Chen H, Liu J, Wen QM *et al.* CytoBrain: Cervical cancer screening system based on deep learning technology. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 36(2): 347–360 Mar. 2021. DOI 10.1007/s11390-021-0849-3

CytoBrain: Cervical Cancer Screening System Based on Deep Learning Technology

Hua Chen[1], Juan Liu[1,*], *SeniorMember*, *CCF*, Qing-Man Wen[1], Zhi-Qun Zuo[1], Jia-Sheng Liu[1], Jing Feng[1] Bao-Chuan Pang[2], and Di Xiao[2]

[1] *Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan 430072, China*
[2] *Landing Artificial Intelligence Center for Pathological Diagnosis, Wuhan University, Wuhan 430072, China*

E-mail: {chen_hua, liujuan, csqmwen, zhiqunzuo, liujiasheng, gfeng}@whu.edu.cn
pangbaochuan@landing-med.com; xiaoboashe@hotmail.com

Received July 30, 2020; accepted February 17, 2021.

Abstract Identification of abnormal cervical cells is a significant problem in computer-aided diagnosis of cervical cancer. In this study, we develop an artificial intelligence (AI) system, named CytoBrain, to automatically screen abnormal cervical cells to help facilitate the subsequent clinical diagnosis of the subjects. The system consists of three main modules: 1) the cervical cell segmentation module which is responsible for efficiently extracting cell images in a whole slide image (WSI); 2) the cell classification module based on a compact visual geometry group (VGG) network called CompactVGG which is the key part of the system and is used for building the cell classifier; 3) the visualized human-aided diagnosis module which can automatically diagnose a WSI based on the classification results of cells in it, and provide two visual display modes for users to review and modify. For model construction and validation, we have developed a dataset containing 198 952 cervical cell images (60 238 positive, 25 001 negative, and 113 713 junk) from samples of 2 312 adult women. Since CompactVGG is the key part of CytoBrain, we conduct comparison experiments to evaluate its time and classification performance on our developed dataset and two public datasets separately. The comparison results with VGG11, the most efficient one in the family of VGG networks, show that CompactVGG takes less time for either model training or sample testing. Compared with three sophisticated deep learning models, CompactVGG consistently achieves the best classification performance. The results illustrate that the system based on CompactVGG is efficient and effective and can support for large-scale cervical cancer screening.

Keywords cervical cancer screening, visual geometry group (VGG), deep learning, artificial intelligence (AI), classification

1 Introduction

Globally, cervical cancer is one of the most lethal cancers in human life, only after breast, colorectal, and lung cancers [1–4]. In 2018, about 570 000 new cases of cervical cancer were diagnosed worldwide, accounting for 3.2% of all new cancer cases; the number of deaths of cervical cancer was approximately 311 000, accounting for 3.3% of all cancer deaths, and nearly 90% of them were in underdeveloped and developing nations due to lack of awareness of the disease and limited access to health services [5]. Without urgent attention, the number of deaths of cervical cancer is expected to reach 400 000 annually by 2030 [6]. Fortunately, regular Pap-smear testing, the most successful and effective approach in medical practice, can facilitate the early detection and screening of cervical cancer and significantly reduce the morbidity and mortality [7–11]. In Pap-smear testing, the cervical cell samples collected from the outer opening of the cervix are placed on

Regular Paper
Special Section on AI and Big Data Analysis in Biology and Medicine
This work was supported by the Major Projects of Technological Innovation in Hubei Province of China under Grant Nos. 2019AEA170 and 2019ACA161, the Frontier Projects of Wuhan for Application Foundation under Grant No. 2019010701011381, and the Translational Medicine and Interdisciplinary Research Joint Fund of Zhongnan Hospital of Wuhan University under Grant No. ZNJC201919.
*Corresponding Author

a glass slide and stained, and then a cytopathologist makes diagnosis by observing them under a microscope with naked eyes. However, a Pap-smear slide usually contains such a large number of cervical cells that it is time-consuming to observe whether there are abnormal cells [12]. Obviously, it is inefficient for a cytopathologist to make diagnosis by observing slides under a microscope. With the increase of Pap-smear slides, the clinical feasibility of manual examination approach is further limited. Additionally, different people or the same person at different times may give different results to the same slide, leading to the strong subjective results. Thus, developing an automatic cervical cytological screening system would be of great significance, and can assist cytopathologists to efficiently evaluate Pap-smear slides and to make objective diagnosis.

Thanks to the development of micro-scanning techniques, it is easy to obtain the cytological images, which make it possible to develop computer-aided cervical cancer diagnosis methods. Over the past several years, machine learning techniques have successfully been applied for screening cervical cancer. Yamal *et al.* utilized the logistic regression (LR) algorithm to distinguish cervical cancer at the cellular level, which achieved the sensitivity of 61% and specificity of 89% on the independent dataset [13]. Su *et al.* applied a two-level cascade integration of C4.5 and LR to identify the cervical cells, yielding the recognition rate of 95.6% [14]. Kurniawati *et al.* reported the accuracy of 80.18% in distinguishing cancers by using a random forest classifier [15]. Sharma *et al.* evaluated the k-nearest neighbor method for the classification of cervical cells, and obtained the accuracy of 82.9% [16]. These methods heavily depend on the handcraft features defined by cytopathologists based on their expertise, as well as accurate segmentation of cells and nuclei. Although many segmentation algorithms have been proposed and achieved good performance [17, 18], there are still great challenges in the accuracy and efficiency of cervical cell segmentation. Moreover, it is usually difficult to know what features and how many features are suitable for building a cell classification model, resulting that above methods are poor of generalization.

Of late, deep learning algorithms have been applied in cervical cell classification. After comparing several neural networks, Gupta *et al.* showed that multiple backpropagation neural networks achieved the best performance with accuracy of 95.62% and sensitivity of 95.6% on their data [19]. Wu and Zhou described a deep learning method based on convolutional neural network (CNN), and obtained the classification accuracy of 93.33% on the augmented dataset [20]. We notice that the existing methods simply gather all cervical cell images together and randomly divide them into training and test sets for training and evaluating models respectively. Since there are usually hundreds of or even tens of thousands of cells in one cytological smear, it would happen that cell images in different sets originate from the same subjects. Accordingly, test sets are not truly independent with training sets, and are unable to objectively evaluate the performance of the models.

To sum up, the following problems exist in computer-aided cervical cancer screening. First, the training and test sets are generated by dividing cervical cells images, rather than by dividing whole slide images (WSIs) corresponding to subjects, resulting that the evaluation results cannot objectively reflect the performance of the models. Second, there is a lack of public large datasets, and thus the built models are prone to over-fitting. Third, on the one hand, building the cell classifying models usually requires accurately segmenting cervical cell images at first; on the other hand, the complexity of the cervical cell morphology makes the low accuracy and efficiency of cell segmentation. Especially, the low efficiency of the cell segmentation cannot meet the needs of large-scale cervical cancer screening in practical application.

To address the above questions, we develop an AI system called as CytoBrain for the large-scale cervical cancer screening. CytoBrain consists of three main modules: the cervical cell segmentation module, deep learning based cervical cell classification module, and visualized human-aided diagnosis module, as shown in Fig.1. The cell image extraction module is responsible for efficiently extracting every single cell image in a whole slide image (WSI). The cervical cell classification module is based on our proposed CompactVGG network and is used for categorizing cells into different types. The visualized human-aided diagnosis module makes diagnosis by integrating the classification results of all cell images in a WSI and displays the results in visual ways. Specifically, the system provides two visual interaction modes: the cell display mode and the WSI display mode. In the cell display mode, the system displays the thumbnails of all cells by categories, and the user can check them one by one by scrolling the screen; the user can also click the mouse button on a cell thumbnail to view its original image and modify its classified category. In the WSI display mode, the sys-

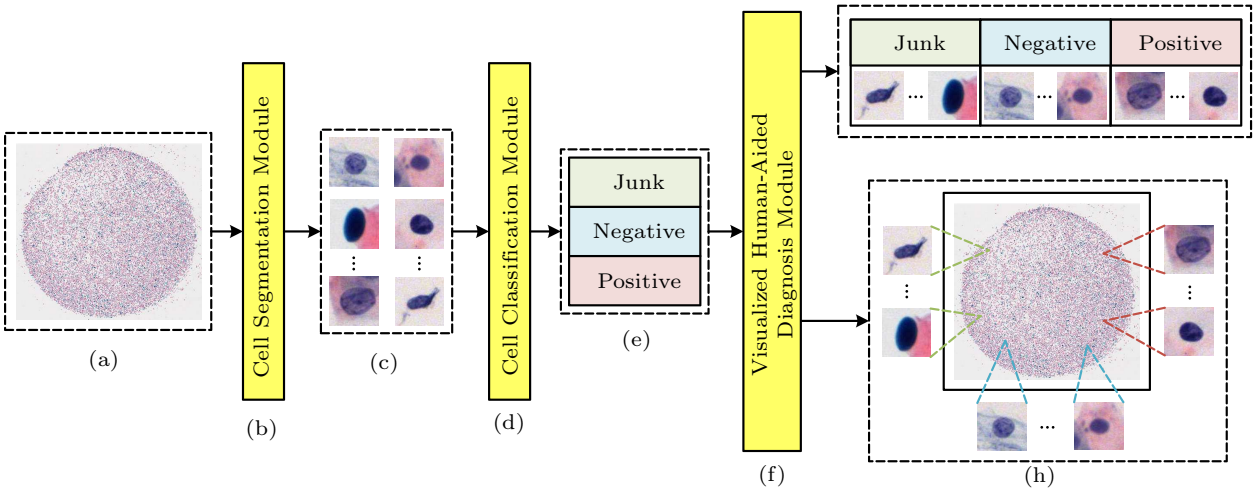


Fig. 1. Outline of the CytoBrain system. (a) Input WSI of cervical cytopathology. (b) Cell localization and segmentation module responsible for extracting every single cell image in a WSI. (c) Extracted cell images in the WSI. (d) Cell classification module that automatically categorizes cells into different types. (e) Illustration of different cell types. (f) Visualized human-aided diagnosis module. (g) Visualization interface (cell display mode). (h) Visualization interface (WSI display mode).

tem displays the WSI corresponding to a whole slice on the screen and its diagnosis result based on the integration of all cell classification results. All cells classified as abnormal are marked with bright boxes in the WSI. Users can view different areas of the WSI by scrolling the screen. The user can also click the mouse button in a marked box to view the enlarged image of the corresponding cell and modify its classified category. In this mode, users can also modify the diagnosis result of the WSI. The details of the visualized human-aided diagnosis module are not discussed in this paper.

In this paper, we focus on critical details involved in the other two modules of the system, including data preparation, cell image extraction and construction of the cell classification model. Our contributions are summarized as follows.

1) We propose a simple cell extraction method centered on the key point of the nucleus, which can extract all cell images from a WSI efficiently. Since there are usually tens of thousands of cells in a WSI, our method is more practical in large-scale cervical cancer screening than sophisticated yet inefficient cell segmentation methods.

2) We generate a large number of cervical cell samples from WSIs collected from hospitals in different geographical locations and build the largest cervical cell images dataset we have known so far. This dataset is not only helpful to build classification models which have strong generalization abilities and can be used in practical applications, but also beneficial for studying machine learning methods as a benchmark dataset.

3) We propose a compact VGG network, CompactVGG, which is more compact and has lower computing cost than existing VGGs, and thus it is more suitable for large-scale applications.

The rest of this paper is organized as follows. Section 2 introduces the data collection of cervical cell images, including WSIs collection, cell images extraction, and cell images labeling. Section 3 describes the proposed cervical cell classifier based on CompactVGG. Section 4 assesses the performance of the developed CompactVGG. Section 5 ends with concluding remarks.

2 Data Collection

2.1 WSIs Collection

The quantity and diversity of data is very important to train a good model with high performance [21]. However, there are few such public cervical cytological images. In this paper, we build a real-world clinical cervical cell images dataset to help promote the development of cervical cytology.

We conduct a retrospective study and collected cervical cell samples from 2 312 participants aged 25-64 years from multiple health care institutions in 2018 and 2019. This study abides by the ethical standards of the institutional research committee and the tenets of the Helsinki Declaration. Since this work is anonymous and retrospective, the need for an informed consent is waived.

Fig.2 provides the process of generating a WSI. For each participant, a small number of cell samples taken

from the cervix of uterus are placed on glass pieces and stained with hematoxylin and eosin (H&E) to get the Pap-smear slide. Every Pap-smear slide is automatically scanned by a micro-scanning device with high resolution to obtain a set of microscopic images, each of which corresponds to a small area on the slide. We use template matching and position fitting algorithms developed by ourselves to stitch these images to get the WSIs. Fig.3 illustrates the WSI generation procedure for the Pap-smear slide by using the above algorithms. In all, we obtained 2 312 WSIs, each of which corresponds to a participant.

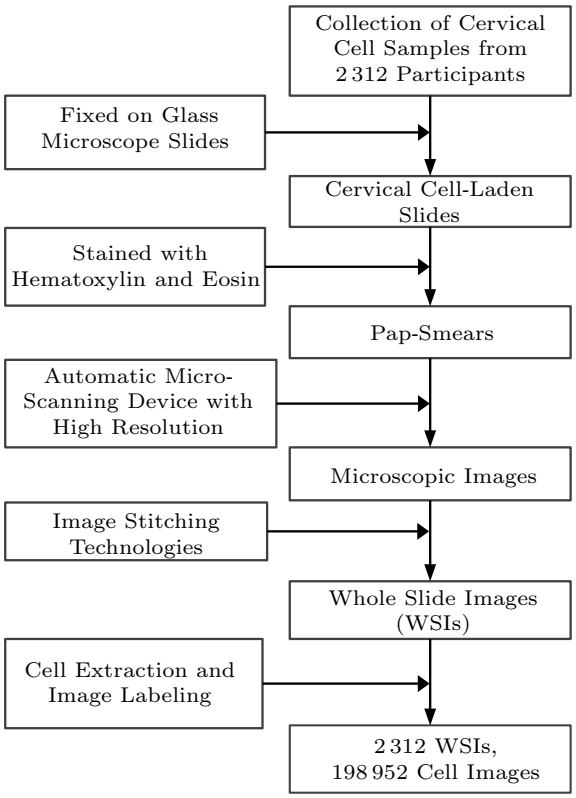


Fig.2. Outline of data collection.

2.2 Cell Images Extraction

The Bethesda system (TBS) rules mention that different cervical cytological abnormalities are related to different nucleus abnormalities, which indicates that the nucleus features in themselves have already included substantial discriminative information [22]. Thus, in order for the efficiency of CytoBrain, we propose a very simple and effective method to extract cells images rather than to accurately segment cells from every WSI in this paper. We extract the cell images from every WSI with 20x magnification. According to the statistical analysis, an area with the size of 128×128 can cover the whole nucleus and part of cytoplasm of a cell, thereby we determine the size of a cell image to be extracted as 128×128 pixels. Our cell extraction approach consists of the following steps.

First, we utilize the SURF (Speeded-Up Robust Features) algorithm [23] to detect feature points that are used for locating nuclei in the WSI (Fig.4).

Second, we utilize the Otsu method [24] to get the contours of the nuclei area and used morphological operations [25] to fill the holes and reduce the isolate noise points in the WSI (Fig.4).

Third, in order to remove stained impurities that are not real nuclei, we discard those with radius smaller than 10 pixels or greater than 150 pixels. Moreover, we further discard those with O/S ratios lower than 0.3 (the O/S ratio refers to the area obtained by using the Otsu method divided by the area gotten via using the SURF algorithm).

Finally, for every nucleus, we select one of the feature points as the key point and extracted the cell image of size 128×128 centered at the key point. In order to avoid extracting the same cell areas as many as possible, the distance between different feature points should be greater than half of the cell radius.

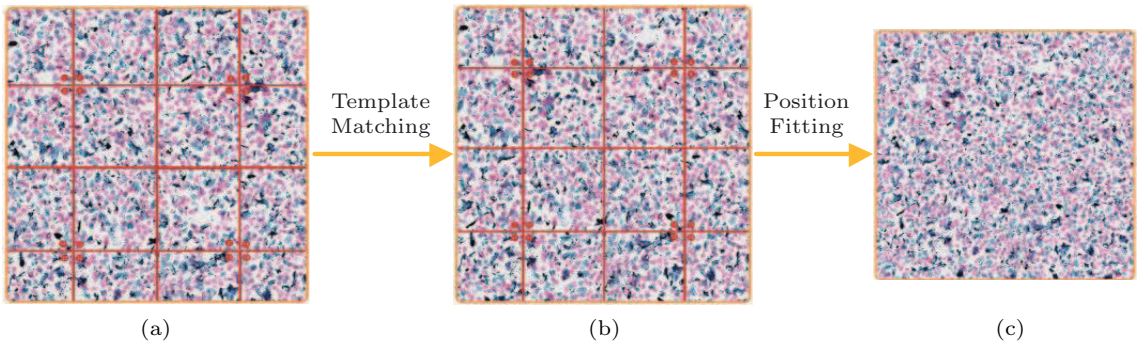


Fig.3. Example of generating WSIs. (a) Set of microscopic images with random order. (b) Template matching. (c) Position fitting to get WSIs.

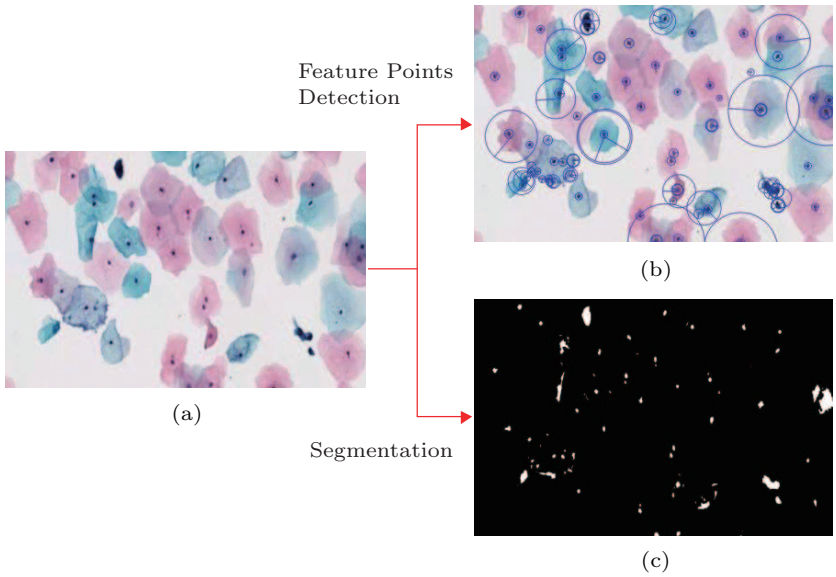


Fig.4. Steps of cell images extraction. (a) Original WSI. (b) Detecting feature points for locating nuclei. (c) Segmentation to get the contours of the nuclei area.

2.3 Cell Images Labeling

Each cell image was labeled as positive (abnormal), negative (normal) or junk by three cytopathologists back-to-back. The junk cell images include those of overlapping cells, mucus, blood cells, and impurities. Considering that junk cell images will inevitably appear in the real clinical circumstances, we incorporate a separate junk category to improve the clinical utility of the computer-aided system. Only images with the same labeling results from three experts were preserved. Ultimately, we collect 198 952 cervical cell images of 2 312 participants in all (60 238 positive, 25 001 negative, and 113 713 junk). In this work, we randomly select 450 participant samples as the independent test. As a result, we build a test set consisting of 39 792 cell images (12 048 positive, 5 001 negative, and 22 743 junk). The rest images from the remaining 1 862 participants were used as the training set. The outline of our developed dataset is listed in Table 1. The training set is used for fitting and optimizing the parameters of the model, while the independent test set is used to assess the model's performance.

Table 1. Number of Cell Images in Our Developed Dataset

Set of Cell Images	Positive	Negative	Junk
Training Set	48 190	20 000	90 970
Test Set	12 048	5 001	22 743

3 Cervical Cell Images Classifier Construction

The cervical cell classification model is the key part of CytoBrain, which is deployed in our own cervical cancer screening cloud platform. Since the accuracy and the efficiency of the model are important for the application of large-scale cervical cancer screening, we propose a compact network based on VGG networks [26] and use the cross-validation strategy to train the network. In order to improve the robustness of the model, we also use the online data augmentation technique during the training process.

3.1 Data Augmentation

Data augmentation has been widely used in the model construction for it can result in better performance, and more generalizable models invariant to certain types of image transformations and variations in image quality [27–29]. In this work, we perform the data augmentation by randomly combining operations of flipping, rotating, translating, changing brightness, and adding Gaussian noise and Gaussian blur. Each augmented cell image is assigned to the same label as the original one. Fig.5 displays an example of an input image and the corresponding augmentation.

3.2 CompactVGG Architecture

Inspired by the VGG16 network [26], in this paper we propose a more compact version of the VGG network, called as CompactVGG, to save computational

costs and speed up the calculation process to meet the needs of large-scale cervical cancer screening. The macro-architecture of CompactVGG is listed in Table 2. It consists of 10 convolution layers, four max-pooling layers, and two fully-connected layers. The convolutional layers pick up distinct features from the input cell image to generate feature maps. The max-pooling layers reduce the output dimensionality of the feature maps by merging semantically similar features into one. The fully-connected layers combine these features and output the prediction probabilities for different classes. Obviously, CompactVGG is narrower (with fewer convolution filter channels, that is, a maximum of 64 channels) and shallower (with three less convolution layers and one less full connection layer) than VGG16, which could reduce the calculation cost.

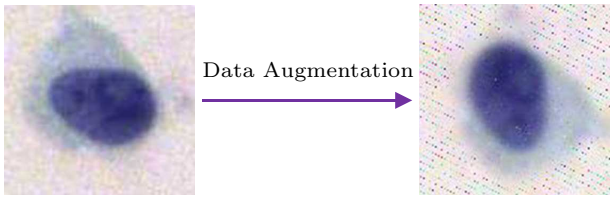


Fig. 5. Example of positive cervical cell image and the corresponding augmented image. (a) Original image. (b) Augmented image.

The essence of training process of the network is to learn the data distribution, and the output of the last layer is directly related to the input of the next layer. If the data distribution of each layer significantly shifts, the learning process needs to adjust the parameters continuously to fit different data, resulting in the long training process and even making the network

Table 2. Details of CompactVGG Architecture

Layer	Number of Neurons	Size of Feature Maps	Size of Kernels	Stride
Input	N/A	128×128×3	N/A	N/A
Conv	32	128×128×32	3×3×3	1
BN	N/A	128×128×32	N/A	N/A
Conv	32	128×128×32	3×3×32	1
BN	N/A	128×128×32	N/A	N/A
MP	32	64×64×32	2×2	2
Conv	32	64×64×32	3×3×32	1
BN	N/A	64×64×32	N/A	N/A
Conv	32	64×64×32	3×3×32	1
BN	N/A	64×64×32	N/A	N/A
MP	32	32×32×32	2×2	2
Conv	64	32×32×64	3×3×32	1
BN	N/A	32×32×64	N/A	N/A
Conv	64	32×32×64	3×3×64	1
BN	N/A	32×32×64	N/A	N/A
Conv	64	32×32×64	3×3×64	1
BN	N/A	32×32×64	N/A	N/A
MP	64	16×16×64	2×2	2
Conv	64	16×16×64	3×3×64	1
BN	N/A	16×16×64	N/A	N/A
Conv	64	16×16×64	3×3×64	1
BN	N/A	16×16×64	N/A	N/A
Conv	64	16×16×64	3×3×64	1
BN	N/A	16×16×64	N/A	N/A
MP	64	8×8×64	2×2	2
FC	128	N/A	N/A	N/A
FC	3	N/A	N/A	N/A

Note: Conv: convolution layer; BN: batch normalization layer; MP: max-pooling layer; FC: fully-connected layer; N/A: not applicable.

difficult to converge. In order to speed up the training process, we propose to reduce the internal covariate shifts by normalizing each convolutional layer in CompactVGG. Concretely, we perform a batch normalization operation [30] after each convolutional layer to perform the layer normalization by fixing the mean and the variance computed as (1) and (2).

$$\mu^l = \frac{1}{M}\sum_{i=1}^{M} a_i^l, \tag{1}$$

$$\sigma^l = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(a_i^l - \mu^l\right)^2}, \tag{2}$$

where a_i^l is the vector representation of the input to the i-th neuron in the l-th layer, and M represents the number of nearons in this paper.

In order to alleviate overfitting, VGG16 performs two dropout operations respectively along with the first and the second full connection layers. Since CompactVGG only contains two full connection layers, we adopt one dropout operation along with the first full connection layer. Moreover, we use the regularization method to further prevent CompactVGG from overfitting. Assume L_0 be the original loss function, we add the $L2$-norm regularization term to get the regularized loss function L as (3):

$$L = L_0 + \frac{\lambda}{2m}\sum_{w} w^2, \tag{3}$$

where λ is $L2$-norm regularization factor, m is the total number of the input, and w represents all weight parameters of the model. Then the updating rule for w becomes:

$$\begin{aligned} w &= w - \eta\frac{\partial L}{\partial w} \\ &= w - \eta\frac{\partial L_0}{\partial w} - \frac{\eta\lambda}{m}w \\ &= \left(1 - \frac{\eta\lambda}{m}\right)w - \eta\frac{\partial L_0}{\partial w}, \end{aligned} \tag{4}$$

where η represents the learning rate.

It can be seen from (4) that the $L2$-norm regularization term can make w smaller, which may help suppress overfitting of the model to the training data.

3.3 Training Process of CompactVGG

As a powerful strategy for model selection, cross-validation has been widely used in machine learning [31, 32]. In this work, we use the five-fold cross validation strategy to train and optimize our CompactVGG. In order to reduce the overfitting of CompactVGG, we only conduct data augmentation on the four-fold data used for training and one-fold validation data used for optimization remains no augmented in each iteration. After five iterations, we obtain five submodels. The final result for an unknown sample is the average output value of the five sub-models.

During the training phase of CompactVGG, we utilize the stochastic gradient descent algorithm with Adadelta optimizer, mini batch size, and the cross-entropy loss function [33, 34]. Besides the use of dropout technology and $L2$-norm regularization method, we also use the early stopping method during the training process to avert overfitting [35–37]. We assess the performance of CompactVGG on the one-fold validation data after each epoch. If the performance did not have improvement in 10 consecutive epochs, we consider that the performance would not improve any more, and ended the training process ahead.

4 Experiments and Results

Since CompactVGG is the key part of CytoBrain, we set up experiments to evaluate the performance of CompactVGG. First, we evaluated the training and the test time performance of CompactVGG. Now that VGG11 has the fastest training speed among the existing VGG networks under the same settings due to its shallow architecture [26], we compared the training and testing speeds of CompactVGG and VGG11. Then we evaluated the classifying performance of our CompactVGG by comparing it with the representative networks (Inception v3 [38], ResNet50 [39], and DenseNet121 [40]) on our data. And then we evaluated the effect of using $L2$-norm regularization by the ablation study. In order to verify the universality of CompactVGG, we further evaluated it on two public datasets. The hyper parameter settings of all networks were the same (Table 3). All the models were trained, validated, and tested on a workstation with dual GPUs, and with 2.20 GHz Xeon® CPU E5-2650 v4, 256 GB RAM, and 12 GB NVIDIA TITAN Xp.

4.1 Evaluation Measure

In this work, we adopted accuracy, sensitivity (recall), specificity, precision, and $F1$-score to assess the performance of the models in detecting abnormal cervical cell images. Therefore, we considered both the

Table 3. Configuration of Training Parameters

Model	Initial Learning Rate	Mini Batch Size	$L2$-Norm Regularization Factor	Dropout	Number of Epochs
CompactVGG	0.001	128	1.0	0.5	160
VGG11	0.001	128	N/A	0.5	160
ResNet50	0.001	128	N/A	0.5	160
DenseNet121	0.001	128	N/A	0.5	160
Inception v3	0.001	128	N/A	0.5	160

Note: N/A: not applicable.

negative and the junk cells as the non-positive and calculated the scores according to the following formulas:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN},$$
$$Sensitivity = Recall = \frac{TP}{TP+FN},$$
$$Specificity = \frac{TN}{TN+FP},$$
$$Precision = \frac{TP}{TP+FP},$$
$$F1\text{-}Score = 2\frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP+FP+FN},$$

where TP (true positive) is the number of correctly detected positive samples, FN (false negative) is the number of positive samples that are incorrectly identified as non-positive, TN (true negative) is the number of correctly-detected non-positive samples, and FP (false positive) is the number of non-positive samples that are incorrectly identified as positive.

4.2 Time Performance of CompactVGG

We adopted the same five-fold cross validation strategy to train the VGG11 and the CompactVGG models on the same platform, respectively, with our developed dataset. Table 4 lists the average training and testing time of VGG11 and CompactVGG in every fold respectively. In Table 4 we can see that although CompactVGG needs more training epochs to converge, the total training time of CompactVGG is less than that of VGG11. Particularly, the training speed of each epoch of CompactVGG is much faster than that of VGG11. Moreover, when the trained models are used to classify an unknown (test) sample, the processing time of the CompactVGG model is also much less than that of the VGG11 model. In Fig.6 we can also see that CompactVGG can generally achieve higher training and validation accuracies than VGG11 in the same training epochs, illustrating the goodness of learning efficiency of CompactVGG once again.

Table 4. Comparison Results of Training and Testing Time on Our Developed Dataset

Comparison Item	VGG11	CompactVGG
Number of training epochs	81.20(±20.17)	133.60(±11.55)
Training time (hr)	27.32(±6.83)	26.33(±2.23)
Time per epoch (min)	20.18(±0.08)	11.83(±0.02)
Testing time (s)	40.49(±0.43)	27.37(±0.04)
Time per sample (ms)	1.02(±0.011)	0.69(±0.001)

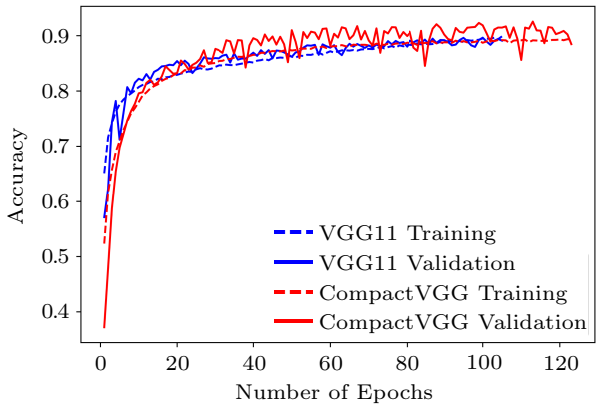


Fig.6. Comparison of the learning efficiency between VGG11 and CompactVGG.

4.3 Classification Performance on Our Developed Dataset

Table 5 summarizes the comparison results of our CompactVGG and three representative models, showing that CompactVGG achieves the highest scores on all measures in the independent test set, though it is more compact than the other networks. The reason may be that the batch normalization operation and the $L2$-norm regularization term help to get better generalization ability of our model. In order to further investigate the details on detecting different cell types, we calculated the accuracy on different cell image, listed in Table 6. In this table we can see that CompactVGG exhibits superior performance to other three models on screening positive cervical cells, negative cervical cells, and junk cells, which shows the promising performance

Table 5. Classification Performance Comparison Results (%) on Our Developed Dataset

Model	Accuracy	Sensitivity	Specificity	Precision	*F*1-Score
Inception v3	79.01±2.70	84.98±2.37	85.11±2.33	71.37±3.18	77.54±2.10
ResNet50	74.77±4.86	87.46±3.22	81.21±5.14	67.38±5.43	75.93±2.54
DenseNet121	83.13±1.54	87.91±1.25	87.58±1.25	75.48±1.79	81.21±1.17
CompactVGG	**88.30**±1.66	**92.83**±3.07	**91.03**±3.56	**82.26**±5.78	**87.04**±1.81

Note: The bold numbers represent the best results.

of CompactVGG once again. Table 6 also shows that all models perform worse on detecting junk cells than on detecting the other two types of cells. The possible reason is that there are overlapping cells and cells deformed by cell membrane rupture in the junk category, which have similar characteristics to the other two types of overlapping cells, and thus it makes the models difficult to distinguish them correctly.

Table 6. Accuracy (%) of Models for Each Cell Image Category

Model	Positive	Negative	Junk
Inception v3	84.98±2.37	84.82±1.68	74.57±4.36
ResNet50	87.46±3.22	85.08±1.12	65.78±9.54
DenseNet121	87.91±1.25	86.16±2.10	79.93±2.37
CompactVGG	**92.83**±3.07	**92.20**±2.02	**85.04**±4.21

Note: The bold numbers represent the best results.

4.4 Ablation Study on *L*2-Norm Regularization

In order to verify whether using *L*2-norm regularization can help to enhance the classification performance of CompactVGG, we conducted the ablation study. Fig.7 shows the comparison results on different performance indicators, illustrating that the *L*2-norm regularization method can effectively suppress the overfitting of our model, and thus improve the classification performance.

4.5 Evaluation on Public Datasets

To investigate the performance of CompactVGG on the third-party data, we further did experiments on two public datasets: Herlev [41] and SIPaKMeD [42]. The Herlev dataset consists of 917 cervical cell images, categorized manually by qualified cytopathologists into seven classes, of which three classes belong to normal (negative) and four classes belong to abnormal (positive). The 4 049 cells in the SIPaKMeD dataset are divided into five types according to the morphology and cellular appearance. Among them, superficial-intermediate and parabasal cells are categorized as normal (negative), metaplastic cells are benign (negative), and dyskeratotic and koilocytotic cells are abnormal (positive). The cell distributions of both datasets are listed in Table 7. In this table we can see that the Herlev dataset is very small, and the number of abnormal cell images is about three times that of the normal cell images, which may lead to the model biased towards the majority class. Although obtaining a high sensitivity is ideal from the medical standpoint, a high false positive rate is not desirable from the practical perspective. In order to balance the data as much as possible, we performed 20 (with a step size of 18 degrees) and 10 rotation transformations (with a step size of 36 degrees) on every negative and positive image respectively. After that, we did data enhancement operations (except for rotation) as described in Subsection 3.1. For the SIPaKMeD dataset, we followed the original method in [42] to do the data augmentation. The five-fold cross validation strategy was still used to evaluate the performance of the models.

Fig. 7. Result of ablation study on *L*2-norm regularization method (mean±standard deviation) where the mean of five folds is drawn by the column and the standard deviation is presented by the yellow vertical line.

Table 7. Cell Distribution of Herlev and SIPaKMeD Datasets

Dataset	Class	Cell Type	Cell Image Count	Category
Herlev	1	Superficial Squamous Epithelial	74	Normal
	2	Intermediate Squamous Epithelial	70	Normal
	3	Columnar Epithelial	98	Normal
	4	Mild Squamous Non-Keratinizing Dysplasia	182	Abnormal
	5	Moderate Squamous Non-Keratinizing Dysplasia	146	Abnormal
	6	Severe Squamous Non-Keratinizing Dysplasia	197	Abnormal
	7	Squamous Cell Carcinoma in Situ Intermediate	150	Abnormal
SIPaKMeD	1	Superficial-intermediate	831	Normal
	2	Parabasal	787	Normal
	3	Metaplastic	793	Benign
	4	Dyskeratotic	813	Abnormal
	5	Koilocytotic	825	Abnormal

4.5.1 *Investigation of Time Performance on Public Datasets*

Once again, we compared the time performance of CompactVGG and VGG11 with the fastest training speed among the existing VGG networks under the same settings. The comparison results on two public datasets are listed in Table 8. Here we can see that under the same settings, the time of CompactVGG in both the training and the testing step is less than that of VGG11 in either dataset. Specifically, the models trained by CompactVGG cost 4.60 and 1.32 milliseconds on average to test one cell image of Herlev and SIPaKMeD respectively, which are faster than those trained by VGG11. We noticed that Zhang *et al.* also tested their ConvNet model on the Hervlev dataset [21], and reported that the average testing time of ConvNet was about 3.5 seconds per image, which is more than 1 300 times slower than model our CompactVGG model. The results illustrate once again that the CompactVGG model is efficient and can be used for large-scale cervical screening system.

4.5.2 *Investigation of Classification Performance on Public Datasets*

First, we compared the classification performance of CompactVGG with that of three representative models, and the results are listed in Table 9. We can see that CompactVGG manifests best in all indicators compared with other three representative methods on the Herlev dataset, and all methods have excellent performance on the SIPaKMeD dataset. It is noticeable that different classes of cell images in the SIPaKMeD dataset are very "typical", with obviously different appearance and morphology. Therefore, it would be easy for the learning algorithms to learn the distinguishing features among different types of cells, so as to obtain good classification performance.

Though all methods' performance on the SIPaKMeD dataset is comparable, ResNet50 achieves the highest scores in three of five evaluation indicators, illustrating it is superior to the other methods. Nevertheless, ResNet50 scores the lowest in the other two indicators, indicating that its performance fluctuates

Table 8. Comparison Results of Training and Testing Time on Two Public Datasets

Dataset	Comparison Item	VGG11	CompactVGG
Herlev	Number of training epochs	87.40(±52.71)	82.00(±21.19)
	Training time (min)	175.95(±107.35)	121.04(±30.96)
	Time per epoch (min)	2.01(±0.025)	1.48(±0.014)
	Testing time (s)	0.88(±0.054)	0.83(±0.074)
	Time per sample (ms)	4.84(±0.298)	4.60(±0.410)
SIPaKMeD	Number of training epochs	196.6(±65.02)	246.40(±62.40)
	Training time (min)	46.02(±15.93)	41.32(±9.99)
	Time per epoch (s)	13.97(±0.33)	10.08(±0.34)
	Testing time (s)	1.10(±0.06)	1.07(±0.18)
	Time per sample (ms)	1.36(±0.075)	1.32(±0.221)

Table 9. Comparison Results (%) of Classification Performance on Public Datasets

Dataset	Model	Accuracy	Sensitivity	Specificity	Precision	$F1$-Score
Herlev	Inception v3	87.85±2.56	88.21±3.31	86.81±4.85	95.04±1.72	91.47±1.93
	ResNet50	85.52±2.19	85.52±2.92	85.53±6.09	94.46±2.21	89.73±1.64
	DenseNet121	92.05±1.85	93.88±1.23	86.81±5.71	95.33±1.96	94.59±1.23
	CompactVGG	**94.81**±1.08	**95.52**±1.18	**92.76**±3.23	**97.42**±1.12	**96.46**±0.74
SIPaKMeD	Inception v3	97.72±0.65	97.62±1.15	98.83±0.99	98.89±0.64	98.24±0.64
	ResNet50	97.62±0.39	97.62±1.25	**99.50**±0.38	**99.26**±0.55	**98.43**± 0.40
	DenseNet121	97.75±0.56	**98.10**±1.25	98.92±0.45	98.41±0.66	98.25±0.80
	CompactVGG	**97.80**±0.50	97.80±0.50	99.17±0.57	98.77±0.83	98.28±0.83

Note: The bold numbers represent the best results.

too much in different indicators. In contrast to ResNet, CompactVGG ranks the first or the second on four of five indicators, suggesting that it is more stable than ResNet50. Now that there are five types of cells in the SIPaKMeD dataset, we then further investigated the identification abilities of these methods for different types of cells, shown in Table 10. We can see that CompactVGG achieves the highest accuracies on koilocytic, superior-intermediate, and parabasal cells, the second highest accuracy on dyskeratotic cells, once again illustrating that CompactVGG performs more stable than ResNet50 and the other methods.

5 Conclusions

In this study, we developed an automatic cervical cancer screening system, CytoBrain, based on deep learning technology. CytoBrain consists of three main modules: cervical cell segmentation module, cervical cell classification module, and visualized human-aided diagnosis module. This paper mainly focuses on the core methods involved in the first two modules. We proposed a very simple, efficient but effective method to extract single cell images from WSIs. In the meanwhile, we proposed a compact VGG network, CompactVGG, to build the cell classifier. We applied three different strategies to alleviate the overfitting problem and improve the robustness of our CompactVGG. First, the dropout technology and the $L2$-norm regularization method were both used as regularization means in our model. Second, the early-stop technology by monitoring the performance on the validation data was utilized to pick our CompactVGG with the highest accuracy. Third, we adopted data augmentation approaches to increase the number and the heterogeneity of cervical cell images within the training set. The evaluation experimental results showed that CompactVGG has not only a faster running speed, but also better classification performance.

Although this paper has made good progress, there are still some limitations that need further improvement. First, in order to ensure the data quality, we directly discarded the cell images labeled differently by different experts. However, such discarded cell images may contain important information on the subtle differences between different cell states and should not be ignored. A further research is how to make use of such data to improve the performance of CompactVGG. Second, we only considered the majority case where a cervical cell image contains only one nucleus; however, there may be overlapping nuclei or tight clusters of cervical cells in the smear slides. Future work will extensively analyze the impact of overlapping nuclei and cell clumps on the detection accuracy of CompactVGG so that it can be expected to improve the classification of such objects. Finally, we currently focused on detect-

Table 10. Average Accuracy (%) of Models for Each Category on the SIPaKMeD Dataset

Model	Dyskeratotic	Koilocytotic	Metaplastic	Superior-Intermediate	Parabasal
Inception v3	97.78±1.78	94.06±2.20	97.85±0.85	99.40±1.04	99.62±0.57
ResNet50	98.02±1.11	93.09±1.94	**98.10**±0.77	99.28±0.50	99.75±0.57
DenseNet121	**98.64**±1.19	94.18±2.12	96.96±1.37	99.28±0.27	99.75±0.57
CompactVGG	98.02±1.02	**94.30**±2.17	97.34±2.30	**99.52**±0.27	**99.87**±0.29

Note: The bold numbers represent the best results.

ing abnormal (positive) cervical cells by using CompactVGG, while according to different stages of cervical precancerous lesion, abnormal cells can be classified as ASC-US (atypical squamous cells of undetermined significance), ASC-H (atypical squamous cells, cannot exclude high-grade squamous intra-epithelial lesion), LSIL (low-grade squamous intra-epithelial lesion), HSIL (high-grade squamous intra-epithelial lesion), CIS (carcinoma in situ) and so on. Future research should further optimize our system to distinguish specific categories of abnormal cervical cells to make it more practical in clinic.

References

[1] Jemal A, Center M M, DeSantis C, Ward E M. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiology Biomarkers and Prevention*, 2010, 19(8): 1893-1907. DOI: 10.1158/1055-9965.EPI-10-0437.

[2] Prat J, Franceschi S. Cancers of the female reproductive organs. In *World Cancer Report 2014*, Stewart B W, Wild C P (eds.), International Agency for Research on Cancer, 2014, pp.465-481.

[3] Adem K, Kiliçarslan S, Cömert O. Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Systems with Applications*, 2019, 115: 557-564. DOI: 10.1016/j.eswa.2018.08.050.

[4] Kurnianingsih, Allehaibi K H S, Nugroho L E, Widyawan, Lazuardi L, Prabuwono A S, Mantoro T. Segmentation and classification of cervical cells using deep learning. *IEEE Access*, 2019, 7(99): 116925-116941. DOI: 10.1109/ACCESS.2019.2936017.

[5] Bray F, Ferlay J, Soerjomataram I, Siegel R L, Torre L A, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 2018, 68(6): 394-424. DOI: 10.3322/caac.21492.

[6] Wittet S, Goltz S, Cody A. Progress in cervical cancer prevention: The CCA report card. Technical Report, Cervical Cancer Action, 2011. https://path.azureedge.net/media/documents/RH_cca_report_card.pdf, Mar. 2020.

[7] Schwaiger C, Aruda M, Lacoursiere S, Rubin R. Current guidelines for cervical cancer screening. *Journal of the American Academy of Nurse Practitioners*, 2012, 24(7): 417-424. DOI: 10.1111/j.1745-7599.2012.00704.x.

[8] William W, Ware A, Basaza-Ejiri A H, Obungoloch J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Computer Methods and Programs in Biomedicine*, 2018, 164: 15-22. DOI: 10.1016/j.cmpb.2018.05.034.

[9] William W, Ware A, Basaza-Ejiri A H, Obungoloch J. Cervical cancer classification from Pap-smears using an enhanced fuzzy C-means algorithm. *Informatics in Medicine Unlocked*, 2019, 14: 23-33. DOI: 10.1016/j.imu.2019.02.001.

[10] Bora K, Chowdhury M, Mahanta L B, Kundu M K, Das A K. Automated classification of Pap smear images to detect cervical dysplasia. *Computer Methods and Programs in Biomedicine*, 2017, 138: 31-47. DOI: 10.1016/j.cmpb.2016.10.001.

[11] McDonald J T, Kennedy S. Cervical cancer screening by immigrant and minority women in Canada. *Journal of Immigrant and Minority Health*, 2007, 9(4): 323-334. DOI: 10.1007/s10903-007-9046-x.

[12] Elsheikh T M, Austin R M, Chhieng D F, Miller F S, Moriarty A T, Renshaw A A. American society of cytopathology workload recommendations for automated Pap test screening: Developed by the productivity and quality assurance in the era of automated screening task force. *Diagnostic Cytopathology*, 2013, 41(2): 174-178. DOI: 10.1002/dc.22817.

[13] Yamal J M, Guillaud M, Atkinson E N, Follen M, MacAulay C, Cantor S B, Cox D D. Prediction using hierarchical data: Applications for automated detection of cervical cancer. *Statistical Analysis and Data Mining*, 2015, 8(2): 65-74. DOI: 10.1002/sam.11261.

[14] Su J, Xu X, He Y, Song J. Automatic detection of cervical cancer cells by a two-level cascade classification system. *Analytical Cellular Pathology*, 2016, 2016: Article No. 9535027. DOI: 10.1155/2016/9535027.

[15] Kurniawati Y E, Permanasari A E, Fauziati S. Comparative study on data mining classification methods for cervical cancer prediction using pap smear results. In *Proc. the 1st International Conference on Biomedical Engineering*, Oct. 2016. DOI: 10.1109/IBIOMED.2016.7869827.

[16] Sharma M, Singh S K, Agrawal P, Madaan V. Classification of clinical dataset of cervical cancer using KNN. *Indian Journal of Science and Technology*, 2016, 9(28): 1-5. DOI: 10.17485/ijst/2016/v9i28/98380.

[17] Liu Y, Zhang P, Song Q, Li A, Zhang P, Gui Z. Automatic segmentation of cervical nuclei based on deep learning and a conditional random field. *IEEE Access*, 2018, 6: 53709-53721. DOI: 10.1109/ACCESS.2018.2871153.

[18] Wang P, Wang L, Li Y, Song Q, Lv S, Hu X. Automatic cell nuclei segmentation and classification of cervical Pap smear images. *Biomedical Signal Processing and Control*, 2019, 48: 93-103. DOI: 10.1016/j.bspc.2018.09.008.

[19] Gupta R, Sarwar A, Sharma V. Screening of cervical cancer by artificial intelligence based analysis of digitized Papanicolaou-smear images. *International Journal of Contemporary Medical Research*, 2017, 4(5): 1108-1113. DOI: 10.21276/ijcmr.

[20] Wu W, Zhou H. Data-driven diagnosis of cervical cancer with support vector machine-based approaches. *IEEE Access*, 2017, 5: 25189-25195. DOI: 10.1109/ACCESS.2017.2763984.

[21] Zhang L, Lu L, Nogues I, Summers R M, Liu S, Yao J. DeepPap: Deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, 2017, 21(6): 1633-1643. DOI: 10.1109/JBHI.2017.2705583.

[22] Nayar R, Wilbur D C. The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes (3rd edition). Springer, 2015.

[23] Bay H, Tuytelaars T, Gool L V. SURF: Speeded up robust features. In *Proc. the 9th European Conference on Computer Vision*, May 2006, pp.404-417. DOI: 10.1007/11744023_32.

[24] Otsu N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems*, 1979, 9(1): 62-66. DOI: 10.1109/TSMC.1979.4310076.

[25] Soille P. Morphological Image Analysis: Principles and Applications (2nd edition). Springer-Verlag Berlin Heidelberg Publisher, 2003. DOI: 10.1007/978-3-662-05088-0.

[26] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014. https://arxiv.org/pdf/1409.1556.pdf, Mar. 2020.

[27] Wang J, Perez L. The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621, 2017. https://arxiv.org/pdf/1712.04621.pdf, Mar. 2020.

[28] Son J, Shin J Y, Kim H D, Jung K H, Park K H, Park S J. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology*, 2020, 127(1): 85-94. DOI: 10.1016/j.ophtha.2019.05.029.

[29] Chalakkal R J, Abdulla W H, Thulaseedharan S S. Quality and content analysis of fundus images using deep learning. *Computers in Biology and Medicine*, 2019, 108: 317-331. DOI: 10.1016/j.compbiomed.2019.03.019.

[30] Ba J L, Kiros J R, Hinton G E. Layer normalization. arXiv:1607.06450, 2016. https://arxiv.org/pdf/1607.06450.pdf, Mar. 2020.

[31] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. the 14th International Joint Conference on Artificial Intelligence*, Aug. 1995, pp.1137-1143.

[32] Refaeilzadeh P, Tang L, Liu H. Cross-validation. In *Encyclopedia of Database Systems*, Liu L, Özsu M T (eds.), Springer, 2016, pp.532-538. DOI: 10.1007/978-0-387-39940-9_565.

[33] Zeiler M D. ADADELTA: An adaptive learning rate method. arXiv:1212.5701, 2012. https://arxiv.org/pdf/1212.5701.pdf, Mar. 2020.

[34] Xiang S, Liang Q, Hu Y, Tang P, Coppola G, Zhang D, Sun W. AMC-Net: Asymmetric and multi-scale convolutional neural network for multi-label HPA classification. *Computer Methods and Programs in Biomedicine*, 2019, 178: 275-287. DOI: 10.1016/j.cmpb.2019.07.009.

[35] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014, 15: 1929-1958.

[36] Nielsen M. Improving the way neural networks learn. http://neuralnetworksanddeeplearning.com/chap3.html, Mar. 2020.

[37] Wang Y, Shen X, Yang Y. The classification of Chinese sensitive information based on BERT-CNN. In *Artificial Intelligence Algorithms and Applications*, Li K, Li W, Wang H, Liu Y (eds.), Springer, 2020, pp.269-280. DOI: 10.1007/978-981-15-5577-0_20.

[38] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. arXiv:1512.00567, 2015. https://arxiv.org/pdf/1512.00567.pdf, Mar. 2020.

[39] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp.770-778. DOI: 10.1109/CVPR.2016.90.

[40] Huang G, Liu Z, Maaten L V D, Weinberger K Q. Densely connected convolutional networks. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp.2261-2269. DOI: 10.1109/CVPR.2017.243.

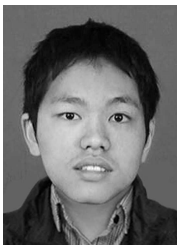
[41] Paul P R, Bhowmik M K, Bhattacharjee D. Automated cervical cancer detection using Pap smear images. In *Proc. the 4th International Conference on Soft Computing for Problem Solving*, Dec. 2014, pp.267-278. DOI: 10.1007/978-81-322-2217-0_23.

[42] Plissiti M E, Dimitrakopoulos P, Sfikas G, Nikou C, Krikoni O, Charchanti A. SIPaKMeD: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images. In *Proc. the 25th IEEE International Conference on Image Processing*, Oct. 2018, pp.3144-3148. DOI: 10.1109/ICIP.2018.8451588.

Hua Chen is a Ph.D. candidate in the School of Computer Science, Wuhan University, Wuhan. His current research interests include deep learning, medical image processing, and image classification and segmentation.

Juan Liu received her B.S. and Ph.D degrees in computer science from Wuhan University, Wuhan, in 1991 and 1996 respectively. She is now a professor in School of Computer Science at Wuhan University, Wuhan. Her research interests include machine learning, data mining, bioinformatics, pattern recognition, and artificial intelligence methods for medicine.

Qing-Man Wen is a Master student in School of Computer Science, Wuhan University, Wuhan. Her current research interest is artificial intelligence methods in medicine.



Zhi-Qun Zuo is a Master student in School of Computer Science, Wuhan University, Wuhan. His current research interest is artificial intelligence methods in medicine.

Jia-Sheng Liu received his Master's degree in School of Computer Science, Wuhan University, Wuhan, in 2020. His research interests are image processing and deep learning.

Jing Feng received his Ph.D. degree in software engineering from Wuhan University, Wuhan, in 2007. His current research interests include bioinformatics, pattern recognition, data mining, machine learning, and artificial intelligence methods in medicine.

Bao-Chuan Pang received his Ph.D. degree from Huazhong University of Science and Technology, Wuhan, in 2009. His is now a senior scientist of Landing Artificial Intelligence Center for Pathological Diagnosis, Wuhan. His current research interests include digital pathology and medical image processing.

Di Xiao is a senior scientist in Landing Artificial Intelligence Center for Pathological Diagnosis, Wuhan. His current research interests include big data and deep learning.