so-called transfer refers to the transfer of neural network weight parameters, convolution kernel parameters and bias. In the same network model, weight parameters are pre-trained in one dataset are pre-loaded as initial weight values of the new dataset training. It refers to the small-scale adjustment of original weight parameters, according to the feature extract mechanism of the new task to achieve a great result in the new task. This procedure reduces the time of model training and can effectively solve the problem of limited data. Because of the limited training dataset, most of methods related to the medical image processing are based on deep learning with transfer learning. In the [22], the results of four different medical image applications confirm that the TL yields better results in comparison to the CNN training from the scratch. There are two main transfer learning methods: (1) fine-tuning a pre-trained CNN on a large, generic image dataset (such as ImageNet). (2) only fine-tuning the weights of the CNN last fully connected layer, and the parameters from other layers are frozen. Nima et al. [22] made a comparative analysis of these two fine-tuning methods and found that the overall effects of the method (1) is better.

In recent years, colposcopy images analysis has not widely used deep learning technology to screening and prevention of cervical cancer. Some work uses a CNN-based approach to classify abnormal cells from cervical epithelial microscopy images obtained with Pap smear to prevent cervical cancer [27–29]. At present, there are only a few works that apply CNN to the analysis of colposcopy images. Similarly, Xu et al. [31,32] used the pre-trained AlexNet [33] network to classify cervical dysplasia with colposcopy images, which shows better performance than that for the classical machine learning algorithm in the CIN2+ classification. Especially in the liter-

Table 1
The distribution of our datasets. Contains the total number of cases and images of the training set and test set. Normal and CIN1 are classified as negative, CIN2 and CIN3+ (include cancer) are classified as positive.

Class	Items	Training datasets		Test datasets	
		Cases	Images	Cases	Images
–	Total	1109	8239	600	600
negative	Normal	335	1806	297	297
	CIN1	280	2531	147	147
positive	CIN2	244	1872	82	82
	CIN3+	250	2030	74	74

ature [31], they proposed a multimodal deep network for the task of cervical dysplasia diagnosis achieved the state-of-the-art performance in visit level classification. Sato et al. [30] used the deep learning to classify the lesions of colposcopy images. The result was not satisfactory, but it suggests that deep learning has the potential to classify images from colposcopy. These works [25,34] like colposcopy image analysis also make good use of transfer learning technology with pre-trained DenseNet. Tan et al. [25] proposed a novel transfer learning method, called sequential fine-tuning method, to diagnosis lung diseases, including cancers and tuberculosis (TB) with the bronchoscopy image. Yuan et al. [34] added rotation invariant and image similarity constrained to the DenseNet model and made great results in polyp detection with endoscopy images. Given the feasibility of these efforts, the study intends to apply the DenseNet model to the classification of cervical images.

3. Materials and methodology

3.1. Datasets

The experimental data is 1709 cases of colposcopy cervical images from the Fujian Maternal and Child Health Hospital. Each case contains the patient's pathological diagnosis report. Data are collected from March 2017 to March 2018 and each patient file contains seven images on average. Data are classified according to the diagnostic report and data details are shown in Table 1. The training set totaled 1109 cases and the test set consisted of 600 cases. Experiments is performed on the CIN2+ diagnosis to distinguish whether the image has a CIN2 or higher lesion. Therefore, negative samples of the training data have 615 cases (4337 images) and positive samples have 494 cases (3902 images). The experiment randomly selected one ROI-processed image for each of the 600 test cases as the test image of each case, considering the workload of manual diagnosis. Finally, test data has 600 images with156 positive sample cases (156 images) and 444 negative sample cases (444 images).

Table 2
Summary of data augmentation methods used in this study.

Methods	parameter settings
Random_Lighting Change	±35 intensity
Random_Blur	(1, 10) block size of Mean filtering
Random_Crop	30% on both width and length
Random_Rotate	$(0, 270^{\circ})$

3.2. Preprocessing

Data preprocessing is an important part of experiments. Fig. 2 shows the data pre-processing flow before the neural network training data input. Firstly, data is divided into four categories through diagnostic reports, namely normal, CIN1, CIN2 and CIN3+ (including cancer). Because of professionalism limitations, experts are asked to label the region of interest (ROI) area of the cervical image. The ROI area is the main area where the lesion occurs, which is called the transformation zone (TZ) in the clinic. Therefore, an original image is first obtained label and ROI by annotations. Secondly, data augmentation methods are used to increase the volume of training data. Data augmentation can generate more training samples to enhance the robustness of the model and reduce the overfitting. Table 2 shows data augmentation methods and their parameter settings. All transformed images data are finally resized to 224 by 224 for the CNN training.

3.3. Method overview

Experiment is to diagnose whether cervical lesions occur above CIN2 (moderate) through deep neural network. Based on the transfer learning, pre-trained DensNet network from two datasets Kaggle and ImageNet are fine-tuned. Fig. 3 shows an overview of our method. Procedure (a) is the training process for DenseNet networks. Procedure (b) is the fine-tuning process for cervical images with pre-trained DenseNet. Procedure (b) contains three steps, data preprocessing, fine-tuning of models and prediction. First, pre-trained model weights are downloaded from the Internet. Second, cervical data for model training are preprocessed, including a series of methods for data augmentation. Finally, a pre-trained DenseNet network is fine-tuned. The trained DenseNet model can diagnose whether the cervical image is higher than CIN2, we called CIN2+ diagnosis.

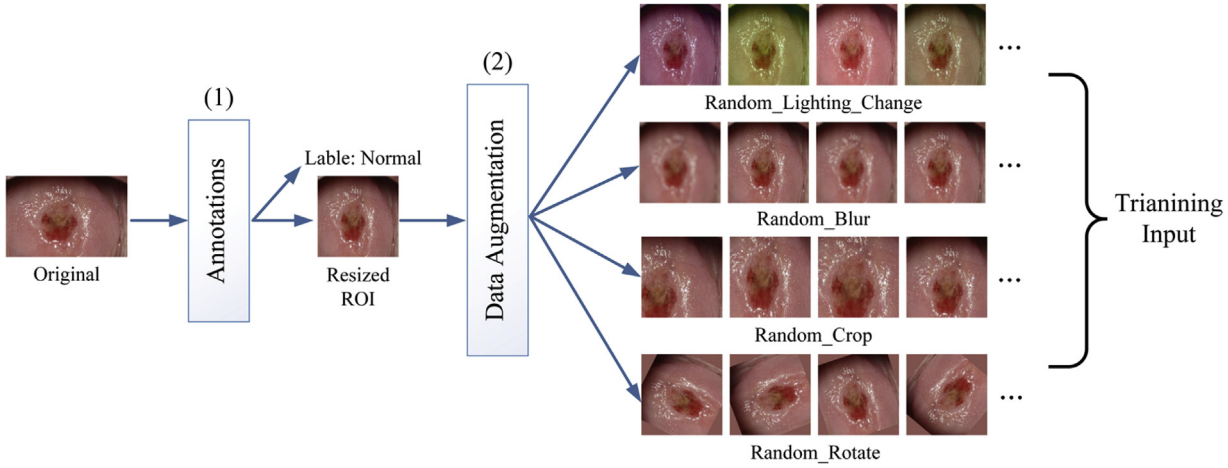

Fig. 2. Data preprocess. Mainly contains ROI extract (1) and data augmentations (2).

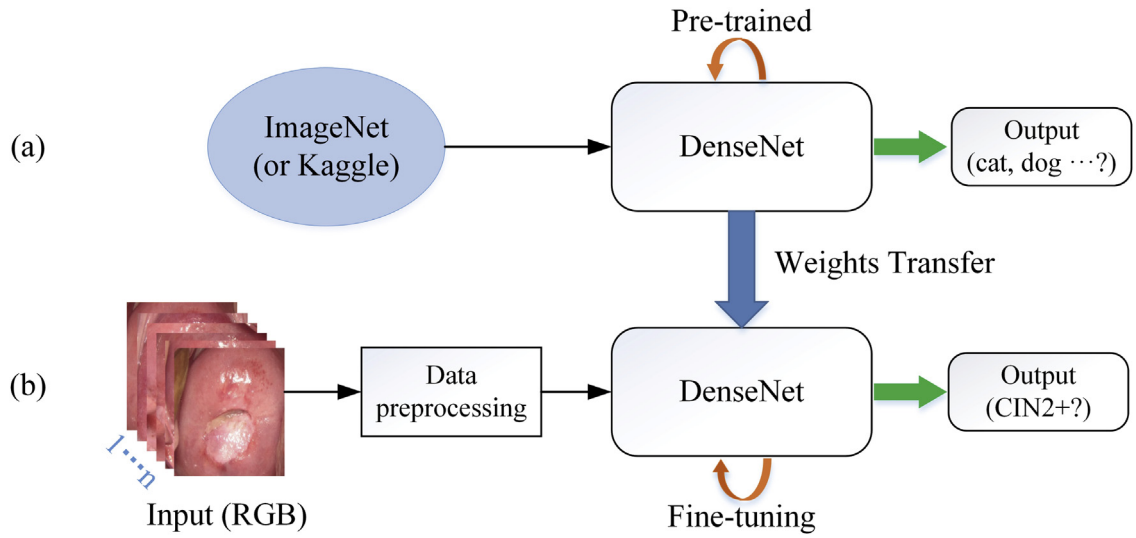

Fig. 3. The Overall flow-chart of our CNN-based framework for cervical precancerous lesions classification. (a) is the procedure of the pre-trained DenseNet model. (b) is our main process of fine-tuning the pre-trained model in our cervical images.

3.4. Pre-trained densely convolutional neural networks

In this study, the pre-trained DenseNet network structure is applied, which won the 2016 ImageNet image classification and recognition competition (ILSVRC 2016). Fig. 4 shows the DenseNet network structure is composed of four Dense blocks and three transition blocks. Based on Resnet [35] (ILSVRC 2015 champion), DenseNet uses a dense connection to improve the connection of context information, its input of each layer contains the output and input of the early layer, which enhances the transmission of information and focusing on the feature reuse. Features of images are often complex and indistinguishable, especially in medical images. Coupled with the limited data, the training of neural networks is prone to gradient disappearance then leads to overfitting. DenseNet mainly benefits from the design of Dense blocks. The method of feature map concatenation enhances the information connection to alleviate the problem of gradient disappearance. It performs well for the extraction of complex features.

The Dense block module has very similar module as the Resnet. However, there are significant differences between the Dense block and Resnet modules. Eq. (1) indicates that in ResNet's residual module, the s̈kip connectionm̈akes the simple summation of the x_{l-1} and nonlinear transformation of x_{l-1}. Therefore, the network can learn the residual mapping of the input, which can effectively alleviate the gradient disappearing problem. On the other hand, Eq. (2) shows that in the Dense block module, the connection of the feature map is not a simple summation, but a concatenation. The number of convolution maps at the i^{th} layer is $K*(i-1)+K_0$, where K is the number of feature maps convolved for each layer, indicating the growth rate. Fig. 4 shows that, in the Dense block, the input of any layer includes feature maps of all early layers. It makes the L layer network have $\frac{L(L+1)}{2}$ connections.

$$x_l = F_l(x_{l-1}) + x_{l-1} \tag{1}$$

$$x_l = F_l([x_0, x_1, \ldots\ldots, x_{l-1}]) \tag{2}$$

Where $F(\cdot)$ and x_l denote a nonlinear transformation and the output of the layer l, respectively.

Two network models used for the classification of cervical precancerous lesions are Densenet121 and Densenet169. The difference lies in the number of convolution layers in the dense block. Table 3 shows the specific parameters of the network structure.

3.5. Transfer learning

Transfer learning (TL) is one of the machine learning methods. Transfer learning also known as fine-tuning pre-trained networks and is widely used in deep learning. Especially in computer vision (CV) and natural language processing (NLP) applications, training a neural network model requires large-scale training data, calculation and the time resource. The technology span in deep learning is also large, random initialization of the training network from scratch within limited data may lead to local optimization [22]. Transfer learning applies a pre-trained model to a new task and serves as a starting point for the new task. This ideal is handled by parameter transfer approaches and each model of the related task should share some structure (parameters or hyperparameters) [36]. Therefore, training a new CNN model from related tasks should be faster, with better solutions or with less amount of labeled data.

Refer to a similar transfer learning ideal like [37], experiments fine-tuning two pre-trained models. One is the commonly used large-scale data ImageNet for CNN training. The other is the pre-trained network from Kaggle competition, which training data is closer to experiment cervical data. Among them, the network parameters of the Kaggle model is fine-tuning from the pre-trained model of ImageNet. This study uses two commonly fine-tuning ways, drawing on the analysis scheme of the [22] and analyzing the impact of different training data on the fine-tuning results. It also compares the results of the fine-tuning and random initialization training methods.

3.6. Implementations

Neural network training is performed in accordance with the Keras [38] deep learning API with TensorFlow [39] backend. Moreover, the model training is running on the NVIDIA GTX 1060ti graphics card. After many training and parameter adjustments, the optimization method uses the Adam algorithm [40] (where " β_1 ", " β_2 ", ε and learning rate (LR) are set to 0.9, 0.999, 1×10^{-8} and 1×10^{-5}, respectively).

In order to increase the robustness of the model and obtain a stable model, cross-validation is used for the model training. The K-fold cross-validation method is used in the experiment and K values are set to 5 and 2, respectively. When K = 5, the training data is divided into five parts, four parts of each training are taken as the

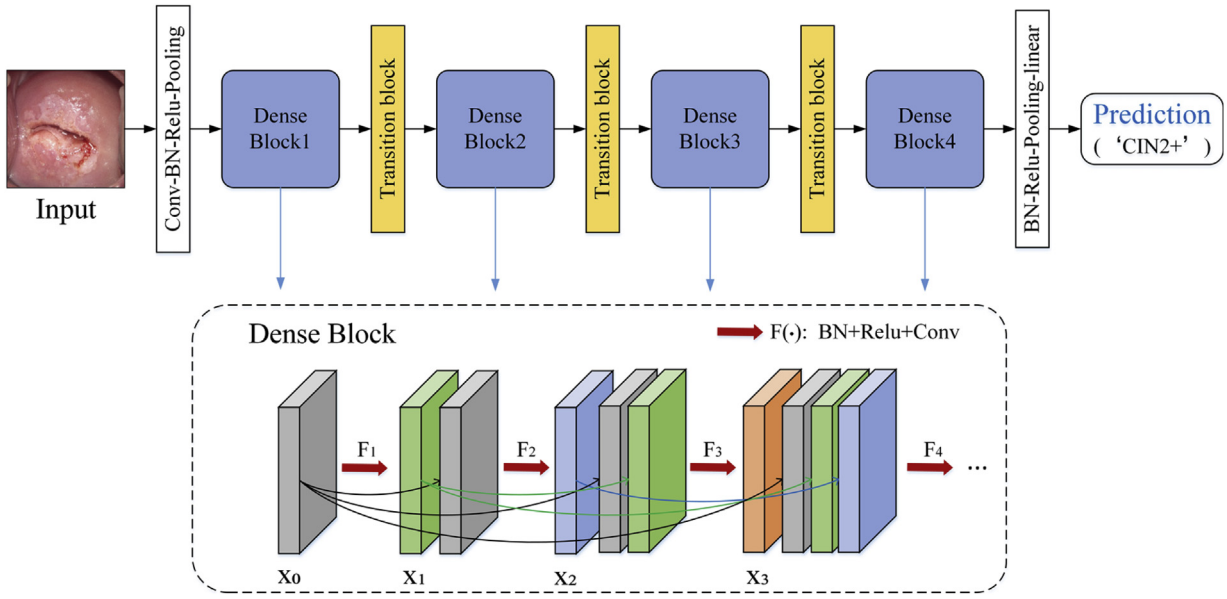

Fig. 4. Schematic configuration of the DenseNet model for cervical classification. Four Dense Blocks are connected to three transition blocks.

Table 3
Network architectures of two DenseNets.

Block	Output size	Layers	Densenet121	Densenet169
Convolution	112*112	7*7 conv	1	1
Pooling	56*56	3*3 max pool	1	1
Dense Block1	56*56	1×1 conv 3×3 conv	6	6
Transition Layer1	28*28	1*1 conv, 2*2 Avg pool	1	1
Dense Block2	28*28	1×1 conv 3×3 conv	12	12
Transition Layer2	14*14	1*1 conv, 2*2 Avg pool	1	1
Dense Block3	14*14	1×1 conv 3×3 conv	24	32
Transition Layer3	7*7	1*1 conv, 2*2 Avg pool	1	1
Dense Block4	7*7	1×1 conv 3×3 conv	16	32
FC Layer	1*1	7*7 avg pool, sigmoid	1	1

training set, and the remaining part is used as the validation set to calculate the accuracy of the model. In other words, the training is repeated five times. Finally, the average of prediction results of each model in the test set is calculated to evaluate the performance of the trained model. When K = 2, the training data is divided into two parts in the same way, which is used as the training set and the validation set respectively, repeatedly trained twice. Obviously, when K is different, the data size of the training set and the validation set is different. Experiment implements two cross-validation methods and analyze which performance is better. In addition, referring to the experimental content of [22], the impact of the dataset size on the fine-tuning performance was also analyzed. The training data are set 100%, 50%, and 10% of all training data separately.

4. Results

4.1. Evaluations of model training and classification

In the training of the model, experiments need to solve a binary classification problem, which is to distinguish whether the cervical image has lesions of CIN2+. The labels for the training data are set to Öänd Ï. ÖÏndicates the negative sample to represent the normal/CIN1 level of the cervical images, and ÏÏndicates the positive sample represent the CIN2/CIN3+ level of the cervical image. The sigmoid activation function is selected at the final output layer of the network. The Logistic Regression (LR) model in machine learning is based on this function and is often used for classification and regression. The activation function can be represented by Eq. (3).

$$S(t) = \frac{1}{1 + e^{-t}} \tag{3}$$

where S ∈ (0, 1), when t → +∞, $S(t) \approx 1$, t → −∞, $S(t) \approx 0$. The corresponding loss function sets a binary-crossentropy loss function, which is generally used in binary classification problem. Eq. (4) indicated that the calculation details of loss value.

$$loss = -\sum_{i=1}^{n} \hat{y}_i \log y_i + (1 - \hat{y}_i) \log(1 - y_i) \tag{4}$$

Where n represents the number of test set images, $\hat{y}_i$ represents the correct value of the desired output, y_i represents the output of the neural network, and loss approaches 0 when y_i is closer to $\hat{y}_i$.

Table 4
The definition of evaluation indicators.

evaluations	definition
ACC	$TP + TN/(TP + FN + FP + TN)$
SEN	$TP/(TP + FN)$
SPEC	$TN/(TN + FP)$
NPV	$TP/(TP + FP)$
PPV	$TN/(TN + FN)$

A final evaluation of the classification with test set is made, based on each training model. It should be indicated that the test set is a patient-based image test. Evaluation indicators include accuracy (ACC), sensitivity (SEN), specificity (SPEC), positive predicted value (PPV) and negative predictive value (NPV) [41]. Table 4 illustrates performance evaluations, where TP, TN, FP and FN are the number of true positives, true negatives, false positives, and false negatives respectively in test results.

4.2. Two pre-trained model for fine-tuning

Pre-trained networks with different dataset may have a certain impact on the results of fine-tuning. Most work uses the pre-trained network model with the big dataset (such as ImageNet) or uses the large-scale data that close to the target training data. This study fine-tuned the two pre-trained network models from ImageNet and Kaggle (close to the experiment's cervical data) and analyzed their differences and connections.

4.2.1. ImageNet pre-trained model

Fig. 5 shows the loss curve of fine-tuning on the ImageNet pre-trained network model for Densenet121 and Densenet169. Applying the Early-Stopping training strategy in Keras, the parameter of patience is set to 10. This strategy is often used in cross-validation to improve the efficiency of the CNN training. The convergence of the network determined according to the change of the loss value on the validation set. DenseNet121 converges at a relatively faster speed, the val_loss reached 0.2316 at the lowest and stops training after 10 epochs. DenseNet169 converges slowly and val_loss reached 0.2304. Two network training processes are similar. The train_loss of DenseNet169 is lower than DenseNet121, but the val_loss curve is unstable. In contrast, DenseNet121 is more stable and smoother.

4.2.2. Kaggle pre-trained model

Recently, Kaggle conducted a cervical transformation zone (TZ) classification competition (https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening, 2017). The image data contains 8215 colposcopy images with three types. Considering the related image classification tasks, this study applied the same method to fine-tuning the pre-trained network model from the Kaggle dataset. Fig. 6 shows the loss curve of CNN training. The loss value during training is like the fine-tuning of the ImageNet pre-trained model.

The experiment used the trained model to classify and test 600 test images and evaluate these models through the test results. Some indicators of the medical evaluation (ACC and SEN etc.) are used to assess the final diagnostic performance of the trained model. Table 5 shows the classification results of DenseNet121, DenseNet169 and SVM models. Each result is the average of the 5-fold cross-validation. Two DenseNet models are fine-tuned with ImageNet and Kaggle pre-trained dataset respectively. In addition, a traditional machine learning based image classification method with the SVM classifier using PHOG, PLBP and PLAB handcraft features is employed for cervical images of the present study in accordance with Xu's article [31]. Comparing these models indicates that the results of the fine-tuned DeneNet121 with ImageNet are generally better than those of other CNN models from ACC, SPEC, and PPV. However, SEN and NPV are lower relatively.

The receiver operating characteristic (ROC) and the area under the ROC curve (AUC) are also used for evaluating the binary class classification performance. Fig. 7 shows the ROC curves for the different CNN model classification results. A prediction results that closest to the average performance of the cross-validation model are used for drawing the ROC curve. According to the ROC and AUC distribution, the overall performance of these four fine-tuned models is equivalent, in which the fine-tuned DenseNet121 with Kaggle reached highest AUC = 0.763. In fact, the Kaggle pre-trained model used in this study is pre-fine-tuned with their own cervical images using pre-trained ImageNet model. In other words, the Kaggle pre-trained model used in the experiment is fine-tuned from ImageNet in the same way as in this study. Although the Kaggle model is obtained by training cervical images, the pre-trained weights all adopt ImageNet large-scale datasets from the root. Table 5 clearly shows that the results of the SVM (PHOG + PLBP + PLVB) are not satisfactory compared with four deep CNN methods. It indicates that transfer learning with the DenseNet model based on deep learning is better than the conventional machine learning method.

4.3. Comparative experiment

Referring to some related work, we also conducted comparative experiments of different training strategies to analysis

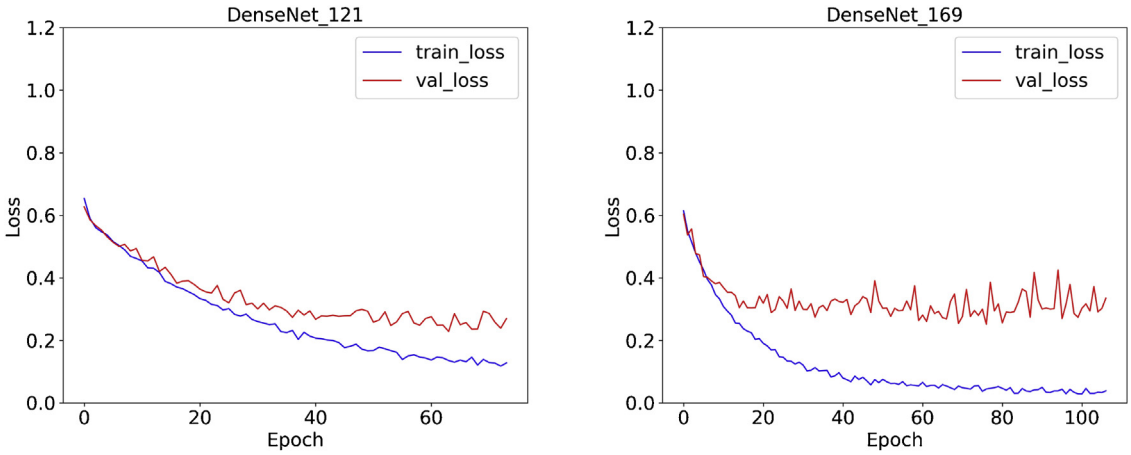

Fig. 5. Training and validation loss curve for DenseNet121 and DenseNet169 models that fine-tuning all network layers with ImageNet pre-trained models.

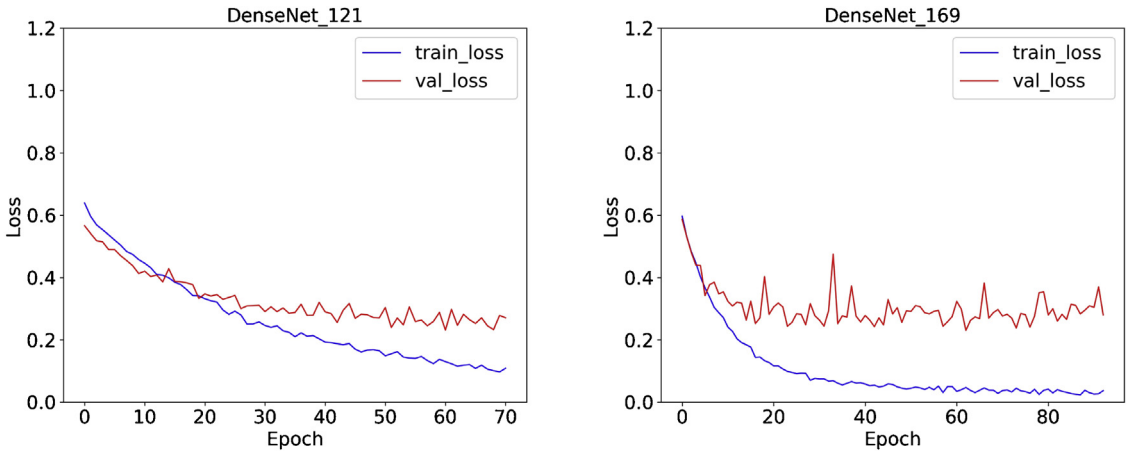

Fig. 6. Training and validation loss curve for DenseNet121 and DenseNet169 models that fine-tuning all network layers with Kaggle pre-trained models.

Table 5
The test results of two fine-tuned network models and the method from [31]. The result is the average of the 5-fold cross-validation.

Methods	Pre-trained dataset	ACC	SEN	SPEC	PPV	NPV
DenseNet_121	ImageNet	0.7308	0.5756	0.7855	0.4879	0.8408
DenseNet_169		0.6901	0.6679	0.6979	0.4393	0.8568
DenseNet_121	Kaggle	0.7242	0.5986	0.7683	0.4839	0.8452
DenseNet_169		0.6979	0.6500	0.7148	0.4484	0.8531
SVM (PHOG + PLBP + PLVB) [31]	–	0.6327	0.3846	0.7185	0.3243	0.7687

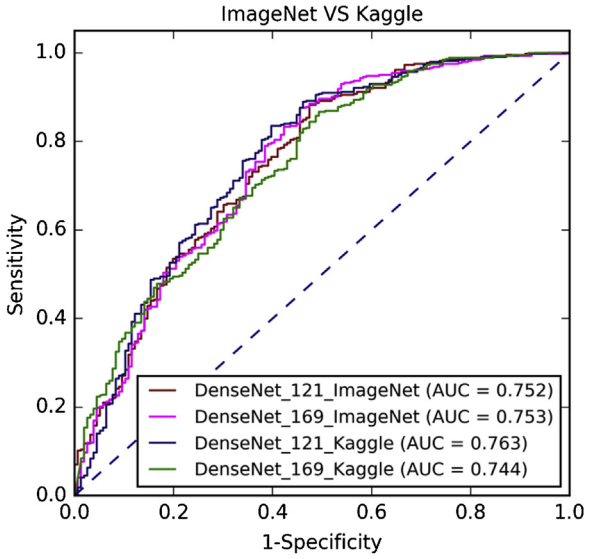

Fig. 7. ROC curves for different model results. The value of AUC represents the classification performance of the classifier.

cervical image classification. It includes random initialization VS fine-tuning, different K-fold cross-validation (K = 2 VS K = 5), and training with different sizes of the dataset.

(a) Random Initialization VS Fine-tuning: To verify the feasibility of the transfer learning, experiments compare two training ways that random initialization training from scratch and Fine-tuning pre-trained models respectively. Fig. 8(a) shows the receiver operating characteristic curve (ROC) for the two training ways with DenseNets. The area under the ROC curve is 0.752 (DenseNet_121 with FT), 0.640 (DenseNet_121 with RI), 0.753 (DenseNet_169 with FT) and 0.695 (DenseNet_169 with RI), respectively.

(b) K-fold cross validation (K = 2 VS K = 5): Fig. 8(b) shows the receiver operating characteristic curve (ROC) for the fine-tuning with DenseNets using the cross-validation method of K = 2 and K = 5 respectively. The area under the ROC curve is 0.752 (DenseNet121 with K = 5), 0.745 (DenseNet121 with K = 2), 0.752 (DenseNet169 with K = 5) and 0.727(DenseNet169 with K = 2), respectively.

(c) Different size of training data. The experiment further analyzed how the size of training samples influences the performance of the CNN classification. Experiments divided the training data into 100%, 50% and 10% of the total data to CNN training. This part training was all DenseNet121 fine-tuning with ImageNet. Fig. 8(c) shows the receiver operating characteristic curve (ROC) for each trained model of different training samples. The area under the ROC curve is 0.752 (100% training data), 0.735 (50% training data) and 0.694 (10% training data).

Table 6 shows the average test results for training in 600 test sets in different trained models. From the above results and ROC curve analysis, the performance of the network model obtained by fine-tuning the two different dataset of ImageNet and Kaggle is not much different. The performance of the two network structures DenseNet_121 and DenseNet_169 are similar in our data (AUC ≈ 0.75). From the comparison of the trained methods FT and RI, the FT mode (AUC ≈ 0.75) are much higher than RI (AUC ≈ 0.65). After this comparative experimental analysis, we have reached the same conclusion as the literature [21], in the medical image training of small dataset, it is still recommended to use fine-tuning to CNN training. Compared with the fine-tuning of the two K-old cross-validation methods, the overall performance comparison is also similar, this indicate that the value of K has little effect on the performance of the model. Compared with the training of different data volumes, there is a small gap between AUC and other diagnostic indicators. According to the reduction of training data size, all indicators are reduced. These results show that the fine-tuning still tend to be more training data for CNN training.

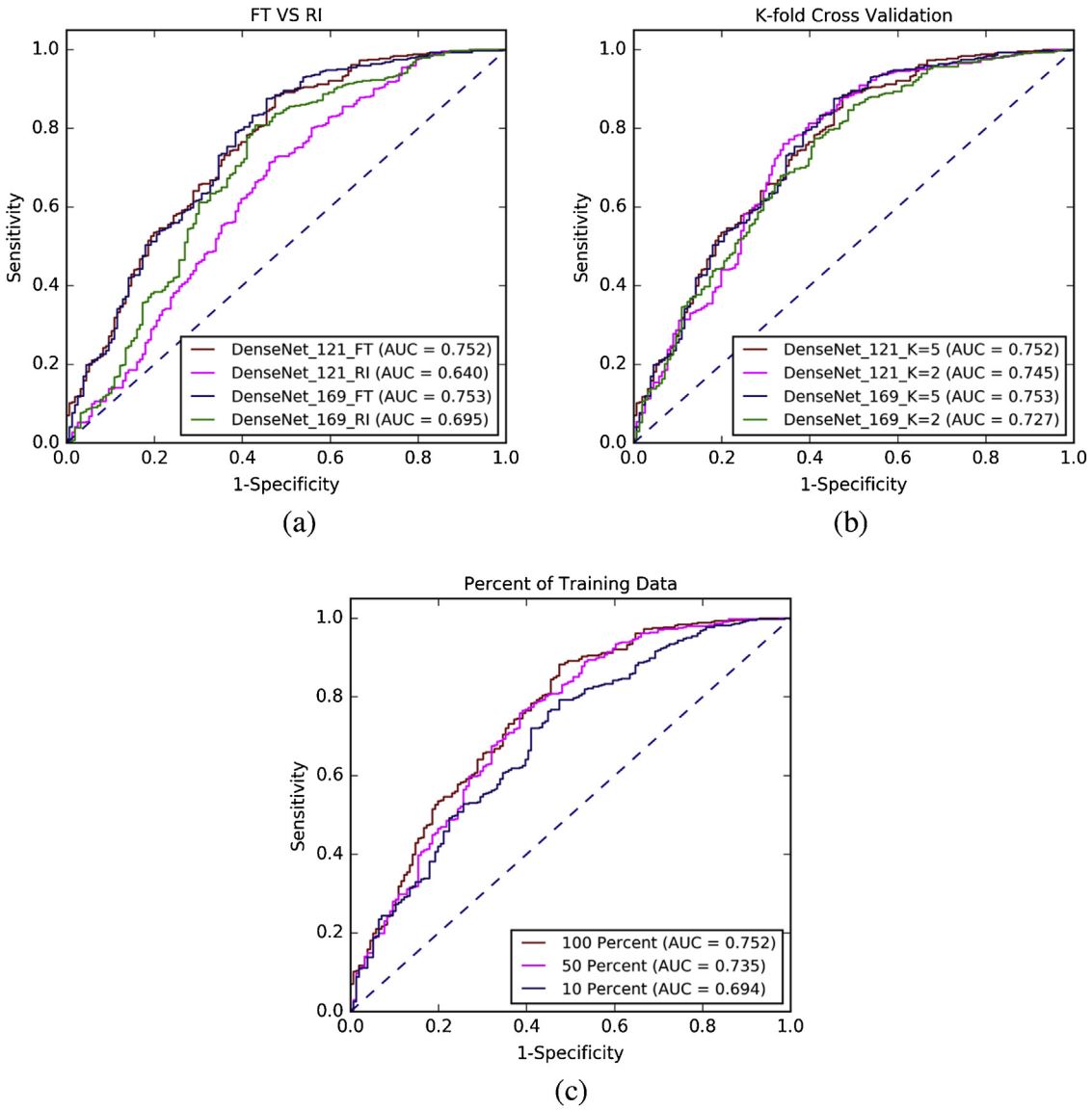

Fig. 8. (a): FT: fine-tuning. RI: Random Initialization. Two DenseNet networks test results with two training ways. (b): Comparison of K-fold cross validation (K = 5, 2). (c): Different size of training data for DenseNet_121 training.

Table 6
Comparative experiment results. Contains the results of training ways FT VS RI, K-fold cross validation (K = 5, 2) and different size of training data (100%, 50%, 10%).

Comparative experiments	Models	ACC	SEN	SPEC	PPV	NPV
(a)	DenseNet121_FT(K = 5, 100% data)	**0.7308**	0.5756	**0.7855**	**0.4879**	0.8408
	DenseNet121_RI(K = 5, 100% data)	0.6580	0.5833	0.6843	0.4006	0.8257
	DenseNet169_FT(K = 5, 100% data)	0.6901	**0.6679**	0.6979	0.4393	**0.8568**
	DenseNet169_RI(K = 5, 100% data)	0.6440	0.6230	0.6514	0.3919	0.8331
(b)	DenseNet121_FT(K = 2, 100% data)	0.6936	0.6442	0.7110	0.4421	0.8506
	DenseNet169_FT(K = 2, 100% data)	0.7011	0.6378	0.7235	0.4557	0.8500
(c)	DenseNet121_FT(K = 5, 50% data)	0.7071	0.6320	0.7336	0.4584	0.8512
	DenseNet121_FT(K = 5, 10% data)	0.6931	0.6525	0.7074	0.4429	0.8532

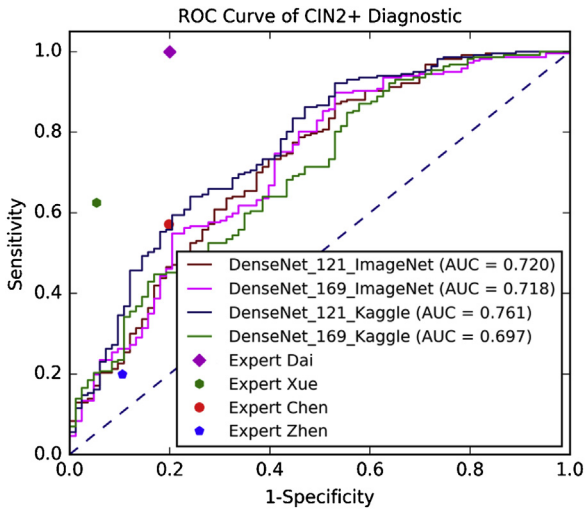

Fig. 9. Three hundred images test results of all methods. The best performance of CNN model is fine-tuned DenseNet121 with Kaggle pre-trained model (AUC = 0.761) and the ROC curve contains the performance points of physician Chen and physician Zhen.

In addition, four colposcopy physicians are invited to diagnose for 300 cervical images (79 positive images and 221 negative images) that randomly selected from the test dataset. They are expert physicians (Dai) and (Xue), senior physician (Chen) and junior physician (Zhen) respectively. Fig. 9 shows the performance comparison of fine-tuned DenseNet models and different physicians for CIN2+ diagnosis with 300 images. The performance of each physician reader plotted in the ROC graph using points. It is observed that physician Chen and physician Zhen performance points fall within the ROC curve for the fine-tuned DenseNet121 with Kaggle. Table 7 details the diagnosis results of the Densenet121_Kaggle and four physicians, where TP represents a true positive image in the diagnosis result, and TN represents a true negative image. It indicates that the sensitivity of the diagnosis varies greatly with the level of expertise. From the overall performance of the model of this study, the proposed method has achieved satisfactory results with an ACC = 0.7633, SEN = 0.4430, and SPEC = 0.8778. The performance of PPV and NPV is still very well despite the relatively low sensitivity. It means that the proposed method can diagnosis CIN2+ and comparable with a senior physician.

5. Discussion

Experiments used two model structures, the fine-tuned DenseNet121 achieving an average accuracy of 0.7308 and 0.7242 with the ImageNet and the Kaggle pre-trained models, which is better than the fine-tuned DenseNet169 achieving an average accuracy of 0.6901 and 0.6979 with two pre-trained models. However, sensitivity is an important criterion of medical diagnosis, especially in the early stages of disease screening. Clinical screening of positive samples should try to avoid missed diagnosis. The sensitivity of the two pre-trained models of the DenseNet169 is 0.6679 and 0.6500, respectively, which is higher than the 0.5756 and 0.5986 of the DenseNet121. It indicates that the model sensitivity and depth of the network seems to be positively correlated. The deeper the neural network structure, the better the fitting of positive lesion features. However, for cervical lesion images with less inter-class difference, it also leads to the decline of overall model classification performance. Experiments show that the depth of the network has a significant impact on the unilateral performance of the model. Setting the depth of the network for specific screening tasks is also worth investigating.

In addition, the comparative experiment of this study analyzed the selection of training strategy and the influence of data size. Results demonstrate that the fine-tuning is still the first choice to training CNN with limited target data. The demand for large-scale data is always a difficult problem for deep learning, especially for medical data. Obviously, the increase in data size can improve the overall classification performance of the model. Data augmentation is a common technique for deep learning to generate more training samples using limited data. However, the original data determines the diversity of features, and the new data generated by the algorithm cannot effectively represent the distribution of features in real life. Therefore, how to design data augmentation methods for specific problems is also a research point of deep learning in medical image processing.

Machine-learning based approaches using colposcopy images to cervical cancer screening yield promising early results. Table 8 list the previous studies mentioned in the section of introduction and related works. It contains the main method, datasets and performance of each paper. These results indicate that the multimodal CNN scheme [30] has achieved state-of-the-art performance in their data analysis. However, non-image data such as Pap and HPV are not readily available in low-resource regions where cervical cancer incidence and mortality is highest, so the real-life application of multimodal aggregation is uncertain. The image only method performs well. Conventional machine-learning methods on RF and AdaBoost [13,14] achieved a high specificity (90%, 90% respectively) but low sensitivity (51%, 42% respectively). A deep learning method [30,31] still yields a promising result. Among them, Xu et al. fine-tuning a CNN and achieving an ACC, SEN, and SPEC of 65.0%, 40.0%, 90.0% respectively. Even though the dataset is different, our study achieved a higher ACC and sensitivity (73.0% and 57.5%) using fine-tuned DenseNet based on deep learning methods. It seems to indicate that the DenseNet model based on deep learning is more suitable for the classification of cervical images.

Unlike the CNN method, manual diagnosis focuses on shallow hand-craft features such as color and texture information and the correctness of clinician's analysis depend on their clinical experience. Although with the limited clinical group, comparison with clinicians indicates that the research method is promising and comparable to that of senior or junior physicians. About 14 positive cases and 55 negative cases were the consensuses achieved between DenseNet121_Kaggle with a senior physician in the same 100 test results. Fig. 10 shows some true positive, true negative, false positive and false negative examples according to the diagnosis given by the manual reading of a senior physician and

Table 7
Experiment results of Densenet121_Kaggle compared with physician's diagnosis results.

Physicians	TP	TN	ACC	SEN	SPEC	PPV	NPV
DenseNet121_Kaggle	35	194	0.7633	0.4430	0.8778	0.5645	0.8151
Expert Dai	79	177	0.8526	1.000	0.8045	0.6475	1.000
Expert Xue	49	209	0.8600	0.6202	0.9457	0.8032	0.8744
Senior Chen	45	177	0.7400	0.5696	0.8009	0.5056	0.8388
Junior Zhen	16	197	0.7100	0.2025	0.8914	0.400	0.7576

Table 8
Published studies of machine-learning based cervical cancer screening.

Paper	Method	Data	Performance
Kim, et al., 2013 [10]	SVM	1000 negative (Normal or CIN1) and 1000 positive (CIN2+) images	ACC: 75.5% SEN:75.0% SPEC:76.0%
Song, et al., 2015 [12]	Multimodal aggregation	140 negative (Normal or CIN1) and 140 positive (CIN2+) cases	ACC: 89.0% SEN: 83.2% SPEC: 94.7%
Xu, et al., 2015 [13]	RI	767 negative (Normal or CIN1) and 345 positive (CIN2+) patient visits.	SEN: 51.0% SPEC: 90.0%
Xu, et al., 2015 [14]	AdaBoost	767 negative (Normal or CIN1) and 345 positive (CIN2+) patient visits.	SEN: 42.0% SPEC: 90.0%
Xu, et al., 2016 [30]	Multimodal aggregation with CNN (FT)	345 negative (Normal or CIN1) and 345 positive (CIN2+) patient visits.	ACC: 88.9% SEN: 87.8% SPEC: 90.0%
Xu, et al., 2017 [31]	CNN (FT)	767 negative (Normal or CIN1) and 345 positive (CIN2+) patient visits.	ACC: 65% SEN: 40.0% SPEC: 90.0%
Sato, et al., 2018 [33]	CNN (RI)	Training: 485 images. Validation: 25 images.	Accuracy of the validation was >50.0%

CNN: Convolutional Neural Network; FT: Fine-tuning; RI: Random Initialization.

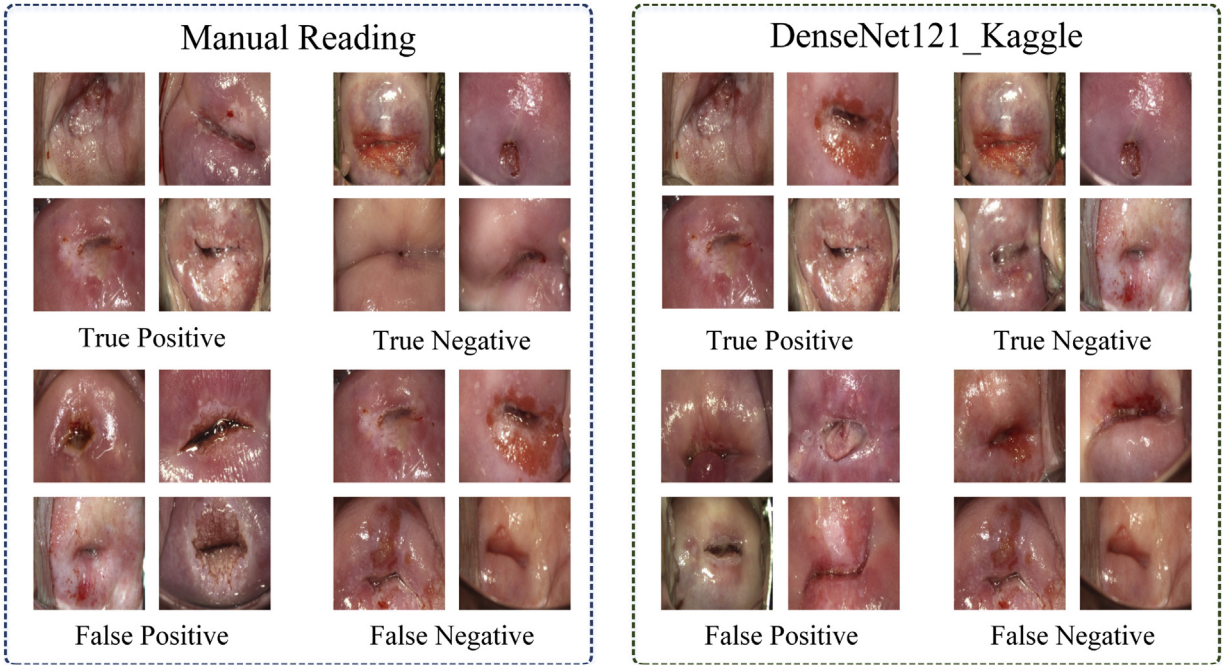

Fig. 10. Diagnosis examples as diagnosed by the manual reading (left) and DenseNet121_Kaggle (right).

DenseNet121_Kaggle. As one can see, the difference between true positive and true negative samples is obvious in the case of consensus. Both of manual reading and DenseNet121_Kaggle, some of the false positive examples are difficult to distinguish from true positive examples. In the diagnosis of false negatives, doctors have fewer false positives. CNN seems to be more focused on cervical epithelial changes and ignores morphological features, its false negative epidermal features are similar to true negatives. According to experts, there is no significant change in the cervical epithelium of some highly pathological lesions. For deep learning, if the image features of limited samples are too sparse, which leads to under-fitting and misdiagnosis. Increasing specific data samples with improved data augmentation methods may improve the classification performance of CNN.

6. Conclusion

This present study proposes a method for classification of precancerous lesions using pre-trained densely connected convolutional networks with colposcopy images. Due to the limited data, the training of CNN is benefited by transfer learning and obtained a satisfactory result. Two pre-trained DenseNet models downloaded from ImageNet and Kaggle fine-tuned with the target cervical images. The final performance of two fine-tuned models is almost the same. The main reason is that Kaggle's colposcopy image dataset is limited and its pre-trained model is also fine-tuning by ImageNet pre-trained model, not training from scratch. In the comparative experiments, this study analyzed the effects of different training strategies. Deep learning in the colposcopy image classification task has still preferred the way of fine-tuning a CNN and uses more training data. In addition, compared with the conventional SVM algorithm, experiments show that the performance of the DenseNet-based algorithm is better than the SVM classifier that training using hand-craft features. Importantly, this method can efficiently test and diagnose 600 images in less than 1 min. Compared with previous related work and clinicians, the proposed method is effective and has promising prospects. In the future, we will increase our data volume and work on the improvement of

the data augmentation and CNN algorithm to create a diagnostic framework for cervical precancerous lesions suitable for new data.

Ethics statement

All experiments were performed in compliance with the ethical standards set by our institutional board. The authors had access to information that could identify individual participants during or after image analysis.

Declaration of Competing Interest

No conflict of interest exits in the submission of this manuscript.

Acknowledgements

This work was supported by Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (No. ZQN-PY518), and the grants from National Natural Science Foundation of China (Grant No. 61605048), in part by Natural Science Foundation of Fujian Province, China under grant 2016J01300, in part by Fujian Provincial Big Data Research Institute of Intelligent Manufacturing, in part by the Quanzhou City Scientific & Technological Program of China (No. 2018C113R and No. 2018G024), and in part by the Subsidized Project for Postgraduates' Innovative Fund in Scientific Research of Huaqiao University under Grant 17014084014.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (2018) 394–424, http://dx.doi.org/10.3322/caac.21492.
[2] WHO: Human Papillomavirus and Related Cancers in the World, Summary Report, ICO Information Centre on HPV and Cancer, 2014.
[3] R. Sankaranarayanan, L. Gaffikin, M. Jacob, J. Sellors, S. Robles, A critical assessment of screening methods for cervical neoplasia, IJGO 89 (2005) 4–12, http://dx.doi.org/10.1016/j.ijgo.2005.01.009.
[4] Q. Ji, J. Engel, E. Craine, Classifying cervix tissue patterns with texture analysis, Pattern Recognit. 33 (2000) 1561–1573, http://dx.doi.org/10.1016/s0031-3203(99)00123-5.
[5] H. Lange, D.G. Ferris, Computer-aided-diagnosis (CAD) for colposcopy, in: Medical Imaging 2005: Image Processing, International Society for Optics and Photonics, 2005, pp. 71–85, http://dx.doi.org/10.1117/12.594658.
[6] Y. Srinivasan, E. Corona, B. Nutter, S. Mitra, S. Bhattacharya, A unified model-based image analysis framework for automated detection of precancerous lesions in digitized uterine cervix images, IEEE J. Sel. Top. Signal Process. 3 (2009) 101–111, http://dx.doi.org/10.1109/jstsp.2008.2011102.
[7] S.Y. Park, D. Sargent, R. Lieberman, U. Gustafsson, Domain-specific image analysis for cervical neoplasia detection based on conditional random fields, IEEE Trans. Med. Imaging 30 (2011) 867–878, http://dx.doi.org/10.1109/tmi.2011.2106796.
[8] Simões, et al., Classification of images acquired with colposcopy using artificial neural networks, Cancer Inform. 13 (2014) 119–124, http://dx.doi.org/10.4137/CIN.S17948.
[9] W. Li, A. Poirson, Detection and characterization of abnormal vascular patterns in automated cervical image analysis, in: International Symposium on Visual Computing, Springer, Berlin, Heidelberg, 2006, pp. 627–636, http://dx.doi.org/10.1007/11919629_63.
[10] E. Kim, X. Huang, A data-driven approach to cervigram image analysis and classification Lecture Notes in Computational Vision and Biomechanics: Color Medical Image Analysis, 6, 2013, pp. 1–13, http://dx.doi.org/10.1007/978-94-007-5389-1_1.
[11] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intel. Syst. Technol. (TIST) 2 (2011) 27, http://dx.doi.org/10.1145/1961189.1961199.
[12] D. Song, E. Kim, X. Huang, Multi-modal entity coreference for cervical dysplasia diagnosis, IEEE Trans. Med. Imaging 34 (2015) 229–245, https://doi.org/10.1109tmi.2014.2352311.
[13] T. Xu, C. Xin, L.R. Long, S. Antani, Z. Xue, E. Kim, X. Huang, A new image data set and benchmark for cervical dysplasia classification evaluation, in: International Workshop on Machine Learning in Medical Imaging, Springer, Cham, 2015, pp. 26–35, http://dx.doi.org/10.1007/978-3-319-24888-2_4.
[14] T. Xu, E. Kim, X. Huang, Adjustable adaboost classifier and pyramid features for image-based cervical cancer diagnosis, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE, 2015, pp. 281–285, http://dx.doi.org/10.1109/isbi.2015.7163868.
[15] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (2002) 18–22.
[16] Z.J. Sun, L. Xue, Y.M. Xu, Z. Wang, Overview of deep learning, Jisuanji Yingyong Yanjiu 29 (2012) 2806–2810, http://dx.doi.org/10.3969/j.issn.1001-3695.2012.08.002.
[17] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, http://dx.doi.org/10.1038/nature14539.
[18] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, http://dx.doi.org/10.1016/j.media.2017.07.005.
[19] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee (2009) 248–255, http://dx.doi.org/10.1109/cvprw.2009.5206848.
[20] K. Fernandes, J.S. Cardoso, J. Fernandes, Transfer learning with partial observability applied to cervical cancer screening, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer, Cham, 2017, pp. 243–250, http://dx.doi.org/10.1007/978-3-319-58838-4_27.
[21] U.K. Lopes, J.F. Valiati, Pre-trained convolutional neural networks as feature extractors for tuberculosis detection, Comput. Biol. Med. 89 (2017) 135–143, http://dx.doi.org/10.1016/j.compbiomed.2017.08.001.
[22] N. Kajbakhsh, J.Y. Shin, S. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (2016) 1299–1312, http://dx.doi.org/10.1109/tmi.2016.2535302.
[23] J.J. Näppi, T. Hironaka, D. Regge, H. Yoshida, Deep transfer learning of virtual endoluminal views for the detection of polyps in CT colonography, Medical Imaging 2016: Computer-aided Diagnosis (2016) 97852B, http://dx.doi.org/10.1117/12.2217260.
[24] R.K. Samala, H.P. Chan, L. Hadjiiski, M.A. Helvie, J. Wei, K. Cha, Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography, Med. Phys. 43 (2016) 6654, http://dx.doi.org/10.1118/1.4967345.
[25] Tao Tan, Zhang Li, Haixia Liu, Ping Liu, Wenfang Tang, Hui Li, Yue Sun, Yusheng Yan, Keyu Li, Tao Xu, Shanshan Wan, Ke Lou, Jun Xu, Huiming Ying, Quchang Ouyang, Yuling Tang, Zheyu Hu, Qiang Li, Optimize transfer learning for lung diseases in bronchoscopy using a new concept: sequential fine-tuning, IEEE J. Transl. Eng. Health Med. 6 (2018) 1–8, http://dx.doi.org/10.1109/jtehm.2018.2865787.
[26] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, CVPR 1 (2017) 3, http://dx.doi.org/10.1109/cvpr.2017.243.
[27] L. Zhang, L. Lu, I. Nogues, R.M. Summers, S. Liu, J. Yao, DeepPap: deep convolutional networks for cervical cell classification, IEEE J. Biomed. Health Inform. 21 (2017) 1633–1643, http://dx.doi.org/10.1109/jbhi.2017.2705583.
[28] H.A. Almubarak, R.J. Stanley, R. Long, S. Antani, G. Thoma, R. Zuna, S.R. Frazier, Convolutional neural network based localized classification of uterine cervical cancer digital histology images, Procedia Comput. Sci. 114 (2017) 281–287, http://dx.doi.org/10.1016/j.procs.2017.09.044.
[29] H. Lin, Y. Hu, S. Chen, et al., Fine-Grained Classification of Cervical Cells Using Morphological and Appearance Based Convolutional Neural Networks, 2018, arXiv Preprint arXiv:1810.06058.
[30] T. Xu, H. Zhang, X. Huang, S. Zhang, D. Metaxas, Multimodal deep learning for cervical dysplasia diagnosis, MICCAI 9901 (2016) 115–123, http://dx.doi.org/10.1007/978-3-319-46723-8_14.
[31] T. Xu, H. Zhang, C. Xin, E. Kim, L. Long, Z. Xue, S. Antani, X. Huang, Multi-feature based benchmark for cervical dysplasia classification evaluation, Pattern Recognit. 63 (2017) 468–475, http://dx.doi.org/10.1016/j.patcog.2016.09.027.
[32] A. Khrizhevsky, A. Sutskever, G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS, 2012.
[33] M. Sato, K. Horie, A. Hara, Y. Miyamoto, K. Kurihara, K. Tomio, H. Yokota, Application of deep learning to the classification of images from colposcopy, Oncol. Lett. 15 (2018) 3518–3523, http://dx.doi.org/10.3892/ol.2018.7762.
[34] Y. Yuan, W. Qin, B. Ibragimov, B. Han, L. Xing, RIIS-DenseNet: rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection, in: MICCAI 2018, Springer, Cham, 2018, pp. 620–628, http://dx.doi.org/10.1007/978-3-030-00934-2_69.
[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778, http://dx.doi.org/10.1109/cvpr.2016.90.
[36] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359, http://dx.doi.org/10.1109/TKDE.2009.191.
[37] V. Iglovikov, A. Shvets, TernausNet: U-Net With VGG11 Encoder Pre-trained on ImageNet for Image Segmentation, 2018, arXiv Preprint arXiv:1801.05746.
[38] J. Moolayil, "An Introduction to Deep Learning and Keras," Learn Keras for Deep Neural Networks, 2018, pp. 1–16, http://dx.doi.org/10.1007/978-1-4842-4240-7_2.
[39] M. Abadi, P. Barham, J. Chen, et al., Tensorflow: A System for Large-scale Machine Learning, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16) (2016) 265–283.
[40] D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, ICLR, 2015.
[41] P. Qiu, The statistical evaluation of medical tests for classification and prediction, Publ. Am. Stat. Assoc. 100 (2004) 705, http://dx.doi.org/10.1198/tech.2005.s278.