IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 3, 2021, accepted April 19, 2021, date of publication May 3, 2021, date of current version May 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076604

Research on Multiple-Instance Learning for Tongue Coating Classification

YONGHUI TANG[1], YUE SUN[1], JOHN Y. CHIANG[2], AND XIAOQIANG LI[1,3], (Member, IEEE)

[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[2]Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
[3]Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China

Corresponding author: Xiaoqiang Li (xqli@shu.edu.cn)

This work was supported in part by the Shanghai Innovation Action Plan Project under Grant 16511101200.

ABSTRACT Tongue coating can provide valuable diagnostic information to reveal the disorder of the internal body. However, tongue coating classification has long been a challenging task in Traditional Chinese Medicine (TCM) due to the fact that tongue coatings are polymorphous, different tongue coatings have different colors, shapes, textures and locations. Most existing analyses utilize handcrafted features extracted from a fixed location, which may lead to inconsistent performance when the size or location of the tongue coating region varies. To solve this problem, this paper proposes a novel paradigm by employing artificial intelligence to feature extraction and classification of tongue coating. It begins with exploiting prior knowledge of rotten-greasy tongue coating to obtain suspected tongue coating patches. Based on the resulting patches, tongue coating features extracted by Convolutional Neural Network (CNN) are used instead of handcrafted features. Moreover, a multiple-instance Support Vector Machine (MI-SVM) which can circumvent the uncertain location problem is applied to tongue coating classification. Experimental results demonstrate that the proposed method outperforms state-of-the-art tongue coating classification methods.

INDEX TERMS Tongue coating classification, multiple-instance learning, deep features.

I. INTRODUCTION

Tongue diagnosis is an effective treatment in Traditional Chinese Medicine (TCM). The tongue is rich in geometric features and texture features, which are closely linked to the physiological information of human organs. For a long time, tongue diagnosis mainly relies on TCM practitioner' clinical experience to make visual judgment and analysis. Obviously, it is imperative to utilize the objective and automatic identification technology to tongue diagnosis. In recent years, a large number of researches based on tongue images have emerged in the field of artificial intelligence.

Tongue image classification is an important task in tongue diagnosis. It is used to identify the type of tongue image to help TCM physicians make further diagnostic decisions. For this type of automatic diagnosis, a feature extractor is usually used to extract the features of tongue, and then a classifier is utilized to do the final classification. The tongue can be classified by examining its color, shape, texture and other features. For example, Hou *et al*. [1] performed tongue color classification by modified CaffeNet [2]. Tooth-marked tongue is identified according to whether there are tooth marks on the edge of the tongue. Li *et al*. [3] fused a Convolutional Neural Network (CNN) variant into the recognition of tooth-marked tongue. Zhang and Zhang [4] examined tongue shape and its relationship to patient status. Thirteen geometry features including measurements, distances, areas, and their ratios were extracted from each tongue image and classified using a decision tree. Tang *et al*. [5] used a cascaded CNN to detect the tongue region and tongue landmarks. Meanwhile, a fine-grained classification network was utilized to recognize the tooth-marked tongue.

Tongue image alignment is a fundamental issue of tongue diagnosis, which is the mapping of points or subregions among different tongue images. Wu *et al*. [6] presented a conformal mapping method for tongue image alignment. And Dai and Wang [7] proposed a method called the conceptual alignment deep autoencoder to analyze tongue images that

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

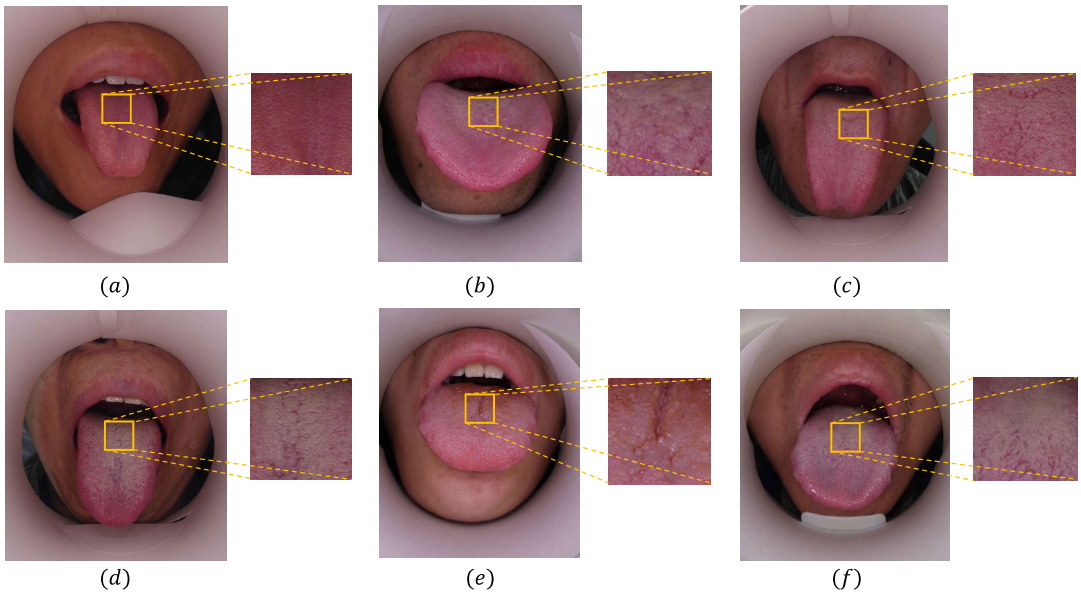

FIGURE 1. Different tongue coatings. (a) Normal tongue coating. (b-f) Typical rotten-greasy tongue coatings.

represent different body constitution types. These methods provided an efficient and accurate tool for deformable medical image alignment and disease diagnosis.

It is appealing that the tongue can be used for biometric recognition because of its rich individual characteristics. Yang *et al*. [8] exploit appearance manifold learning method to represent the dynamic shape changes of the tongue for identity verification.

From the TCM perspective, tongue coatings are closely related to people's health. Tongue coatings correspond to the floating layers of the tongue, whose colors, shapes and textures can reflect the internal state of the body and organs [9]. This paper focuses on how to distinguish rotten-greasy tongue coating from normal tongue coating based on tongue image. Rotten-greasy tongue coating is thick and loose, resembles residues of bean curd and always locates in the middle and rear of the tongue body [10]. Normal tongue coating is usually thin and pale white. Fig. 1 shows normal and different rotten-greasy tongue coatings, (a) is a normal tongue coating image and (b-f) correspond to typical rotten-greasy tongue coating images.

Tongue diagnosis is a kind of inspection. TCM practitioners observe tongue of the patients and make a decision according to its color, shape, and texture. To the best of our knowledge, this is the first time that "TCM vision concept," proposed in this paper, is applied to the simulation of TCM inspection diagnosis. These concepts are only implicitly meaningful to TCM practitioner, but not to others. In fact, the rotten-greasy coating is a unique vision concept in TCM. The classification of tongue coatings is a challenging task for three reasons. First, the classification of tongue coatings into different subclasses can be viewed as a fine-grained classification problem since normal and rotten-greasy tongue coatings are only different symptom of the floating layer of the tongue. Fine-grained classification refers to the task of differentiating objects that belong to the same base class [11]. It demands a powerful algorithm to discriminate among object classes with a high degree of similarity that are often differentiated by only subtle differences such as different species of birds or dogs. Second, there lacks further information (such as the location or size of the tongue coating patch) because a tongue image is always either labeled as normal coating tongue or rotten-greasy coating tongue. Third, in a rotten-greasy coating tongue, normally, the area covered by rotten-greasy coating may account for only a small proportion, while the normal coating area constitutes a larger proportion. Therefore, there is too much noise in the rotten-greasy coating tongue, which affects the classification accuracy.

In this paper, we try to solve these aforementioned problems by multiple-instance learning (MIL) [12] and deep learning.

MIL is first proposed by Dietterich *et al*. [12]. In the MIL method, a classifier is learned based on a training set of bags, where each bag contains multiple feature vectors (called instances in the MIL terminology) [13]. As described in [14], the classification of tongue coatings maps naturally to a multiple-instance problem. Along this line of thought, only coarsely labeled images are required to train a MIL model. In the medical image classification task, Support Vector Machine (SVM) is a common tool for high dimensional feature classification. Lin *et al.* [15] used SVM for diabetic tongue image recognition. And Wan *et al*. [16] proposed using SVM for the identification of fissured tongue. In our paper, we embed MIL into SVM as a classifier. On the other hand, the quality of feature extraction, an important step of

tongue coating classification, directly determines the final classification performance. Inspired by the success of CNN, a fine-tuned CNN, instead of handcrafted feature extraction, is utilized to extract deep features of the tongue coating patches. The contributions of this paper can be summarized as follows.

- MIL is introduced into the tongue coating classification task to alleviate the coarsely labeling problem of tongue coating images
- We propose a method to find important patches related to the rotten-greasy information in a tongue body.
- Extensive experiments is conducted to select appreciate patch size and CNN [17] structure, the results demonstrate effectiveness of our method.

The remaining of this paper is organized as follows. In section. II, we review methods of tongue coating classification in recent years. In section. III, the proposed method is elaborated. Experimental results are presented in section. IV. Finally, we make a conclusion and discuss the future work in section. V. Our implementation has been released[1] to facilitate further developments on MIL based tongue coating classification.

II. RELATED WORK

TCM tongue diagnosis relies heavily on the clinical experience of TCM practitioners. Nowadays, there have been some researches combining modern technology to tongue coating classification, which mainly uses machine learning methods to objectively analyze the distribution of tongue images. We will introduce them in this section and the corresponding experimental results are shown in Table 6.

Wei *et al.* [18] used the improved subspace method to analyze the texture density of tongue coating. The tongue body was divided into fixed-size blocks, and each block was classified. Based on the classification results, the rotten-greasy index and description of the entire tongue image were given. When classifying tongue coating blocks, an improved subspace method was used, and the projection length ratio was used as the classification distinguishing feature to analyze the density of texture structure.

Zhang *et al.* [19] extracted four feature vectors: contrast (CON), angular second moment (ASM), entropy (ENT) and correlation (COR) of the gray-level co-occurrence matrix (GLCM) in tongue coatings to determine the features of rotten-greasy tongue coatings.

Qu *et al*. [20] proposed a tongue coating classification method based on Gabor wavelet transform. First, the whole tongue image was transformed by Gabor wavelet. After weaking the edge of the tongue body, the mean value and standard deviation were extracted as the texture features to recognize the rotten-greasy tongue coatings.

Choraś *et al*. [21] used a bank of filters built from the real part of Gabor expression, named even symmetric Gabor filter. By selecting different center frequencies and orientations, a family of Gabor kernels was formed to extract features from a tongue image. Li *et al*. [22] extracted the center patch of a tongue body and classified tongue coating using Gabor and Tamura features of a patch.

Deep learning has good performance in many computer vision tasks, such as image classification, object detection, semantic segmentation and so on. Deep learning method has a strong self-learning ability, which can obtain an implicit expression of image rules through repeated learning. Therefore, deep learning method has unique advantages in extracting features and classifying images. Fu *et al*. [23], combining basic image processing with deep learning, derived tongue coating features through deep neural networks. Zhang *et al*. [24] realized automatic classification of tongue texture color and tongue coating color through a neural network. Yang and Zhang [25] proposed a tongue image classification method based on transfer learning and fully connected neural network, which can effectively improve the accuracy of tongue image classification.

The methods mentioned above, however, have some drawbacks. Firstly, handcrafted features used in the methods of Wei *et al.* [18], Zhang *et al.* [19], Qu *et al.* [20], Choraś *et al.* [21] and Li *et al.* [22] cannot describe the salient characteristic of tongue coating. Secondly, although the methods of Fu *et al.* [23], Zhang *et al.* [24] and Yang and Zhang [25] based on deep neural networks can extract deep features, it only focuses on global information rather than local one, which may adversely affect the classification result due to more irrelevant information captured.

III. METHOD

As shown in Fig. 2, the proposed method is composed of three stages. First, rotten-greasy tongue coating information is utilized to select suspected tongue coating patches. Then, a CNN is used to extract fixed-length feature vectors for each tongue coating patch. At last, feature vectors belong to one tongue image are grouped into a bag to obtain an arranged feature vector, and the label of the bag is the label of the original image. Then, a multiple-instance Support Vector Machine (MI-SVM) [26] is used to perform the classification. The details will be described in this section.

A. OBTAINING SUSPECTED ROTTEN-GREASY COATING PATCHES

The goal of this step is to obtain suspected rotten-greasy coating patches for training the MI-SVM. The suspected rotten-greasy coating patches are not clearly labeled, but there exists at least one tongue coating patch with rotten-greasy information in a rotten-greasy coating tongue. Therefore, we manually select patches in the area with a high probability of rotten-greasy information. According to the TCM perspective and our observation (as shown in Fig. 4), the rotten-greasy coating always appears in the middle and rear of a tongue body, while the rest of the tongue can be ignored.

[1] https://github.com/Hazel-4/litangsun

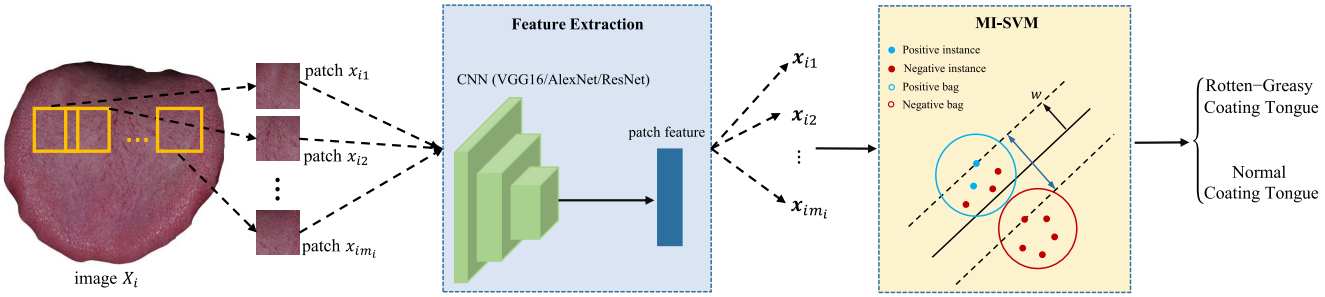

FIGURE 2. The system diagram of the proposed method. Left: Patch selection. Middle: Feature extraction. Right: A multiple-instance SVM is trained to classify the tongue.

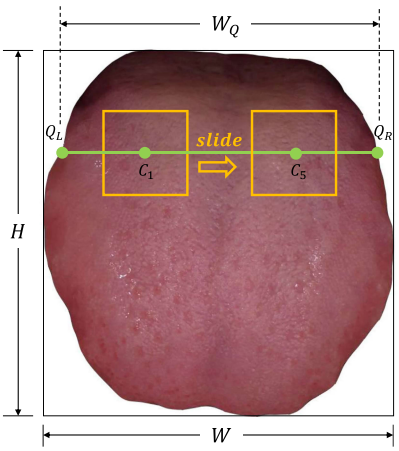

FIGURE 3. The diagram illustrating the process of patch selection.

Thus, these patches selected from a rotten-greasy coating tongue can satisfy the assumption of multiple-instance binary classification that there exists at least one positive instance in a positive bag. Patches obtained from a healthy coating tongue include only healthy ones.

The steps of obtaining the patches proceed as follows.

Step 1: The rectangle circumscribing the tongue is denoted as R. The height and width of the rectangle are denoted as H and W, respectively.

Step 2: Draw a horizontal line $\frac{H}{3}$ from the top of the tongue, denoted as Q. Use Q_L, Q_R to denote its left and right intersection point with the edge of the tongue. The width of the intersection line is denoted as W_Q.

Step 3: Find C_1 on line Q. C_1 is located $\frac{W_Q}{3}$ to the right of Q_L. Take C_1 as the starting point and $\frac{W_Q}{3n}$ the step length, find C_i $(i = 2, 3, \ldots, n+1)$ rightwards. For each C_i, draw a square with C_i as its center and $\frac{W_Q}{6}$ as its side length. The squares represent the selected patches.

As shown in Fig. 3, by changing the side length and the step length, we can obtain tongue coating patches of different sizes and quantities. The experiment to select the optimal sizes and quantities of tongue coating patches is shown in section. IV-C.

B. CONVOLUTIONAL NEURAL NETWORK

In our method, robust texture features are needed to describe tongue coating patches. Motivated by the characteristics of CNN which can naturally integrate low/mid/high level features, we use a CNN to extract fixed-length feature vectors of the tongue coating patches instead of traditional handcrafted methods.

AlexNet [27], VGG16 [28], and ResNet [29] are three popular CNN models in recent years, and we tried all three models in the experiment. Details of the comparative experiments of the three models will be described in section. IV-D1. In Table 1, we compared the characteristics of the three networks in terms of structure and calculation speed. The symbol (+) indicates the method is used and the symbol (−) indicates the method is not used. In the table, GFLOPS (giga floating-point operations per second) represents the computation speed of the model and FC represents the fully connected layers.

TABLE 1. System layouts of three different of CNN models.

Comparison	AlexNet	VGG16	ResNet
Data Augmentation	+	+	+
Number of Layers	8	16	50
Size of Kernel	11,5,3	3	7,1,3,5
Number of FC	3	3	1
Size of FC	4096,1000	4096,1000	1000
Dropout	+	+	+
Local Normalization	+	-	+
Batch Normalization	-	-	+
GFLOPS	7.2	19.6	3.8

Hosny *et al.* [30] utilized transfer learning with pretrained AlexNet, a highly accurate method, for skin lesion classification. AlexNet has the smallest layer numbers and the fastest operating speed. The activation function of AlexNet adopted the Rectified Linear Units (ReLU) instead of the traditional sigmoid function and the tanh function, which can solve the gradient dispersion problem in the deep network. The dropout technology was used to avoid model overfitting. For AlexNet, its first five layers are convolution layers and the remaining three layers are fully connected layers. The local response normalization (LRN) layer following the first and the second convolution layers fixes the means and variances of layer inputs to ensure that the distribution of each batch is close

to the true distribution [31]. Pooling layers follow the first two LRN layers and the fifth convolution layer. The ReLU non-linearity is applied to the output of each convolution layer and fully connected layer [27].

VGG16 is proposed by the Visual Geometry Group at Oxford University. VGG16 performs well in transfer learning task. So far, this model is still used to extract image features. For example, [32] shown that VGG16, as a transfer learning model based on convolutional neural network, can be used to construct an effective extractor of abdominal ultrasound images. VGG16 improves AlexNet by replacing large-size convolutional kernels (11×11 and 5×5 kernels in the first and second convolution layer, respectively) with multiple 3×3 convolutional kernels one after another. With a given receptive field (the effective area size of input image on which output depends), multiple stacked smaller size kernels are preferred than one with a larger size kernel. Although multiple non-linear layers increase the depth of the network, it enables the network to learn more complex features at a lower cost.

ResNet, with 50 layers, is the deepest among the three models. It consists mostly of 3×3 kernels, which is similar to VGG16. Each convolution layer is followed by a batch normalization layer and a ReLU activation function to alleviate vanishing gradient. Shortcut connection is inserted to VGG16 to form a residual network. Thus, ResNet with a depth of up to 50 layers still possesses lower complexity and the degradation problem can be well addressed. Residual networks are characterized by their ease of optimization and their ability to increase accuracy by adding considerable depth. Wang *et al*. [33] proposed an artificial intelligence framework using ResNet for the recognition of tooth-marked tongue. The model had good effectiveness and the overall accuracy was over 90%. Therefore, we also used a typical ResNet architecture consisting of 50 layers to classify the tongue coating images in the present study.

C. FEATURE EXTRACTION

In this stage, we use a CNN to extract fixed-length feature vectors of tongue coating patches instead of using the whole tongue image. In this paper, a fine-tuned ResNet is applied as the feature extractor and the fine-tune experimental details are delineated in section. IV-B2. The ResNet model has 50 weight layers, 49 of which are convolutional layers and the remaining 1 is a fully connected layer. There are a total of 2048 units in the last pooling layer and the outputs of this layer are used as features.

The tongue coating patches obtained according to the method described in section. III-A are used as inputs, and the network outputs a 2048-deimension vector. Thus, a 2048-dimension feature vector for every suspected tongue coating patch is obtained.

D. CLASSIFICATION

In this stage, a MI-SVM is trained to classify the tongue images. In the MIL task, a classifier based on a training set of bags, where each bag contains multiple feature vectors, is learned [13]. The main idea of MI-SVM is to maximize bag margin which serves as an extension of the instance margin of standard SVM, and the details are thoroughly introduced in [26].

In our method, a tongue image is represented as a bag and a patch are represented as a instance in the bag. we denote $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ as a set of bags, the *ith* bag $\boldsymbol{X}_i = \left\{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{im_i}\right\}$, where N and m_i denote the number of bags and number of instances in bag $\boldsymbol{X}_i$ respectively, and $\boldsymbol{x}_{ij}$ denotes the *jth* instance of the *ith* bag. A binary lable Y_i is associated with the bag $\boldsymbol{X}_i$, while the instance lable y_{ij} in the bag is ambiguous. $Y_i = -1$ indicates a negative bag (rotten-greasy tongue coating image) where all instance are negative. And $Y_i = +1$ represents a positive bag (normal tongue coating image) containing at least one positive instance. In MI-SVM, the function margin of a bag is defined as:

$$\gamma_i = Y_i \max_{1 \le j \le m_i} \left(\left\langle \boldsymbol{\omega}, \boldsymbol{x}_{ij} \right\rangle + b \right). \tag{1}$$

Parameters $\boldsymbol{\omega}$ and b are the normal vectors and intercepts of the hyperplane respectively.

The MI-SVM aims at maximizing the bag margin, which is defined as follows:

$$\begin{aligned} &\min_{\boldsymbol{\omega}, b, \xi} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C \sum_{i=1}^{N} \xi_i, \\ &\text{s.t. } \forall i : Y_i \max_{1 \le i \le m_i} \left(\langle \boldsymbol{\omega}, \boldsymbol{x}_i \rangle + b \right) \ge 1 - \xi_i, \quad \xi_i \ge 0. \end{aligned} \tag{2}$$

ξ is the relaxation factor and C is the penalty parameter. When the value of C is large, the penalty for misclassification increases, and when the value of C is small, the penalty for misclassification decreases. In MI-SVM, the bag margin is determined by only one of its instance. For a positive bag, the margin is decided by the most positive instance, while the margin of a negative bag is up to the least negative instance [34]. The label of the bag is then the label of the image.

IV. EXPERIMENT AND DISCUSSION

In this section, our tongue coating classification experiment are introduced. The limitations and requirements of dataset collection is discussed in section. IV-A. The pre-trained ResNet is fine-tuned to learn deep features of tongue coating patches (section. IV-B). The selection of suspected rotten-greasy coating patches is discussed in section. IV-C. We conduct some comparative experiments, including comparing the performance of different feature extractors, different classifiers and the proposed method with other methods (section. IV-D, section. IV-E and section. IV-F, respectively).

A. DATASET COLLECTION

Data collection is challenging for tongue coating classification. Two problems often arise when collecting data. The first problem is the standardization of illumination and imaging camera. Different light sources and environments always

affect illumination, while the lens and the Charge Coupled Device (CCD) of the imaging camera will influence the captured picture [9]. The second problem involves privacy, when we intend to use patients' tongue images as dataset for the study of tongue diagnosis, their consents are required. As a result, the number of tongue images we can collect is limited.

Fortunately, our dataset has no problem with the effects of illumination, as all of the tongue images are captured under a standard light source. In addition, the images are collected clinically by DS01-B Information Collection System of Tongue and Face Diagnosis, professional equipment provided by Shanghai Daosheng Medical Technology Co., Ltd.

In many pattern recognition and matching problems, for example the identification and verification of identities, the annotation of single sample is straightforward and objective. However, in TCM, since some symptoms are not visually apparent and the texture of the rotten-greasy tongue coating is not very obvious, there may be noise labels in the tongue images. To solve the problem above, the label of a tongue image is proceeded voted by five TCM practitioners and only tongue images with a judgment rate of 80% or above were accepted into the sample set. The dataset used in our paper includes 274 samples of tongue images, 186 of them are normal tongue coating images and 88 are rotten-greasy ones. On the other hand, in order to better analyze the tongue images, we used Photoshop to extract the tongue body part from the original image.

B. TRAINING ON CNN

We hope that by training the ResNet network, it will be able to identify the characteristics of different tongue coatings. The features of suspected rotten-greasy coating patches can be extracted by the fine-tuning ResNet model. And the feature extraction method is as described in section. III-C.

1) DEEP FEATURES OF ROTTEN-GREASY COATING PATCH

CNN can be a powerful feature extractor. We hope that CNN can effectively extract deep features combining color, shape and texture information to describe tongue coating patches. Therefore, when training a CNN, we manually label patches with salient features in each tongue image as input.

In order to obtain deep features of rotten-greasy coating patch, the method of obtaining tongue coating patches is described as follows. The tongue coating patches that contain rotten-greasy features are generated using the bounding boxes on rotten-greasy tongue coating images provided by TCM practitioners. For normal tongue coating images, patches are chosen in the area with normal coating characteristics. And for each tongue image, 10-15 patches are obtained and each patch is about 180-300 pixels wide and 240-400 pixels high. Unlike the suspected rotten-greasy tongue coating patches, these patches for training a CNN are clearly labeled.

2) TRAINING

In our paper, we use the ResNet described in [29] as a feature extractor. Since tongue coating classification is a binary classification problem, we drop the last 1000-way fully connected layer and replace it with a 2-way fully connected layer during the network fine-tuning.

The network is first pretrained on ILSVRC [35] dataset and then followed by fine-tuning on tongue coating patches. All tongue coating patches are obtained using the method shown above. These patches are only used for fine-tuning the network. There are 3333 tongue coating patches in total, among which 1183 are rotten-greasy coating patches and 2150 are normal coating patches, which are, however, not enough to train such a high-capacity network. The network would fail to converge if it is not pretrained.

Size of tongue coating patches is variable, while our system demands a constant input dimension. Therefore, for the input of neural network models (AlexNet, VGG16 and ResNet), we sample the patches to a fixed size of 224×224. Given a rectangular patch, we first scale the patch so that the length of the short side is 256, and then crop out the central 256×256 patch from the resulting patch. At last a fixed-size 224×224 sub-patch is randomly cropped from the 256×256 patch to be trained on the ResNet.

During training, we took seven tenths of the total number of tongue coating patches as the training set, and the remaining three tenths as the test set, and kept the balance between the number of rotten-greasy coating patches and normal coating patches in the training set and test set. We use stochastic gradient descent to fine-tune the network with a batch size of 32 and a learning rate of 0.00001. We stop the training after 30 epoches since the accuracy ceases increasing. Thus, the fine-tuned ResNet has the ability to extract deep features of different tongue coatings.

C. DISCUSSION ON THE SELECTION OF SUSPECTED ROTTEN-GREASY COATING PATCHES

In this section, we introduce the method of choosing the reasonable position, size and quantity of the suspected rotten-greasy coating patches in each tongue body for the classification stage. All the tongue images were segmented manually from the background for better analysis. According to the TCM perspective and our observation, rotten-greasy coatings always appear in the middle and rear of a tongue body. Our experiments will subsequently lead to an objective conclusion about this.

Our dataset includes 274 tongue coating images. We perform fivefold cross-validation on this dataset to prove the effectiveness of the proposed method. The details of the dataset are described in section. IV-A. The dataset is randomly shuffled and partitioned into five subsets and the balance between the amount of rotten-greasy tongue coating images and normal tongue coating images is maintained in each subsets. Each time, four subsets are selected for training and the remaining one for testing. It should be noted that the samples of MI-SVM training are suspected tongue coating patches selected using the method described in section. III-A.

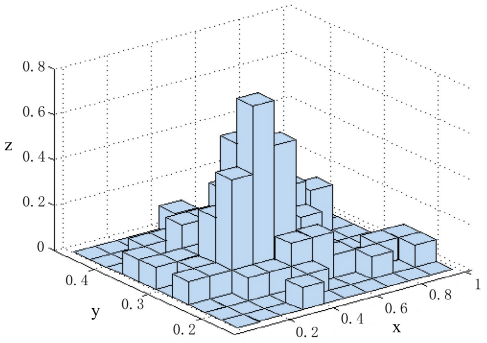

FIGURE 4. Statistical histograms of the position, height and width of the rotten-greasy tongue coating in the tongue body.

Firstly, for all rotten-greasy tongue coating images, distribution of tongue coating areas is counted and the location information is integrated. Statistical histogram of the position, height and width of the rotten-greasy tongue coatings in the rotten-greasy tongue coating body is shown in Fig. 4. If a position Q on the tongue body is determined, we denote A as the vertical distance from Q to the root of the tongue and B as the horizontal distance from Q to the left boundary of the tongue. If the height of the tongue body is H and the width is W, then the value of X-axis is $A \times H$ and the value of Y-axis is $B \times W$. The Z-axis is the probability of having a rotten-greasy tongue coating at this position. For each rotten-greasy tongue image, the rotten-greasy tongue coating is counted according to whether the position Q contains rotten-greasy tongue coating. Let M denotes the number of the rotten-greasy tongue images and m denotes the number of the rotten-greasy tongue coating at this position, then the probability of the occurrence of rotten-greasy tongue coating at the position is defined as $z = \frac{m}{M}$. This statistical result verifies our pervious claim about the position of the rotten-greasy tongue coatings.

Secondly, in order to determine the optimal size and quantity of the suspected rotten-greasy coating patches in each tongue body for the classification stage, we conduct a series of experiments which use different sizes and quantities of tongue coating patches in each tongue image. As mentioned above, we select patches in the middle and rear of the tongue. All experiments are based on the proposed methods, in which the suspected rotten-greasy tongue coating patches are obtained using the method detailed in section. III-A, ResNet is used as a feature extractor and MI-SVM is utilized to perform the final classification. Experimental results in Table 2 show that our method has the highest accuracy when the size of the suspected tongue coating patches is 300×300 and the quantity is 5 in each tongue images. Therefore, in the following experiments, the suspected tongue coating patches are all obtained from the middle and rear of the tongue body, the quantity of patches of each tongue image is 5, and the size is 300×300.

D. COMPARISON BETWEEN DIFFERENT FEATURES

In this section, we conduct experiments on the dataset described in section. IV-A using the proposed method and compare the performance between ResNet and other feature extractors in extracting the features of tongue coating patches. The same MI-SVM classifier is used to test the performance of different feature extraction methods. And fivefold cross-validation are performed on the dataset for all the experiments.

TABLE 2. Comparison of eight experiments with different patch quantities and sizes.

quantity	250×250	300×300	350×350	400×400
3	78.45%	79.13%	80.27%	81.76%
4	80.32%	79.04%	80.65%	81.61%
5	**83.76%**	**85.0%**	82.67%	81.00%
6	82.16%	82.01%	82.87%	81.39%
7	81.37%	83.27%	82.16%	83.18%
8	81.85%	83.18%	82.68%	83.13%
9	81.56%	83.32%	**84.23%**	83.43%
10	82.24%	83.37%	83.35%	**83.63%**

Experimental results are evaluated by the following three metrics: 1) accuracy (ACC); 2) true positive rate (TPR); 3) true negative rate (TNR).

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

$$TPR = \frac{TP}{TP + FN}. \tag{4}$$

$$TNR = \frac{TN}{TN + FP}. \tag{5}$$

True Positive (TP) and False Negative (FN) are samples which are positive and predicted as positive or negative, respectively. On the other hand, False Positive (FP) and True Negative (TN) are samples which are negative and predicted to be positive or negative, respectively.

1) COMPARISON BETWEEN DIFFERENT CNN FEATURES

CNN is mainly used to extract features of images based on the idea of shared weights and good nonlinear learning ability. For the task of tongue coating classification in this paper, in order to show the rationality of selecting ResNet as feature extractor, AlexNet and VGG are applied as two representative deep learning methods, and are compared in performing tongue coating classification. In which VGG uses the 16-layer model and ResNet the 50-layer model. The structural details of ResNet, AlexNet and VGG16 are shown in section. III-B. The method of MI-SVM is used to test the performance of these CNN extractors.

We extracted features from ResNet, AlexNet and VGG16 using the method described in section. III-C. Similar to ResNet, AlexNet and VGG16 are pretrained on ILSVRC dataset. In the training stage, the last layers of both AlexNet and VGG16 are modified to be a 2-way fully connected layer because tongue coating classification is a binary classification problem. And they are also fine-tuning on the same tongue coating patches described in section. IV-B1. In the feature extraction stage, the outputs of the second fully connected layer for both AlexNet and VGG16 are used as features.

Compared with AlexNet, VGG16 is deeper. To ensure the same perception as AlexNet and reduce network parameters, three 3 × 3 convolutional kernels are used instead of 7 × 7 convolutional kernels in VGG16. But the performance is not as good as expected in our dataset, as shown in Fig. 5. This cannot be interpreted as overfiting, because overfiting networks perform better in the training sets. The reason for the poor performance of VGG16 is that the gradient vanishes with the increased depth. Meanwhile, it is necessary to optimize more spatial parameters when training deeper networks, which will result in higher training errors.

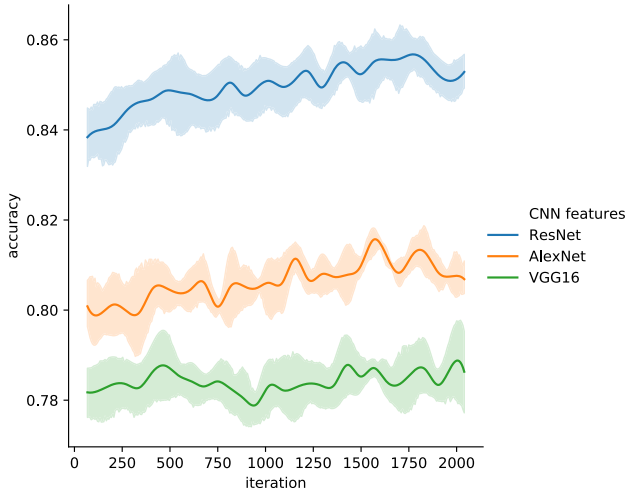

FIGURE 5. The accuracies of ResNet, AlexNet, and VGG16.

ResNet, however, performs better than AlexNet and VGG16. Although ResNet is the deepest, thanks to its residual module, its accuracy and computational efficiency are the highest. At the same time, ResNet uses 3 × 3 convolutional kernels instead of large-size kernels. Consequently, the network learns more detailed features under the same receptive field and reduces the amount of parameters. Thus, in the following experiments, ResNet is employed as a feature extractor.

2) COMPARISON WITH HANDCRAFTED FEATURES

Handcrafted feature extractions are utilized to extract features for each suspected rotten-greasy coating patch. 30 features are extracted for GLDM [36], 6 features for Tamura *et al.* [37], 80 features for Gabor [38], and 20 features for Subspace [39]. These features are also grouped into bags to train MI-SVM. Experimental results in Table 3 show that deep features perform better than handcrafted features such as the features extracted by GLDM, Tamura, Gabor and Subspace. It demonstrates that ResNet can effectively extract useful feature information to describe tongue coating patches.

E. COMPARISON WITH OTHER CLASSIFIERS

Different classifiers are evaluated for tongue coating classification using the features extracted by the fine-tuned ResNet. The categories of tongue coating can be predicted by the aggregation method [3] used by other classification algorithms. The aggregation method also satisfies the multiple-instance learning assumption that if a bag is positive then at least one instance of this bag is positive [3]. Suppose there are m instances in a bag, and p_i is the prediction result of the ith instance in the bag, then the test accuracy of the tongue represented by this bag is as follows.

$$P_B = \max\{p_0, \ldots, p_m\}. \tag{6}$$

Firstly, we compared the experimental results between using MI-SVM with CNN features and using CNN directly. As shown in Table 4, the results demonstrate that our method achieves an accuracy of 85.0% and a recall rate (TPR) of 89.8% which are 11% and 6% higher respectively than those of using CNN directly.

TABLE 3. Comparison between different features.

Method	Accuracy
MI-SVM with GLDM features	54.0% ± 0.034
MI-SVM with Tamura features	60.2% ± 0.042
MI-SVM with Gabor features	68.5% ± 0.037
MI-SVM with Subspace features	55.7% ± 0.045
MI-SVM with ResNet features	**85.0% ± 0.037**

TABLE 4. Comparison between using MI-SVM with CNN and using CNN directly.

Method	Accuracy	TPR	TNR
ResNet [27]	74.2% ± 0.045	83.1% ± 0.041	70.0% ± 0.046
ResNet+MI-SVM (Ours)	**85.0% ± 0.037**	**89.8% ± 0.031**	**82.8% ± 0.043**

The performance of other classifiers with the same features extracted using ResNet model is shown in Table 5. It can be seen that the accuracy of the proposed method is superior to that of Decision Tree [40], Random Forest [40], K-nearest neighbor (KNN) [41] and EMDD [42] – a multiple instance learning algorithm combines expectation maximization (EM) with the diverse density (DD).

TABLE 5. Comparison with other classifiers.

Method	Accuracy
Decision Tree with ResNet features	61.8% ± 0.044
Random Forest with ResNet features	67.3% ± 0.036
KNN (k=5) with ResNet features	67.7% ± 0.040
EMDD with ResNet features	69.2% ± 0.043
MI-SVM with ResNet features	**85.0% ± 0.037**

The reasons why MI-SVM performs better are summarized as follows. 1) For fine-grained classification problem, MI-SVM can effectively exclude irrelevant information. 2) The exact location of the rotten-greasy tongue coating in the tongue body is irrelevant.

F. COMPARISON WITH OTHER METHODS

We conduct experiments on the dataset described in section. IV-A using the proposed method and three other methods. The three methods are: Li's work [22], Qu's

TABLE 6. Comparison with other methods.

Method	Accuracy
Li's [22]	75.6% ± 0.052
Qu's [20]	67.9% ± 0.044
Fu's [23]	58.3% ± 0.035
ResNet+MI-SVM (Ours)	**85.0% ± 0.037**

work [20] and Fu's work [23]. Li obtained the center patches of the tongue images and classified rotten-greasy tongue coatings using Gabor and Tamura features. Without knowing the exact location of tongue coating, this method may lead to unstable performance. Qu proposed a tongue coating recognition method using Gabor wavelet transform on the whole tongue image. It inevitably captures more irrelevant information, deteriorating the recognition results. Fu *et al*. trained a CNN to extract features of the tongue images and used the softmax layer of the CNN directly to classify the tongues. Although Fu's method can extract deep features for tongue coatings, feature extraction was still conducted on the whole image. The results of the above experiments are listed in Table 6. It can be concluded from the table that our approach outperforms all other methods and has the highest accuracy.

V. CONCLUSION

In this paper, we have presented a new method for tongue coating classification using MIL and deep features. The method is divided into three stages. First, suspected rotten-greasy tongue coating patches are selected. Then, a deep CNN is used to extract features of each patch. At last, tongue coating is represented by a bag consisting of multiple feature vectors and MI-SVM is used to perform the final classification. Experiment results demonstrated that the proposed method outperforms previous methods. Future work includes four aspects: 1) Collecting more tongue samples. Since we use a deep CNN as feature extractor, the proposed model always benefits from a larger dataset. 2) Adopting more advanced network architecture to further improve the accuracy. 3) Some fine-grained classification models can be used as feature extractors, because the classification of the tongue coating is a naturally classification task. 4) The multiple-instance learning method can be embedded into the neural network to realize the end-to-end classification model.

REFERENCES

[1] J. Hou, H.-Y. Su, B. Yan, H. Zheng, Z.-L. Sun, and X.-C. Cai, "Classification of tongue color based on CNN," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 725–729.

[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.

[3] X. Li, Z. Yin, Q. Cui, X. Yi, and Z. Yi, "Tooth-marked tongue recognition using multiple instance learning and CNN features," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 1–8, Feb. 2018.

[4] B. Zhang and H. Zhang, "Significant geometry features in tongue image analysis," *Evidence-Based Complementary Alternative Med.*, vol. 2015, pp. 1–8, Jul. 2015.

[5] W. Tang, Y. Gao, L. Liu, T. Xia, L. He, S. Zhang, J. Guo, W. Li, and Q. Xu, "An automatic recognition of tooth-marked tongue based on tongue region detection and tongue landmark detection via deep learning," *IEEE Access*, vol. 8, pp. 153470–153478, 2020.

[6] J. Wu, B. Zhang, Y. Xu, and D. Zhang, "Tongue image alignment via conformal mapping for disease detection," *IEEE Access*, vol. 8, pp. 9796–9808, 2020.

[7] Y. Dai and G. Wang, "Analyzing tongue images using a conceptual alignment deep autoencoder," *IEEE Access*, vol. 6, pp. 5962–5972, 2018.

[8] Y. Xin, Y. Cao, Z. Liu, Y. Chen, L. Cui, Y. Zhu, H. Hou, G. Zhao, and M. Wang, "Automatic tongue verification based on appearance manifold learning in image sequences for the Internet of medical things platform," *IEEE Access*, vol. 6, pp. 43885–43891, 2018.

[9] X. Wang, B. Zhang, Z. Yang, H. Wang, and D. Zhang, "Statistical analysis of tongue images for feature extraction and diagnostics," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5336–5347, Dec. 2013.

[10] B. Kirschbaum, *Atlas of Chinese Tongue Diagnosis*. Seattle, WA, USA: Eastland Press, 2010.

[11] B. Yao, G. Bradski, and L. Fei-Fei, "A codebook-free and annotation-free approach for fine-grained image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3466–3473.

[12] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.

[13] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[14] X. Li, Y. Tang, and Y. Sun, "Tongue coating classification based on multiple-instance learning and deep features," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2019, pp. 504–511.

[15] X. Lin, Z. Yu, Z. Li, and W. Liu, "Machine learning based tongue image recognition for diabetes diagnosis," in *Proc. Int. Conf. Mach. Learn. Cyber Secur.* Cham, Switzerland: Springer, 2020, pp. 474–484.

[16] C. Wan, Y. Zhang, C. Xia, P. Qian, and Y. Wang, "Fissured tongue image recognition based on support vector machine," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, 2019, pp. 1–5.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[18] B. Wei, L. Shen, Y. Cai, and X.-F. Zhang, "Research on curdy and greas tongue fur analysis for traditional Chinese medicien," *Acta Electronica Sinica*, vol. 31, no. 12, pp. 2083–2086, 2003.

[19] J. Zhang, G. Hu, and X. Zhang, "Extraction of tongue feature related to TCM physique based on image processing," in *Proc. 12th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2015, pp. 251–255.

[20] T. T. Qu, C. M. Xia, Y. Q. Wang, and M. L. M. Zhu, "Recognition of greasy or curdy tongue coating based of wavelet transformation," *Comput. Appl. Softw.*, vol. 33, no. 10, pp. 162–166, 2016.

[21] R. S. Choraś, "Automatic tongue recognition based on color and textural features," in *Proc. Int. Conf. Image Process. Commun.*, 2016, pp. 16–26.

[22] X. Li, Q. Shao, and J. Wang, "Classification of tongue coating using Gabor and tamura features on unbalanced data set," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Dec. 2013, pp. 108–109.

[23] S. Fu, H. Zheng, Z. Yang, B. Yan, H. Su, and Y. Liu, "Computerized tongue coating nature diagnosis using convolutional neural network," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 730–734.

[24] K. Zhang, H. Zhang, and M. Zhu, "Automatic classification of tongue texture color and tongue coating color based on bp neural network," *J. Phys., Conf. Ser.*, vol. 1423, no. 1, 2019, Art. no. 012056.

[25] J.-D. Yang and P. Zhang, "Tongue image classification method based on transfer learning and fully connected neural network," *Acad. J. Second Mil. Med. Univ.*, vol. 29, no. 8, p. 19, 2018.

[26] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, vol. 15, no. 2, pp. 561–568.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[30] K. M. Hosny, M. A. Kassem, and M. M. Fouad, "Classification of skin lesions into seven classes using transfer learning with alexnet," *J. Digit. Imag.*, vol. 33, no. 5, pp. 1325–1334, 2020.
[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[32] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *J. Digit. Imag.*, vol. 30, no. 2, p. 234, 2017.
[33] X. Wang, J. Liu, C. Wu, J. Liu, Q. Li, Y. Chen, X. Wang, X. Chen, X. Pang, B. Chang, J. Lin, S. Zhao, Z. Li, Q. Deng, Y. Lu, D. Zhao, and J. Chen, "Artificial intelligence in tongue diagnosis: Using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 973–980, Jan. 2020.
[34] S. Manivannan, C. Cobb, S. Burgess, and E. Trucco, "Sub-category classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 308–316.
[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[36] E. S. Gadelmawla, "A vision system for surface roughness characterization using the gray level co-occurrence matrix," *NDT & E Int.*, vol. 37, no. 7, pp. 577–588, Oct. 2004.
[37] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 6, pp. 460–473, Jun. 1978.
[38] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Automat. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.
[39] E. Oja, "Subspace methods of pattern recognition," *Signal Process.*, vol. 7, no. 1, p. 79, 1983.
[40] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[41] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[42] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1073–1080.

YONGHUI TANG received the B.S. degree in computer science from Anqing Normal University, Anhui, China, in 2018. She is currently pursuing the M.S. degree in computer science with Shanghai University, Shanghai, China.

YUE SUN received the M.S. degree in computer science from Shanghai University, Shanghai, China, in 2019. Her research interests include machine learning and computer vision.

JOHN Y. CHIANG received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1985, the M.S. and Ph.D. degrees in electrical engineering from the Northwestern University, Evanston, IL, USA, in 1987 and 1990, respectively. He is currently a Professor with the Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan. His research interests include deep vision, pattern recognition, image processing, and automatic tongue diagnosis systems.

XIAOQIANG LI (Member, IEEE) received the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2004.

He is currently an Associate Professor of computer science with Shanghai University, Shanghai. His current research interests include image processing, pattern recognition, computer vision, and machine learning.