Morphologic Classification and Automatic Nugent Scoring of Bacterial Vaginosis by Deep Neural Networks

Zhongxiao Wang[b,c,*], Lei Zhang[a,*], Ying Wang[a], Yufeng Wang[a], Zhaohui Liu[d], Huihui Bai[d], Wei Wu[c], Weike Mo[c], Ruifang An[g], Jiao Li[g], Na Li[g], Ping Li[f], Xin Zeng[f], Can Rui[f], Chong Fan[f], Li Geng[j], Xinhuan Liu[j], Min Zhao[e], Weipei Zhu[h], Lin Qi[h], Qiao Qiao[i], Zitao Wang[i], Yanyan Si[l], Andrea Feng[k], Mingxuan Li[c], Qiongqiong Zhang[a,n], Mengdi Wang[m], Qinping Liao[a,n,**] and Wei Xu[b,**]

[a] *Department of Obstetrics and Gynecology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing 102218, China*

[b] *Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China*

[c] *Suzhou Turing Microbial Technologies Co., Ltd, Suzhou 215021, China*

[d] *Beijing Obstetrics and Gynecology Hospital,Capital Medical University Beijing Maternal and Child Health Care Hospital, Beijing 100000, China*

[e] *Peking University First Hospital, Beijing 100035, China*

[f] *Women's Hospital of Nanjing Medical University, Nanjing Maternity and Child Health Care Hospital, Nanjing 210004, China*

[g] *The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China*

[h] *The Second Affiliated Hospital of Soochow University, Suzhou 215004, China*

[i] *The Affiliated Hospital of Inner Mongolia Medical University, Hohhot 010050, China*

[j] *Peking University Third Hospital, Beijing 100191, China*

[k] *Beijing HarMoniCare Women's and Children's Hospital, Beijing 100029, China*

[l] *Binzhou Medical University Hospital, Binzhou 256600, China*

[m] *Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544*

[n] *School of Clinical Medicine, Tsinghua University, Beijing 100084, China*

ARTICLE INFO

Keywords:
Bacterial Vaginosis
Deep Learning
Convolutional Neural Network
Nugent Score
Microscopic Images

ABSTRACT

Background: Bacterial vaginosis (BV) was the most common condition for women's health caused by the disruption of normal vaginal flora and an overgrowth of certain disease-causing bacteria, affecting 30-50% of women at some time in their lives. Gram stain followed by Nugent scoring (NS) was long considered golden standard and based on bacterial morphotypes under the microscope. This conventional manual method often gave variable results among different technologists.

Methods: We created a convolutional neural network (CNN), and evaluated its ability to automatic identify vaginal bacteria and classify Nugent scores from microscope images. All the CNN models were first trained with 23280 microscopic images diagnosed and archived either positive or negative for BV. A separate set of 5815 images were evaluated by the CNN model and technologists/obstetricians independently. The CNN model's generalization ability was evaluated on total independent test sets of 1082 images collecting from three medical institutions.

Results: Our model could classify Nugent Scores at the image-level with high sensitivity (82.4%) and specificity (96.6%), which was more consistent and had better diagnostic yield than the top-level technologists and obstetricians in China. The speed of our CNN model was much faster than human reader. The generalization ability of our model was strong and the model could be deployed in more medical institutions.

Conclusion: The CNN model over performed human readers on accuracy, efficiency and stability for BV diagnosis using microscopic image-based Nugent scores.

1. Introduction

Abnormal vaginal discharge and odor were vaginitis symptoms affected millions of women globally and represented the most commen reasons for women to visit clinics. Bacterial vaginosis (BV, 40-50%), vulvovaginal candidiasis (VVC,

*These authors contributed equally to this work
**corresponding authors
ORCID(s):

20-25%) and trichomoniasis (TV, 15-20%) were the leading causes of vaginitis. Bacterial vaginosis (BV) represented a dysbiosis of the vaginal microbiome that was associated with significant adverse healthcare outcomes, including preterm labor resulting in low birth weight, pelvic inflammatory disease, acquisition of the human immunodeficiency virus and increased susceptibility to sexually transmitted infections [1–8]. In the United States, women had a high BV incidence rate of 29.2% with the prevalence varies with race: African-American (51%), Hispanic (32%), and Whites (23%) [9]. In China, the prevelance of BV in a few cities with survey data was between 15-20%, which represented more than 100 million women [10].

In 1991, Nugent et al [11] reported the use of a numerical score to diagnose BV by semiquantization of gram-positive rods, gram-negative coccobacilli forms, and curved gram-negative rods after Gram staining. These morphotypes were thought to represent Lactobacillus spp., Gardnerella vaginalis and Mobiluncus spp., respectively. Nugent scoring had since then become the 'gold standard' for laboratory diagnosis of BV [7, 8, 13]. In Nugent scale, scores of 0–3 were considered to have normal vaginal flora (Lactobacillus dominant); scores of 4–6 were labeled as altered vaginal flora (mixed morphotypes); and scores of 7–10 were indicative of BV (absence of lactobacilli and predominance of the other 2 morphotypes). Alternative diagnostic methods such as molecular diagnostic assays, enzymatic assays, and chromogenic point-of-care test (POCT) were compared to Nugent criteria [12]. However, the determination of a Nugent score by a microbiologist was influenced by individual skill and was time consuming [12, 13]. In addition, the number of experienced microbiologists or technologist performing the microscopic work was extremely imbalanced among different countries and districts [14]. Needed was a more efficient method to classify Nugent scores.

Here, we provided a proof of concept for a deep-learning-based model to quantify Gram stain and automated classification of Nugent scores. Recently, a traditional image processing method was developed for automatic bacterial vaginosis diagnosis. However, the sensitivity (58.3%) and specificity (79.1%) was relatively poor in comparison to expert due to its limitation in the image processing algorithm [13]. Deep learning method especially convolutional neural network (CNN) models had demonstrated excellent performance on computer vision tasks including image classification, image semantic segmentation and image object detection. For the image classification, various CNN models were constructed with increasing performance on natural image classification. Numerous CNN models, including LeNet-5, AlexNet, VGGNet, and ResNet, were developed to improve performance on natural image recognition [15–23]. Many models were proved effective for medical image processing, from identifying diabetic retinopathy in retinal fundus photographs [24–26], endoscopic images [27], to microbiology recognitions [28, 29]. We hypothesized that CNN based deep learning models can be used to diagnose BV using Nugent score classifications efficiently and accurately. First we developed several CNN models to learn images previously diagnosed and curated by obestetricians and microbiologiests. Second, a trained model were used to test on a separate image set from the same hospital. Our trained model was subsequently evaluated for accuracy in comparison to expert classification. Finally, three independent test sets collected from three different medical institutions were used to verify the CNN model's generalization ability.

2. Material and Methods

2.1. Image Data Preparation

A total of 29095 microscope images including associated medical records from January, 2018 to September, 2019 at Beijing Tsinghua Changgung Hospital were retrieved. The diagnosis of BV was made by the experts including two chief obestetricians and three microbiologists based on patients' symptoms and microscopic images. Each patient was diagnosed by the microbiologist then reviewed by the chief obstetrician. One-fifth of the samples were randomly selected as the test set (5815 samples), and the rest samples (23280 samples) were used as the training set. The study protocol was approved by the Ethics Committee of the Beijing Tsinghua Changgung Hospital.

The resolution of microscope images was 1024×768 pixels. The labeled nugent scores ranged from 0 to 10. Due to the fact that determination of a nugent score was subjectived, the eleven scores were divided into three groups: normal vaginal flora (0-3 scores), altered vaginal flora (4-6 scores) and BV(7-10 scores). Fig.1(a) showed the representative microscope images with different groups.

The training set contained 16490 normal vaginal flora microscope images, 4660 microscope images with altered vaginal flora and 2130 BV microscope images. The test set included 4120 normal vaginal flora microscope images, 1164 microscope images with altered vaginal flora and 531 BV microscope images. Patients with altered vaginal flora were considered positive or negative by obstetricians taking the clinical symptoms into account. For our algorithm, it was considered altered vaginal flora and BV as positive when we calculated the sensitivity and specificity. Fig.1(c) illustrated the distribution of the three classes of images.

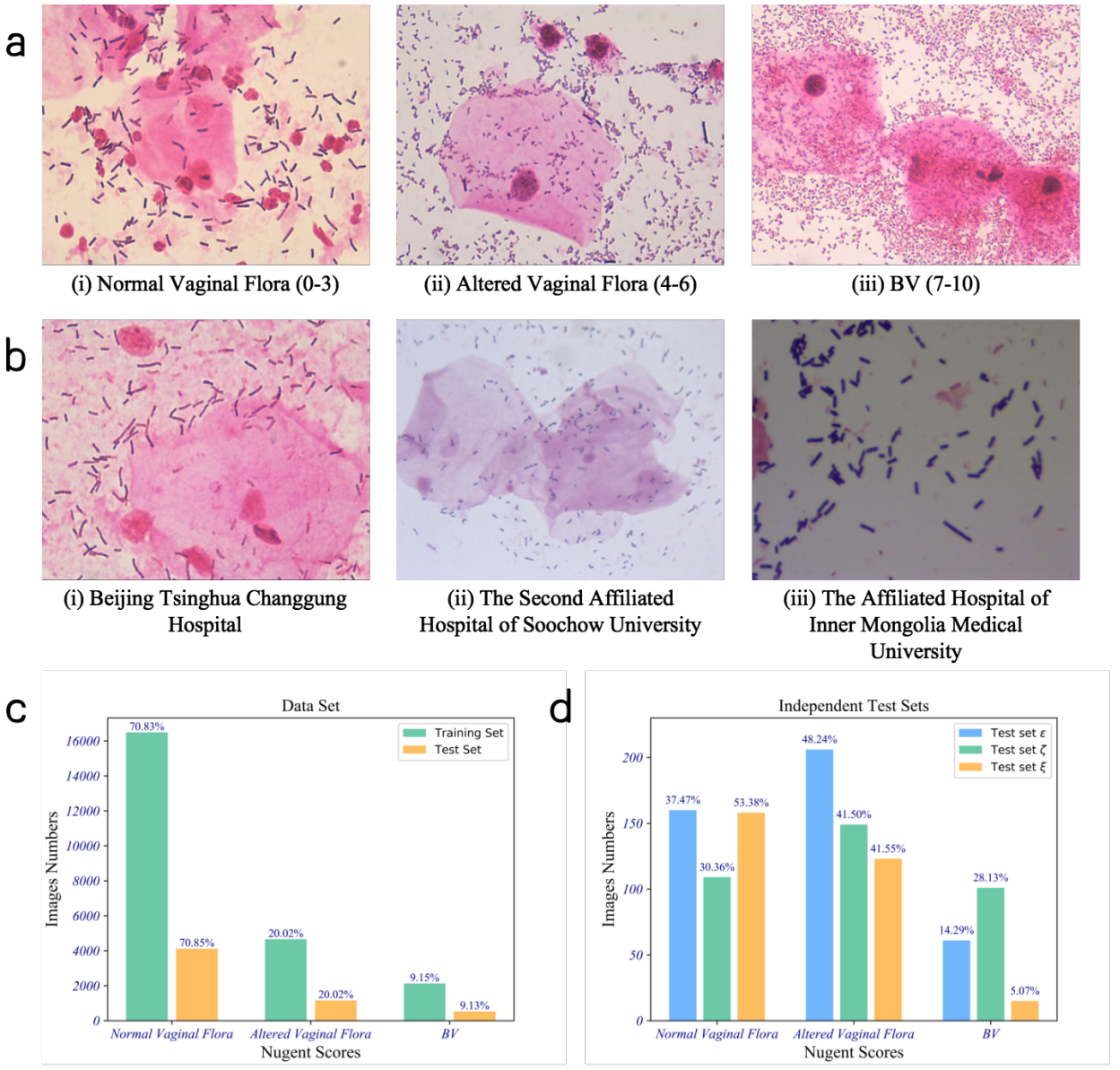

Figure 1: The data set information. a) Three typical samples for (i) Normal Vaginal Flora, (ii) Altered Vaginal Flora, and (iii) BV collecting from Beijing Tsinghua Changgung Hospital. b) Three typical samples collecting from (i) Beijing Tsinghua Changgung Hospital, (ii) The Second Affiliated Hospital of Soochow University, and (iii) The Affiliated Hospital of Inner Mongolia Medical University. c) The distribution of the data set. d) The distribution of the three independent test sets ϵ from Beijing Tsinghua Changgung Hospital, ζ from The Second Affiliated Hospital of Soochow University, and ξ from The Affiliated Hospital of Inner Mongolia Medical University.

For verifying our model's generalization ability and comparing our model with experts, three independent test sets ϵ, ζ, ξ were constructed. ϵ was randomly selected from the test set above, ζ and ξ were collected from other two medical institutions containing The Second Affiliated Hospital of Soochow University and The Affiliated Hospital of Inner Mongolia Medical University. The resolution of the images in ζ and ξ was 1280×1024 pixels. To standardize the samples, the center 1280×960 pixels were cropped and resized to 1024×768 pixels. Fig.1(b) showed three typical samples from the three different medical institutions above. The distribution of the three test sets were illustrated in Fig.1(d). The study protocol was also approved by the Ethics Committees of The Second Affiliated Hospital of Soochow University and The Affiliated Hospital of Inner Mongolia Medical University.

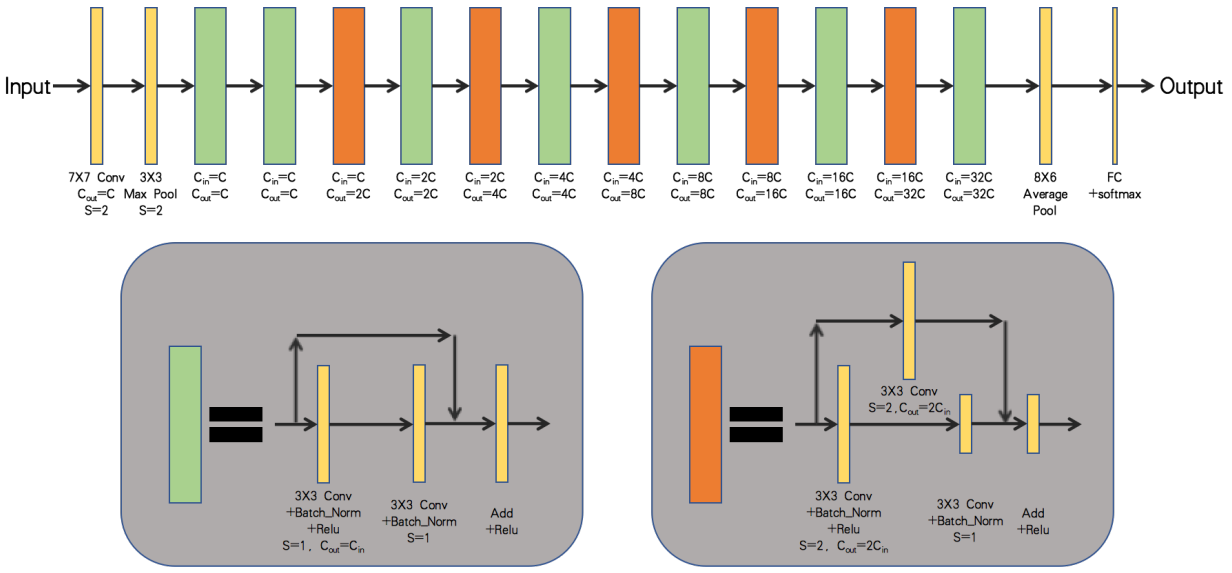

Figure 2: The architecture of our models. Conv represented convolutional layer and batch norm was the batch normalization layer. S represented the stride of the convolution operation. FC was the fully connected layer. C_{in} and C_{out} were the input channel and output channel. C determined the width of the network

2.2. The CNN Method

We developed four CNN models with different network widths (different channel numbers in each layers) to predict nugent scores based on microscope images. The models mainly employed the residual module used in Resnet [18]. In the training process, color jittering, scale jittering and horizontal / vertical flip were used as the data augmentation methods. By comparing the performance of the four models, the best model with the best network width was selected.

In clinical practice, firstly, the inspectors inspected multiple fields of vision of a sample under a microscope, each field of vision was diagnosed with a Nugent score. Secondly, a comprehensive diagnostic result was given by considering each vision's result. Finally, a field of vision representing the comprehensive diagnosis result would be selected and saved as image. Our model could automatically achieved each vision's Nugent score and further automatically obtained the comprehensive diagnostic result by using automatic scanning microscope.

2.2.1. The Basic CNN Model

The classic classification convolutional neural network was only suitable for input pictures with a resolution of 224×224, such as VGG, GooLeNet, ResNet and DenseNet [16–19]. But the resolution of our microscope images was 1024×768. Therefore, we developed a new CNN model named NugentNet to adapt the input of our microscope images. The architecture of our model was plotted in Fig.2. In this model, conv represented convolutional layer and batch norm was the batch normalization layer. S represented the stride of the convolution operation. FC was the fully connected layer. C_{in} and C_{out} were the input channel and output channel, C equalled to 64 for the basic model.

The model was trained to minimize a cross entropy loss function given by

$$J(\theta) = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} y_{ji}^{label} log(y_{ji}^{prediction}) \tag{1}$$

where m, n were the number of the classes and the batch size; y^{label} was the one-hot encode vector of the label; $y_{ji}^{prediction} = f(\theta; x_j)$ was a vector with the elements represented the prediction probabilities for each class, which was obtained by using a softmax after the last fully connected layer in the CNN model. x_j was the input data and θ were the variables for updating. We used momentum optimizer [30] to train the model on the labeled images.

The data set only had 29095 samples, but the number of parameters of the basic model for training was much larger than Resnet18. The basic model needed more samples to train and showed overfit on the data set. Therefore, we further developed three compression model to obtain the best fit model.

Table 1
Different hardwares used by different hospitals to generate microscope images and the main differences of the samples.

Hospital	Camera Models	Camera Sensor Size	C-Mount Magnification	Target Physical Area	Resolution	Test Set Pixel Average Values (B,G,R)
Beijing Tsinghua Changgung Hospital	MC170HD	$6.16mm \times 4.62mm$	0.7×	$87.02um \times 65.26um$	1024×768	(187, 172, 226)
The Affiliated Hospital of Inner Mongolia Medical University	UI-3240LE-C-HQ	$7.18mm \times 5.32mm$	1×	$67.84um \times 54.27um$	1280×1024	(155, 111, 117)
The Second Affiliated Hospital of Soochow University	UI-3240LE-C-HQ	$7.18mm \times 5.32mm$	0.5×	$135.68um \times 108.54um$	1280×1024	(208, 181, 194)

2.2.2. The Compression CNN Models

The compression models (1/2 NugentNet, 1/4 NugentNet, 1/8 NugentNet) were based on the basis model. In the compression models, the number of channels for every convolutional layers were reduced that the value of C in Fig.2 was reduced to 32 for 1/2 NugentNet, 16 for 1/4 NugentNet and 8 for 1/8 NugentNet.

The speeds of the compression models were faster than the basic models in the inference process. The used memory of GPU was also reduced. Therefore, the compression models could significantly save the diagnosis times and use less compute resources.

2.3. Comparing with Human Reader

To compare the performance of our model and human readers on BV diagnosis, we randomly selected test set ϵ with 427 samples from the original 5815 test sets. The test set were further independent labeled by two senior experts [1] with more than 10 years of diagnosis experience, the disagreement results were arbitrated by a professional microbiologist [2] with more than 30 years of diagnosis experience. The 427 samples were labeled as 160 normal vaginal flora, 206 altered vaginal flora and 61 BV images. Our model and five human readers, including three senior inspectors and two top experts, were independently tested on this test set. The three senior inspectors were from Beijing HarMoniCare women's and children's Hospital, Binzhou Medical University Hospital, and Women's Hospital of Nanjing Medical University, Nanjing Maternity and Child Health Care Hospital. The two top experts were from The First Affiliated Hospital of Xi'an Jiaotong University and The Second Affiliated Hospital of Soochow University.

2.4. Verifying the Generalization Ability of the Model

To verify the generalization ability of our model, two additional independent test sets ζ and ξ were collected. ζ from The Second Affiliated Hospital of Soochow University and ξ from The Affiliated Hospital of Inner Mongolia Medical University. ζ had 359 samples containing 109 normal vaginal flora images, 149 altered vaginal flora images and 101 BV images and ξ had 296 samples including 158 normal vaginal flora images, 123 altered vaginal flora images and 15 BV images. These two test sets were labeled by the same experts that labeled test set ϵ.

The three independent test sets ϵ, ζ, ξ collected from different hardwares. Tab.1 showed the different hardwares used by the above three hospitals to generate microscope images. Fig.1(b) showed three different typical samples collecting from the three hospitals above. Fig.1(b[i]) in ϵ collected from Beijing Tsinghua Changgung Hospital, Fig.1(b[ii]) in ζ collected from The Second Affiliated Hospital of Soochow University and Fig.1(b[iii]) in ξ collected from The Affiliated Hospital of Inner Mongolia Medical University. The pixel distribution of these three types of samples were significantly difference because they were generated by different hardwares. Fig.1(b) showed there were huge differences between the samples in ζ, ξ and the samples in the training set. The main differences were the actual physical area represented by the image and the brightness of the image. The actual physical area and the statistical average of the pixel values of three channels on the three test sets were showed in Tab.1. The actual physical area of

[1] From Beijing Tsinghua Changgung Hospital and Beijing Obstetrics and Gynecology Hospital, Capital Medical University Beijing Maternal and Child Health Care Hospital.

[2] From Beijing Tsinghua Changgung Hospital.

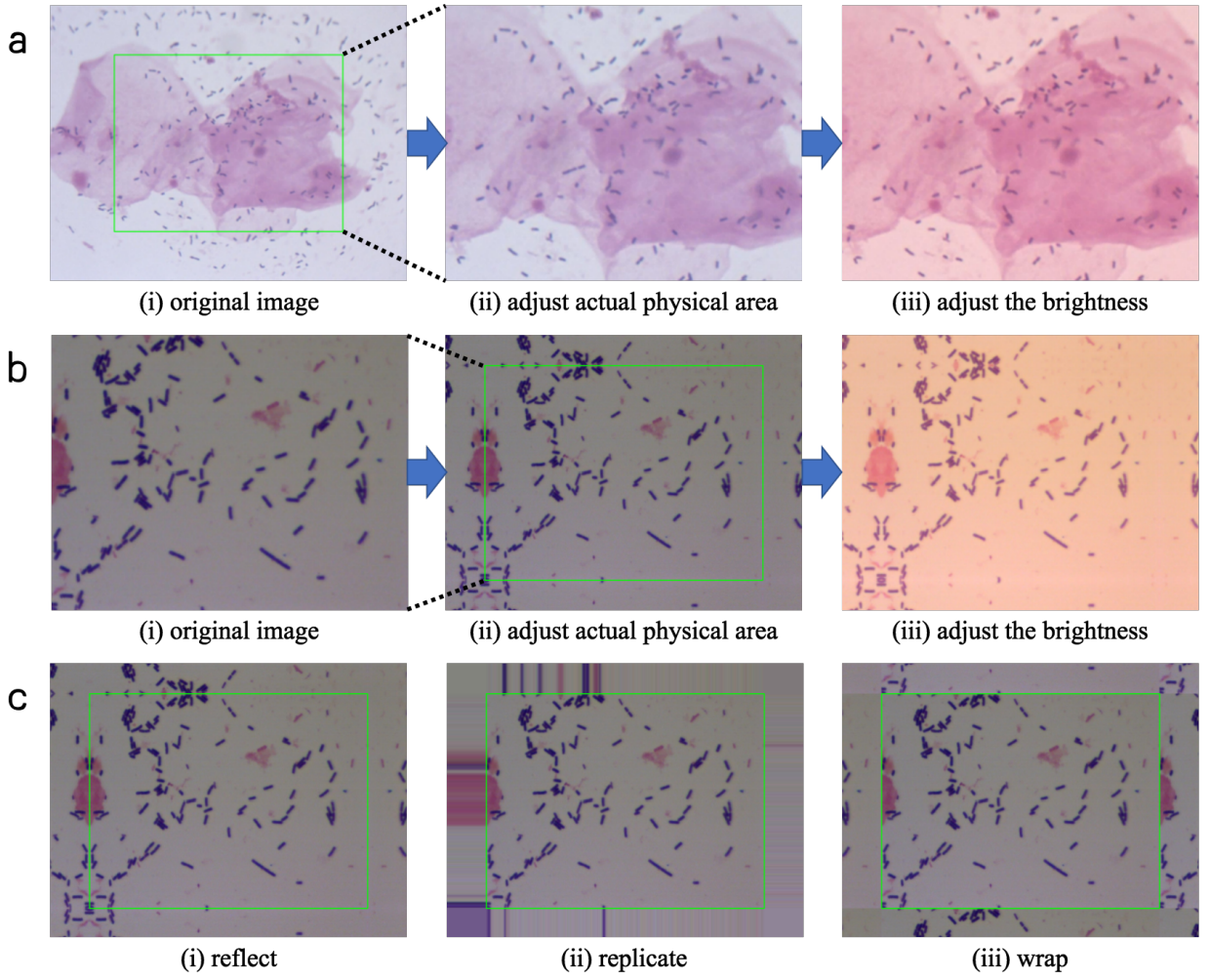

Figure 3: The preprocessing for test set ζ and ξ. a) was the preprocessing for ζ. b) was the preprocessing for ξ. c) was three typical edge expansion methods used in b[ii].

the sample in ζ was twice that of the sample in the training set and the images were brighter. In contrast, the images in ξ represented only half actual physical area and were darker.

In the inference process, we used preprocessing to eliminate sample differences in different test sets. The preprocessing include two steps: first, standardize the actual physical area of the images; second, adjust the brightness of the images. Fig.3 showed the preprocessing. The preprocessing for test set ζ was showed in Fig.3(a). First, the center 656×492 pixels were cropped and resized to 1024×768 pixels. Second, all pixel values increased 32 for red channel, -9 for green channel and -21 for blue channel. Fig.3(b) showed the preprocessing of test set ξ. First, the image was resized to 798×598 pixels and fellowed edge expansion, Fig.3(c) showed three typical edge expansion methods: replicate, wrap and reflect. Our results showed the reflect method was the best. Second, all pixel values increased 109 for red channel, 61 for green channel and 32 for blue channel.

2.5. Training with Different Sample Size

To investigate the performance of our best model with different training sample sizes, we trained the best model with five different sample sizes including 5000 images, 10000 images, 15000 images, 20000 images, and 23000 images.

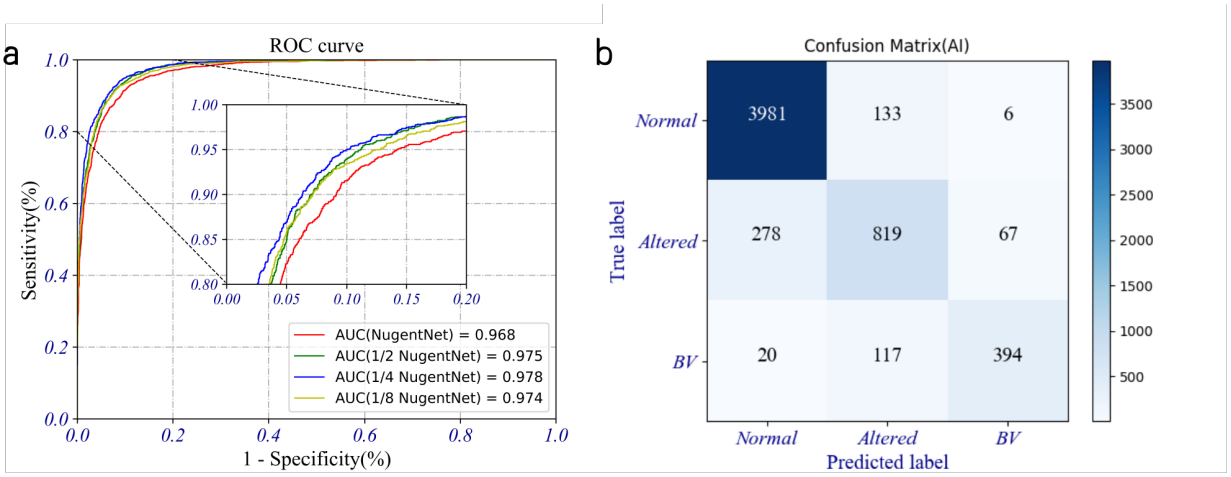

Figure 4: The performance of our models on the test set. a) was the ROC curves of our four models, the 1/4 NugentNet was the best model getting AUC=0.978. b) was the confusion matrix of the best points of the best model.

2.6. The Metrics Methods

The performance of the experts were usually measured by sensitivity and specificity. Sensitivity represented the true positive rate and specificity represented the true negative rate. The two diagnostic indexes was calculated by

$$\begin{aligned} Sensitivity &= \frac{P_T}{P_T + N_F}, \\ Specificity &= \frac{N_T}{N_T + P_F}, \end{aligned} \tag{2}$$

where P_T was the number of the positive samples being correctly diagnosed (True Positive), N_F was the number of the positive samples being diagnosed as negative (False Negative), N_T was the number of the negative samples being correctly diagnosed (True Negative), P_F was the number of the negative samples being diagnosed as positive (False Positive). In this study, we considered normal vaginal flora as the negative samples, altered vaginal flora and BV as the positive samples.

The performance of our models was illustrated by AUC (area under ROC curve). ROC curve was a graphical plot that illustrated the diagnosis ability of a binary classifier system as its discrimination threshold was varied [31]. The ROC curve was created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [31]. The true positive rate was known as sensitivity and the false positive rate was equal to 1-specificity.

To show more performance details of our models and human readers, the confusion matrix was employed to illustrate the prediction results of the three groups. In the confusion matrix, each row of the matrix represented the instances in a predicted class while each column represented the instances in an actual class (or vice versa) [32]. The accuracy of three groups of classifications was also provided as detail information.

3. Results

The test set was used to evaluate the performance of BV diagnosis. The Nugent sores were classified into three groups as described above. The performance of the results could be illustrated by the ROC curve. For the purpose of screening, we considered altered vaginal flora and BV as positive samples and normal vaginal flora as negative samples.

Fig.4(a) showed the result of the four models. All the AUCs were greater than 0.95. The best model was the 1/4 NugentNet with AUC = 0.978. The NugentNet and 1/2 NugentNet showed lower AUC because they had more training variables than the best model, therefore, these two models showed overfitting on our data set. The number of training variables of 1/8 NugentNet was only a quarter of 1/4 NugentNet. The AUC of the 1/8 NugentNet was 0.004 lower than the best model, therefore, it showed underfitting on our data set. To show more detail information of the performance

Table 2
The performance of the best model and five human readers on the independent test set ζ and the total independent test sets.

	Independent Test Set ϵ			Total Independent Test Sets		
	Sensitivity	Specificity	Three Classification Accuracy	Sensitivity	Specificity	Three Classification Accuracy
Best Points(AI)	91.4%	91.3%	80.3%	89.0%	85.0%	75.1%
Senior Inspector1	96.6%	65.0%	66.0%	97.0%	60.0%	67.8%
Senior Inspector2	96.3%	50.0%	62.1%	95.3%	60.9%	67.5%
Senior Inspector3	97.4%	67.5%	67.5%	97.3%	65.6%	70.3%
Top Experts1	93.3%	91.9%	79.6%	94.4%	93.9%	80.9%
Top Experts2	88.0%	91.3%	74.7%	90.4%	92.7%	78.7%
Average (Senior Inspectors)	96.8%	60.8%	65.2%	96.5%	62.2%	68.5%
Average (Human Readers)	94.3%	73.1%	70.0%	94.9%	74.6%	73.0%

of the best model, the three classification results (Confusion Matrix) of the best points of the best model was showed in Fig.4(b). The best points obtained 89.3% three classification accuracy, which was 5.6% higher than the microscopists and only 1.4% lower than the top experts' performance showing in Ref. [13]. The results showed only 3.8% (20/531) BV samples were predicted as normal vaginal flora and 0.1% (6/4120) normal vaginal flora samples were predicted as BV.

The performance of our model and five human readers on the single independent test set ϵ (427 sample from Beijing Tsinghua Changgung Hospital) showed in Fig.3(a). Our model obtained AUC=0.975 and overperformed all senior inspectors and better than one of the two top experts. Tab.2 showed the detail information. The average performance of all the human readers was 94.3% sensitivity and 73.1% specificity. Our model overperformed the human readers' average level. When setting sensitivities equal, the model's specificity was 16.3% higher than the average result, when specificity was setted equal, the specificity was 5.3% higher. Our model obtained three classification accuracy of 80.3%, which was 10.3% higher than the human's average result and 3.1% higher than the top experts' average result.

Table 3
The performance of the best model using preprocessing on test sets ζ and ξ.

	Independent Test Set ζ				Independent Test Set ξ			
	Original Image	Adjust Physical Area	Adjust Bright-ness	Adjust Both	Original Image	Adjust Physical Area	Adjust Bright-ness	Adjust Both
AUC	0.8552	0.9375	0.8626	0.9396	0.5137	0.7136	0.8894	0.9450

The performance of our model using preprocessing on test set ζ and ξ was illustrated in Tab.3. The test results of ζ showed that for larger actual physical area and brighter samples, adjusting physical area greatly improved (8.23%) the performance, but adjusting brightness only improved (0.74%) a little. The test results of ξ showed that for smaller actual physical area and darker samples, both adjusting methods greatly improved the performance. The performance of three edge expansion methods in adjusting physical area step on ξ was showed in Tab.4. The results showed reflect was the best edge expansion method. The total comparison results of the machine and the five human readers (three senior inspectors and two top experts) on the total test set containing test set ϵ, ζ, ξ were showed in Fig.5(b). Both adjusting physical area and adjusting brightness greatly improved model performance on the total independent test set. The best result, obtained AUC=0.94, overperformed all senior inspectors. Tab.2 showed the detail information. The average performance of all the senior inspectors was 96.5% sensitivity and 62.2% specificities. When setting sensitivities equal, the model's specificity was 8.8% higher than the average result, when specificity was set equal, the specificity was 2.1% higher. Our model obtained three classification accuracy of 75.1%, which was 2.1% higher than the human's average result and 6.6% higher than the senior inspectors' average result. The results showed our model

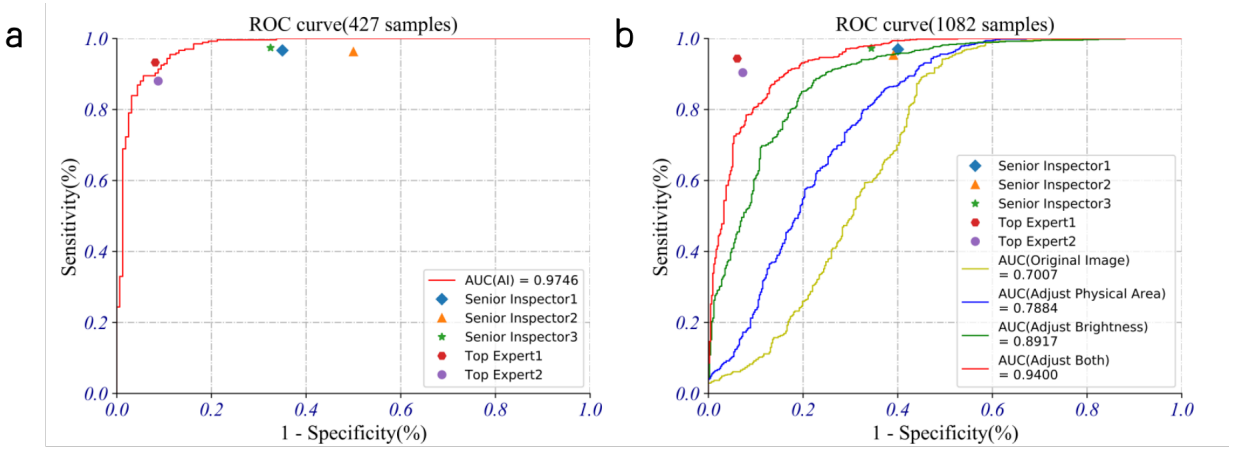

Figure 5: Comparing the results of the best model and five human readers. a)was the performance on the independent test sets ϵ. b) was the performance on the total independent test sets, different ROC curves represented the model's results using different preprocessing methods.

obtained strong generalization ability by using preprocessing.

Table 4
The performance of three edge expansion methods in adjusting physical area step on ξ.

	Original Image	Replicate	Wrap	Reflect
AUC	0.5137	0.6623	0.7038	0.7136

The performance of our models with different training sample size was illustrated in Fig.6. All the AUCs were greater than 0.95. As expected, a larger training set could produce a model with better performance. When the training set had more than 15000 samples, the performance of the model would improved slowly as the amount of sample increased.

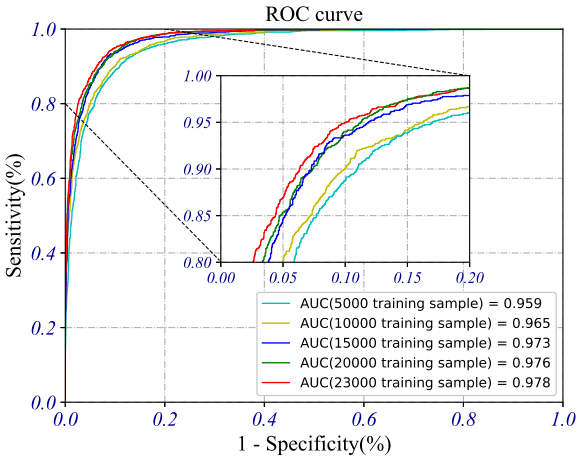

Figure 6: The performance of the best model with different training sample sizes: 5000 images, 10000 images, 15000images, 20000 images, and 23000 images.

4. Discussion

There were lots of methods for diagnosing BV, but gram-stained microscopy was the best method[33]. As Ref [34, 35] showed, clinical criteria for diagnosis of BV only had a sensitivity of 60-72%, other method including OSOM BV Blue and BD MAXTM Vaginal Panel had sensitivities of 91.7% and 90.7% for the diagnosis of BV[36, 37]. Therefore, The Guidelines Group recommended that the current best test to diagnose BV in women was gram-stained microscopy[33].

There were several challenges for manual diagnosis of BV: 1) It was time consuming and labor intensive. Usually it took a skilled inspector about two minutes to identify the morphology and count the number of bacteria. 2) The diagnostic accuracy was still low. The sensitivity was usually below 65% with the specificity was below 80% for general inspectors [13]. 3) The inspector's diagnosis results were subjective and unstable. An inspector may obtain inconsistent results for the diagnosis of the same microscope image at different times. We random selected 110 images to test the diagnosis consistent of two human readers. The human readers diagnosed the same images at intervals of two months. The test results showed than the average consistent diagnosis rate was 89.2%. Traditional automatic diagnosis methods tried to solve these problems, but the effect was limited[13]. By using deep learning techniques, our models could solve these problems well.

Traditional automatic diagnostic methods required three difficult diagnostic steps to get the diagnosis results[13]. The first step involved the segmentation of the infected area, which required a series of artificially designed algorithms to extract the foreground. In the second step, the overlap clumps would be split from the infected area and the individual bacterium morphotypes were obtained. In the third step, the features of the individual bacterium morphotypes were extracted and the bacterium morphotypes were classifed using traditional machine learning methods. In our model, the features of the microscope images could be automatically extracted and the diagnosis was made readily, which avoided complex diagnostic steps in the traditional methods.

The performance of the traditional automatic diagnostic methods was low. For example, the sensitivity, specificity and average accuracy were only 58.3%, 87.1% and 79.1%[13]. Our best model improved all the three diagnostic performance indicators: 33.1% improving for sensitivity, 4.2% increasing for specificity, 1.2% enhancing for average accuracy. Furthermore, our model could simultaneously adjust the sensitivity and specificity by adjusting the predicting probability threshold. The diagnostic performance could not be further improved with the same traditional automatic diagnostic methods, but could be further improved in our model with more training data.

Our model showed strong generalization ability, the performance was overperformed senior inspectors when the samples was standardized by the preprocessing. Standardizing the actual physical area and the brightness of the samples made our model perform very well on samples with large differences. We further studied the impact of the clarity of the samples on the model. The results showed the image sharpening method could not improved the model's performance on the independent test sets, and the performance decreased when the samples became more blurred. The results illustrated the clarity of the samples was good enough for our model.

In addition, we used a NVIDIA GeForce GTX 1080Ti GPU for training and inference. The best model was faster than the NugentNet and 1/2 NugentNet in both training process and inference process. In the training process, the best model could be obtained within 10000 iterations and could be completed in 2.4 hours by one GPU. But it needed several years to train an inspector. In the inference process, our model only needed 2.4 seconds to diagnose 100 images, while it needed more than one hour to diagnose 100 microscope image for a human reader. The traditional automatic diagnostic methods needed 30 seconds to obtain the diagnosis result for a single microscope image[13]. By using the same hardware, our model could diagnose 5 microscope images per second. The inference speed of our model was more than 1500 times faster than manual diagnosis and 150 times faster than traditional automatic diagnostic methods. The diagnostic efficiency of our model was much higher than manual diagnosis and traditional automatic diagnostic methods.

The microscope image contained lots of local details and global information. The local details included various types of bacteria. The global information included the distribution density of various bacteria and the distribution ratio of various bacteria, etc. In the diagnosis process, all the information should be used for a more accurate diagnosis. The details could be extracted by the first few layers of the CNN model and the global information could be extracted by the last few layers. The last fully connected layer could use all the information extracting from the convolution layer in front to obtain the Nugent score. The CNN model could accurate extract details and global information for diagnosis [22, 23, 38]. Therefore, the CNN model was very suitable for BV automatic diagnosis.

In conclusion, this study first used deep learning techniques to diagnose bacterial vaginosis. we constructed the

convolutional neural network models for automatic BV diagnosis. For image-level BV diagnosis, our models had better performance in terms of accuracy, efficiency and stability than experts and traditional automated diagnostic methods. In addition, lots of gynecological lower genital tract infections including aerobic vaginitis (AV), vulvovaginal candidiasis (VVC) and trichomonas vaginitis (TV) were diagnosed by the same microscope images. Our model could be further used to diagnose these three infections. Furthermore, our model could be used for diagnosis of other microscope images for infection diagnosis in other areas. It could be developed into an automatic diagnostic device for inflammatory infections, which would be more precise, more efficient and more stable than manual diagnosis and will standardize diagnostic process.

Data sharing statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding sources

This work was supported by the National Natural Science Foundation of China (Grant No.81671409) and Beijing Municipal Administration of Hospitals Clinical Medicine Development of Special Funding (Grant No. XMLX201605).

Author contributions

Z. Wang, L. Zhang, Z. Liu, Q. Liao and W. Xu proposed the research, Z. Liu, R. An, P. Li, L. Geng, Q. Qiao, W. Zhu and Q. Liao led the multicenter study, Y. Wang, Z. Wang, and L. Qi collected data, Y. Wang, H. Bai, and M. Zhao performed the data annotation, Z. Wang, M. Li and W. Wu wrote the deep learning code and performed the experiment, J. Li, N. Li, C. Rui, C. Fan, X. Liu, Y. Si, L. Qi and A. Feng evaluated the algorithm, Z. Wang, L. Zhang and Q. Zhang wrote the manuscript, M. Wang, W. Mo, Q. Liao and W. Xu reviewed the manuscript.

Declaration of Competing Interest

Authors Z. Wang, W. Mo, W. Wu and M. Li were employed by the company Suzhou Turing Microbial Technologies Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgment

Thanks to Suzhou Turing Microbial Technologies Co., Ltd for technical support.

References

[1] Wang J. Bacterial vaginosis. Prim Care Update Ob Gyns 2000;7:181-5.
[2] Leitich H, Bodner-Adler B, Brunbauer M, Kaider A, Egarter C, Husslein P. Bacterial vaginosis as a risk factor for preterm delivery: a meta-analysis. Am J Obstet Gynecol 2003;189:139-47.
[3] Hillier SL, Krohn MA, Cassen E, Easterling TR, Rabe LK, Eschenbach DA. The role of bacterial vaginosis and vaginal bacteria in amniotic fluid infection in women in preterm labor with intact fetal membranes. Clin Infect Dis 1995;20:Suppl 2:S276-S278.
[4] Peipert JF, Ness RB, Blume J, et al. Clinical predictors of endometritis in women with symptoms and signs of pelvic inflammatory disease. Am J Obstet Gynecol 2001; 184:856-63.
[5] Hillier SL, Kiviat NB, Hawes SE, et al. Role of bacterial vaginosis-associated microorganisms in endometritis. Am J Obstet Gynecol 1996;175:435-41.
[6] Martin HL, Richardson BA, Nyange PM, et al. Vaginal lactobacilli, microbial flora, and risk of human immunodeficiency virus type 1 and sexually transmitted disease acquisition. J Infect Dis 1999;180:1863-8.
[7] Moodley P, Connolly C, Sturm AW. Interrelationships among human immunodeficiency virus type 1 infection, bacterial vaginosis, trichomoniasis, and the presence of yeasts. J Infect Dis 2002;185:69-73.
[8] Klebanoff MA, Hillier SL, Nugent RP, et al. National Institute of Child H, Human Development Maternal-Fetal Medicine Units N (2005) Is bacterial vaginosis a stronger risk factor for preterm birth when it is diagnosed earlier in gestation? Am J Obstet Gynecol 192: 470-477.
[9] Koumans EH, Sternberg M, Bruce C, McQuillan G, Kendrick J, Sutton M, Markowitz LE, The prevalence of bacterial vaginosis in the United States, 2001-2004; Associations With Symptoms, Sexual Behaviors, and Reproductive Health. Sex Transm Dis, 2007; 34 (11), 864-9
[10] Liao QP, Zhang D. Current status and research progress of diagnosis and treatment of female genital tract infections in China. J Int Obstet Gynecol, 2011; 38: 469-474.

[11] Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of Gram stain interpretation. J Clin Microbiol 1991;29:297–301.
[12] Coleman JS, Gaydos CA. Molecular Diagnosis of Bacterial Vaginosis: an Update. Journal of Clinical Microbiology, 2018:JCM.00342-18.
[13] Song Y, He L, Zhou F, et al. Segmentation, Splitting, and Classification of Overlapping Bacteria in Microscope Images for Automatic Bacterial Vaginosis Diagnosis. IEEE journal of biomedical and health informatics, vol. 21, no. 4, July 2017, 1095-1104.
[14] Dimopolous S, Christian EM, Fabian R, and Joerg S. Accurate cell segmentation in microscopy images using membrane patterns. Bioinformatics, vol. 30, no. 18, 2014;2644–2651.
[15] LeCun L, Bottou L, Bengio Y, and Haffner P. Gradient-based learning applied to document recognition. Proceeding of the IEEE, 86(11):1998;2278-2324.
[16] Krizhevsky A, Sutskever I, and Hinton G. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, 2012;1106-1114.
[17] Simonyan K, and Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In CVPR, 2014.
[18] He K, Zhang X, Ren S, and Sun J. Deep residual learning for image recognition. In CVPR, 2016.
[19] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. CVPR 2015:1-9.
[20] Chollet F. Xception: Deep learning with depthwise separable convolutions. In CVPR, 2017.
[21] Long J, Shelhamer E, and Darrell T. Fully convolutional networks for semantic segmentation. In CVPR, 2015.
[22] Zhao H, Shi J, Qi X, Wang Q, and Jia J. Pyramid scene parsing network. In CVPR, 2017.
[23] Ronneberger O, Fisher P, and Brox T. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.
[24] Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402-2410.
[25] Maji D, Santara A, Mitra P, Sheet D. Ensemble of Deep Convolutional Neural Networks for Learning to Detect Retinal Vessels in Fundus Images. arxiv: 1603.04833.
[26] Poplin R , Varadarajan A V , Blumer K , et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning[J]. Nature Biomedical Engineering, 2018.
[27] Shichijo S, et al. Application of Convolutional Neural Networks in the Diagnosis of Helicobacter pylori Infection Based on Endoscopic Images, EBioMedicine, 25 (2017): 106-111.
[28] Smith KP, Kang AD, Kirby JE. Automated Interpretation of Blood Culture Gram Stains by Use of a Deep Convolutional Neural Network, Journal of Clinical Microbiology. 2018: 56:e01521-17.
[29] Lori D, Racsa, Rita M, Gander, Paul M, Southern, et al. Detection of intracellular parasites by use of the CellaVision DM96 analyzer during routine screening of peripheral blood smears[J]. journal of clinical microbiology, 2015, 53(1):167-171.
[30] Sutskever I, Matrens J, Dahl G, and Hinton G. On the importance of initialization and momentum in deep learning. In ICML, 2013.
[31] Detector Performance Analysis Using ROC Curves - MATLAB & Simulink Example. www.mathworks.com. Retrieved 11 August 2016.
[32] Powers DMW. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies. 2008;2(1):37–63.
[33] Sherrard J, Wilson J, Donders G, Mendling W, and Jensen JS. 2018 European (IUSTI/WHO) International Union against sexually transmitted infections (IUSTI) World Health Organisation (WHO) guideline on the management of vaginal discharge. International Journal of STD & AIDS, 2018;0(0):1–15.
[34] Gallo MF, Jamieson DJ, Cu US, et al. Accuracy of clinical diagnosis of bacterial vaginosis by human immunodeficiency virus infection status. Sex Transm Dis, 2011;38:270–274.
[35] Singh RH, Zenilman JM, Brown KM, et al. The role of physical examination in diagnosing common causes of vaginitis: a prospective study. Sex Transm Infect, 2013;89:185–190.
[36] Paavonen J, and Brunham RC. Bacterial Vaginosis and Desquamative Inflammatory Vaginitis. N Engl J Med, 2018;379:2246-2254.
[37] Gaydos CA, Begaj S, Schwebke J, et al. Clinical valida- tion of a test for the diagnosis of vaginitis. Obstet Gynecol, 2017;130:181–189.
[38] Goodfellow I, Bengio Y, and Courville A. Deep Learning. MIT Press. 2016;pages 316-356.