Journal Pre-proofs


# Artificial intelligence in tongue diagnosis: using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark

Xu Wang, Jingwei Liu, Chaoyong Wu, Junhong Liu, Qianqian Li, Yufeng Chen, Xinrong Wang, Xinli Chen, Xiaohan Pang, Binglong Chang, Jiaying Lin, Shifeng Zhao, Zhihong Li, Qingqiong Deng, Yi Lu, Dongbin Zhao, Jianxin Chen

PII: S2001-0370(20)30032-5
DOI: https://doi.org/10.1016/j.csbj.2020.04.002
Reference: CSBJ 498

To appear in: *Computational and Structural Biotechnology Journal*

Received Date: 23 January 2020
Revised Date: 25 March 2020
Accepted Date: 3 April 2020

Please cite this article as: 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, J. Chen, Artificial intelligence in tongue diagnosis: using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark, *Computational and Structural Biotechnology Journal* (2020), doi: https://doi.org/10.1016/j.csbj.2020.04.002

**Title:** Artificial intelligence in tongue diagnosis: using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark

**Author names:** Xu Wang[a,#], Jingwei Liu[a,#], Chaoyong Wu[a], Junhong Liu[b], Qianqian Li[a], Yufeng Chen[a], Xinrong Wang[a], Xinli Chen[a], Xiaohan Pang[a], Binglong Chang[a], Jiaying Lin[a], Shifeng Zhao[c], Zhihong Li[a], Qingqiong Deng[c], Yi Lu[d], Dongbin Zhao[d], Jianxin Chen[a,*]

**Author affiliations:**

a. Being University of Chinese Medicine, Beijing 100029, China.

b. Beijing University of Posts and Telecommunications, Beijing 100876, China.

c. Beijing Normal University, Beijing 100875, China.

d. State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

**Corresponding author:**

Jianxin Chen. Email: cjx@bucm.edu.cn; Tel: +86 10 6428 6398; Fax: +86 10 6428 6398.

**Abstract**

Tongue diagnosis plays a pivotal role in traditional Chinese medicine (TCM) for thousands of years. As one of the most important tongue characteristics, tooth-marked tongue is believed relating to spleen deficiency and can greatly contribute to the symptoms differentiation and treatment selection. Yet, the tooth-marked tongue recognition for TCM practitioners is subjective and challenging. Most of the previous studies have concentrated on subjectively selected features of the tooth-marked region and gained accuracy under 80%. In the present study, we proposed an artificial intelligence framework using deep convolutional neural network (CNN) for the recognition of tooth-marked tongue. First, we constructed relatively large datasets with 1548 tongue images captured by different equipments. Then, we used ResNet34 CNN architecture to extract features and perform classifications. The overall accuracy of the models was over 90%. Interestingly, the models can be successfully generalized to images captured by other devices with different illuminations. The good effectiveness and generalization of our framework may provide objective and convenient computer-aided tongue diagnostic method on tracking disease progression and evaluating pharmacological effect from a bioinformatics perspective.

**Key words:** tooth-marked tongue; traditional Chinese Medicine; convolutional neural network; tongue diagnosis; artificial intelligence.

## 1 Introduction

Tongue diagnosis plays a pivotal role in traditional Chinese medicine (TCM) for thousands of years. Tongue characteristics, such as tongue shape and color, can reflect the internal health status of the body (e.g., organs, qi, blood, cold, heat) and the severity or progression of the diseases. By observing tongue characteristics, TCM practitioners can differentiate clinical symptoms and choose proper treatments. However, traditional tongue diagnosis is based on practitioners' subjective eye observation, which is often biased by personal experience, environmental lighting variations, and etc. Therefore, it is necessary to develop an objective and quantitative tongue diagnostic method that can aid practitioners' diagnosis [1].

Tooth-marked tongue recognition may provide an ideal example to achieve these goals. As one of the most important tongue characteristics, tooth-mark along the lateral borders results from a fatter tongue body compressed by adjacent tooth. Tooth-marked tongue is often related to spleen deficiency, yang vacuity with cold dampness, phlegm and retained fluid, and blood stasis according to TCM theory. In addition, changes in the microcirculation of the dentate tongue include blood supply disorders, local hypoxia, and tissue edema. The clinical manifestations in individuals with tooth-marked tongue include loss of appetite, borborygmus, gastric distention, and loose stool. The diagnosis of tooth-marked tongue can greatly contribute to the symptoms differentiation and treatment selection [2]. Yet, the tooth-marks on the tongue have various types (e.g., different colors and shapes), which makes the recognition of tooth-marked tongue challenging for TCM practitioners [3].

Actually, researchers have attempted to build computerized tooth-marked tongue recognition models using image processing, statistical, and machine learning methods in recent years [3–8]. Most of the studies have concentrated on local color and concavity-convexity features of the tooth-marked region. For example, Hsu and colleagues have conducted RGB color composition of the tongue region image and found that G color spectrum of the tooth-marks is lower than tongue body and tongue fur [4,5]. Li et al. have used concavity information to generate suspected tooth-marked regions, then extracted features from these regions, and at last used a multi-instance support vector machine (SVM) classifier for final tooth-marked tongue classification [3]. Recently, with the continuous development of artificial intelligence and deep learning technology, convolutional neural network (CNN) models have gradually been applied to the classification of tooth-marked tongue. CNN can extract high-level semantic features automatically and perform well in many image classification tasks [9–11]. For instance, Sun et al. have proposed a 7 layer CNN model with the whole tongue region images as input to recognize tooth-marked tongue and achieved the accuracy of 78.6% [7].

Overall, although these previous studies have obtained lots of achievements in the field of automatic tooth-marked tongue recognition, there are still some important

issues to be explored. First, the accuracy of the previous models is usually under 80%. Then, the datasets only come from the identical equipment, indicating that the generalization of the models to classify tongue images captured by other devices remains unknown. Third, the sample size of the datasets is relatively small (e.g., 645 for [3] and [7]) which may also restrict the generalization of the trained models. Finally, researchers only use tongue region images isolated from raw tongue images to train and test the models, without exploring the specific influence of irrelevant facial and surrounding portions.

In the present study, we expanded current tooth-marked tongue classification techniques in the following ways. First, we constructed larger tooth-marked tongue datasets with more than 1500 raw tongue images captured by different equipments, and also labeled tongue region for each image resulting tongue region image datasets. Second, to fully embody the advantages of deep learning, we used CNN models with deeper layers to extract features and perform classifications.

## 2 Materials and Methods

### 2.1 Datasets construction and preprocessing

To build a relatively consistent and stable tongue image dataset, we acquired tongue images using standard equipments designed by Shanghai Daosh Medical Technology Ltd (DS01-B) or Shanghai Xieyang Intelligent Technology Ltd (XYSM01). Then, the images were transferred to personal computer for assessment. The tongue images were differentiated into tooth-marked or non-tooth-marked tongue by three professional TCM practitioners (with 20 years of clinical experience) from Beijing University of Chinese Medicine. All professionals were well trained and had normal or corrected-to-normal vision. The tongue images were sequentially assessed by professionals using HP P223 monitor (21.5 inches, 1920 × 1080) in the same computer room. The detailed assessment procedure included three steps in this study. First, professionals discussed the diagnostic criteria for tooth-marked tongue. Second, one professional labeled all 1548 images to “tooth-mark” or “non-tooth-mark” folder. Third, the other two professionals reviewed the labeling results respectively. For

instances of disagreement, three professionals should discuss and make the final decisions. The images with the consensuses by three professionals were included in the dataset for developing artificial intelligent model. The resultant dataset contained 672 tongue images with tooth-mark and 876 tongue images without tooth-mark. In addition, we also manually labeled tongue region for each raw tongue image. The purpose of isolating the tongue region is to facilitating model performance by controlling the influence of irrelevant facial portions and background surrounding the tongue. As a result, two datasets, including raw tongue image dataset and tongue region image dataset, were constructed. The exemplar of acquired raw tongue image and tongue region image were shown in Figure 1.

**2.2 Network architecture**

CNN performs well in image classification tasks. However, as the depth of the CNN increases, the training becomes more difficult and training error becomes higher. As one of the typical CNN architectures, the deep residual learning network (ResNet) model enables the network robust to the vanishing gradient and degradation problems caused by network depth, and performs better than traditional network model [9]. Therefore, we used a typical ResNet architecture consisted of 34 layers (ResNet34) to classify the tongue images in the present study. The visualization of the ResNet34 model structure was shown in Figure 2. Rectified linear unit (ReLU) [12] was used as an activation function after each convolutional layer (Equation. 1).

$$ReLU(x) = max\{0, x\} = \begin{cases} x, & x > 0 \\ 0, & x < 0 \end{cases} \quad (1)$$

**2.3 Model training and testing**

Models were developed and trained using PyTorch (https://pytorch.org) Python framework on Windows10 system with 1 NVIDIA 1080 GPU and i7 8700K CPU. The network was initialized using pretrained weights on ImageNet datasets (https://pytorch.org/docs/stable/torchvision/models.html) [13] and fine-tuned on our tongue image dataset. Since the fundus tongue images from different devices may have various resolutions, all available images were randomly resized and cropped to 416 × 416 pixels, and also horizontally flipped before model training. Then, the

network was fine-tuned for 40 epochs using a batch size of 16. Stochastic gradient descent (SGD) with learning rate of 0.001 and momentum of 0.9 was used as an optimizer. In the testing stage, the input testing images of the trained network were resized to 420 × 420 pixels and center cropped to 416 × 416 pixels.

**2.4 Statistics for model evaluation**

The accuracy (Equation. 2), sensitivity (Equation. 3), and specificity (Equation. 4) were used to evaluate the performance of the model [14–17]. True positive (TP) represents the number of correctly classified as tooth-marked tongue, true negative (TN) represents the number of correctly classified as non-tooth-marked tongue, false positive (FP) is the number of incorrectly classified as tooth-marked tongue, and false negative (FN) is the number of incorrectly classified as non-tooth-marked tongue.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4}$$

The k-fold cross-validation is a robust and less biased method for evaluating the performance of a model. The general procedure is as follows: 1) Randomly split the data into k subsets; 2) Reserve one subset and train the model on all other subsets; 3) Test the model on the reserved subset and record the evaluation metrics; 4) Repeat above processes until each of the k subsets has served as the test dataset; 5) Summarize the performance by compute the average and variance of the k models' evaluation metrics. The choice of k is usually 5 or 10 in artificial intelligence studies. Five-fold cross-validation was performed in the current study. The 1548 raw tongue images (described in Section 2.1) were randomly shuffled and then divided into 5 subsets. First 3 subsets contained 310 tongue images and last 2 subsets contained 309 tongue images. Note that the partitions of 1548 tongue region images were in line with raw tongue images. In each validation, we used 4 subsets for training and the

other 1 subset for testing (Figure. 1). Then, the average and standard deviation (SD) of the 5 models' accuracy, sensitivity, and specificity were calculated.

All aforementioned statistical metrics were calculated using Python software.

**2.5 Model validating**

Four types of experiments were used to prove the effectiveness of the model in the present study.

First, to further evaluate our model, we also constructed a new testing dataset with 50 tongue images captured by an ordinary camera without strictly controlling for the surrounding circumstance. That is, the images in this dataset had various illuminations. Thus, it may increase the difficulty in classification and can validate the proposed method more strictly. The images were also differentiated to 27 tooth-marked and 23 non-tooth-marked tongue images by the same procedures described in Section 2.1. In addition, we also labeled the tongue region for each raw tongue image. Then, all raw tongue images here were classified using the aforementioned 5 models trained by raw tongue images, and all tongue region images in this dataset were classified using the 5 models trained by tongue region images. The main procedures were shown in Figure 1.

Second, VGG16, which was proposed by the Visual Geometry Group at University of Oxford [10], was used for comparative experiments. The "16" means 13 convolutional layers and 3 fully connected layers. Because the input image size here was 416 × 416 instead of 224 × 224, an adaptive average pooling layer with output size of 7 × 7 was used before fully connected (FC) layer. The training parameters here were consistent with ResNet34 aforementioned. The VGG16 have also been pretained on ImageNet datasets that can greatly reduce training time on our datasets.

The third is the comparison between our proposed models with other works. Here, we focused on the method proposed by Sun et al. [7] which has achieved better tooth-marked tongue classification performance than other previous studies. Since Sun et al. have not released their code to the readers, we conservatively replicated their CNN model with 7 layers and used the best setting stated in their description. We conducted experiments on both of our raw tongue image dataset and tongue

region image dataset. The five-fold cross-validation settings were the same as in the aforementioned paragraph.

Finally, we used Gradient-weighted Class Activation Mapping (Grad-CAM) [18] method to visualize the most indicative regions for tooth-marked tongue and interpreted the model predictions. Here, Grad-CAM uses the gradients of the target concept (i.e., tooth-marked tongue) flowing into the final convolutional layer of the CNN to produce a coarse localization map highlighting the important regions in the image for predicting the tooth-marked tongue.

## 3 Results

### 3.1 Five-fold cross-validation on raw tongue image dataset and tongue region image dataset

The five-fold cross-validated tooth-marked tongue classification results using ResNet34 architecture on 1548 raw tongue images are shown in Table 1. First, we found that the performance of the classification model is relatively good and stable. The overall accuracy is 90.50% which proves the effectiveness of the method. Second, the overall sensitivity is 87.25% and specificity is 93.00%, indicating that the models have relatively high sensitivity and specificity. In order to control the influence of irrelevant facial and surrounding portions in the images, experiments were carried out using tongue region image dataset. As we expected, the overall classification accuracy on tongue region image dataset is 91.47% (Table 1), which is 0.97% higher than the average accuracy on raw tongue image dataset.

**Table 1.** Five-fold cross-validation results of the ResNet34 architecture

| | Raw tongue image dataset (n = 1548) | | | Tongue region image dataset (n = 1548) | | |
|---|---|---|---|---|---|---|
| | Acc | Sens | Spec | Acc | Sens | Spec |
| Fold 1 | 88.71% | 82.22% | 93.71% | 90.97% | 86.67% | 94.29% |
| Fold 2 | 93.23% | 90.48% | 95.11% | 92.58% | 88.10% | 95.65% |
| Fold 3 | 89.35% | 84.85% | 92.70% | 90.97% | 84.09% | 96.07% |

| | | | | | | |
|---|---|---|---|---|---|---|
| Fold 4 | 91.26% | 90.70% | 91.67% | 92.88% | 88.37% | 96.11% |
| Fold 5 | 89.97% | 88.00% | 91.82% | 89.97% | 87.33% | 92.45% |
| Average (SD) | **90.50%** (1.60%) | 87.25% (3.29%) | 93.00% (1.28%) | **91.47%** (1.09%) | 86.91% (1.53%) | 94.91% (1.40%) |

Abbreviations: Acc, accuracy; Sens, sensitivity; Spec, specificity; SD, standard deviation.

### 3.2 Validation on new testing dataset

To further evaluate our model, we also conducted experiments on new testing datasets. The new raw tongue image dataset consisted of 50 tongue images, and new tongue region image dataset contained 50 tongue region images manually isolated form raw images. The average accuracy of the trained models in Section 3.1 is 83.20% and 88.80% for the 2 datasets, respectively (Table 2). Since the images from this testing dataset were captured by camera under various light conditions, the overall accuracy of higher than 85.00% indicating that our models can be generalized to images from different devices with different illuminations.

**Table 2.** Classification results on a new testing dataset

| | New raw tongue image dataset (n = 50) | | | New tongue region image dataset (n = 50) | | |
|---|---|---|---|---|---|---|
| | Acc | Sens | Spec | Acc | Sens | Spec |
| Model 1 | 78.00% | 66.67% | 91.30% | 90.00% | 92.59% | 86.96% |
| Model 2 | 86.00% | 74.07% | 100% | 80.00% | 85.19% | 95.65% |
| Model 3 | 86.00% | 77.78% | 95.65% | 88.00% | 92.59% | 82.61% |
| Model 4 | 88.00% | 85.19% | 91.30% | 92.00% | 85.19% | 100% |
| Model 5 | 78.00% | 85.19% | 69.57% | 84.00% | 96.30% | 69.57% |
| Average (SD) | **83.20%** (4.30%) | 77.78% (7.03%) | 89.56% (10.50%) | **88.80%** (2.71%) | 90.37% (4.44%) | 86.96% (10.65%) |

Abbreviations: Acc, accuracy; Sens, sensitivity; Spec, specificity; SD, standard deviation.

### 3.3 Comparison with VGG16 architecture

To investigate whether the CNN architecture influence the experimental results, VGG16 was used for comparison. The results are shown in Table 3. The average accuracy of five-fold cross-validation is 89.40% and 90.96% on raw tongue image dataset and tongue region image dataset, respectively. Therefore, ResNet34 architecture can increase the accuracy of tooth-marked tongue classification by 1.10% on raw tongue images and 0.52% on tongue region images.

**Table 3.** Five-fold cross-validation results of the VGG16 architecture

| | Raw tongue image dataset (n = 1548) | | | Tongue region image dataset (n = 1548) | | |
|---|---|---|---|---|---|---|
| | Acc | Sens | Spec | Accuracy | Sens | Spec |
| Fold 1 | 88.39% | 81.48% | 93.71% | 90.97% | 82.22% | 97.71% |
| Fold 2 | 90.97% | 85.71% | 94.57% | 91.29% | 88.10% | 93.48% |
| Fold 3 | 88.71% | 83.33% | 92.70% | 90.32% | 86.36% | 93.26% |
| Fold 4 | 91.26% | 86.05% | 95.00% | 92.88% | 93.02% | 92.78% |
| Fold 5 | 87.70% | 84.00% | 91.19% | 89.32% | 89.33% | 89.31% |
| Average (SD) | **89.41%** (1.44%) | 84.11% (1.67%) | 93.43% (1.37%) | **90.96%** (1.17%) | 87.81% (3.55%) | 93.31% (2.67%) |

Abbreviations: Acc, accuracy; Sens, sensitivity; Spec, specificity; SD, standard deviation.

### 3.4 Comparison with other work

Most of previous methods are based on local concave features and set threshold subjectively to classify the tooth-marked tongue. A recent work, using 7 layers' CNN to automatically extract features, has gained accuracy higher than other previous works. [7]. We conducted experiments on our datasets using Sun's method. The results

are shown in Table 4. The average accuracies, 70.61% on raw tongue image dataset and 71.77% on tongue region image dataset, are nearly 20% lower than our methods. In addition, due to the differences in data distributions and model architectures, Sun's method is usually failed to recognize tooth-marked tongue. Thus, the sensitivity is much lower than specificity, which may result in the low overall accuracy.

**Table 4.** Five-fold cross-validation results of the Sun's architecture

| | Raw tongue image dataset (n = 1548) | | | Tongue region image dataset (n = 1548) | | |
|---|---|---|---|---|---|---|
| | Acc | Sens | Spec | Acc | Sens | Spec |
| Fold 1 | 71.61% | 54.07% | 85.17% | 70.65% | 48.89% | 87.43% |
| Fold 2 | 74.19% | 56.35% | 86.41% | 75.16% | 63.49% | 83.15% |
| Fold 3 | 67.42% | 43.18% | 85.39% | 70.65% | 54.55% | 82.58% |
| Fold 4 | 74.11% | 57.36% | 86.11% | 73.46% | 58.91% | 83.89% |
| Fold 5 | 65.70% | 48.67% | 81.76% | 68.93% | 61.33% | 76.10% |
| Average (SD) | **70.61%** (3.47%) | 51.93% (5.31%) | 84.96% (1.67%) | **71.77%** (2.23%) | 57.43% (5.20%) | 82.63% (3.68%) |

Abbreviations: Acc, accuracy; Sens, sensitivity; Spec, specificity; SD, standard deviation.

Notably, the input images are downscaled to 256 × 256 and randomly cropped to 224 × 224 in Sun's method. To eliminate the influence of the image size, we also performed experiments with the input image size in our method (416 × 416). As we can see from Table 4 and 5, the input image size may not significantly affect the model's classification results.

**Table 5.** Five-fold cross-validation results of the Sun's architecture with input image size of 416

| | Raw tongue image dataset (n = 1548) | | | Tongue region image dataset (n = 1548) | | |
|---|---|---|---|---|---|---|
| | Acc | Sens | Spec | Acc | Sens | Spec |

| | | | | | | |
|---|---|---|---|---|---|---|
| Fold 1 | 68.06% | 44.44% | 86.29% | 69.35% | 57.78% | 78.29% |
| Fold 2 | 73.87% | 56.35% | 85.87% | 74.84% | 53.17% | 89.67% |
| Fold 3 | 68.06% | 50.00% | 81.46% | 70.65% | 56.82% | 80.90% |
| Fold 4 | 73.79% | 65.12% | 80.00% | 73.79% | 51.16% | 90.00% |
| Fold 5 | 66.67% | 58.67% | 74.21% | 68.61% | 60.00% | 76.73% |
| Average (SD) | **70.09%** (3.10%) | 54.92% (7.13%) | 81.57% (4.41%) | **71.45%** (2.45%) | 55.79% (3.20%) | 83.12% (5.64%) |

Abbreviations: Acc, accuracy; Sens, sensitivity; Spec, specificity; SD, standard deviation.

In sum, the average accuracy of our ResNet34, VGG16, and Sun's methods with different input image sizes are shown in Figure 3. Our models can increase the accuracy of tooth-marked tongue classification by about 20%.

### 3.5 Visualization of the indicative regions for tooth-marked tongue classification

To ensure whether tooth-marked region contribute more to the tooth-marked tongue classification, Grad-CAM with respect to final convolutional layer of our model was performed. In the CNN, deeper layers can capture higher levels of the semantic information. Therefore, the final convolutional layer contains the best correspondence between semantic and spatial information of the images. As shown in Figure 4, the Grad-CAM highlights the indicative regions, which are usually tooth-marked regions along the lateral borders, for tooth-marked tongue classification. The visualization by using Grad-CAM can help us to evaluate the model and also provide TCM practitioners more intuitive information for aiding diagnosis.

## 4 Discussion

The tooth-mark characteristic of the tongue is a crucial indicator in the TCM assessment. Here, we proposed a framework for the recognition of tooth-marked tongue. First, we captured 1548 raw tongue images by different standard equipments, and differentiated these images into 672 tongue images with tooth-mark and 876

tongue images without tooth-mark. We also labeled tongue region for each image resulting tongue region image dataset. Then, we used ResNet34 CNN models to extract features and perform classifications. The overall accuracy of the models was over 90% on both raw tongue image dataset and tongue region image dataset. Interestingly, the models can be generalized to images captured by other devices with different illuminations very well. These results show that the method in the present study greatly improve the accuracy than previous studies and prove the effectiveness of the models even when the images are from different sources.

Our study may shed new light on symptoms or diseases diagnosis and pharmacological evaluation based tongue characteristics. Several previous studies have reported encouraging results using the tongue image features for differentiating healthy and unhealthy tongue [19,20], diagnosing type 2 diabetes mellitus [21], early-stage breast cancer [22] and gastritis [23], but they usually include more preprocessing steps, extract features empirically, and use traditional statistic and machine learning methods. Our CNN architectures in general can automatically extract features avoiding feature selection and reduce manual steps, which are key elements to enable translation of such systems in to the clinical practice. Moreover, the good effectiveness (higher than 90%) and generalization (not rely on specific device) of our framework may provide objective and convenient computer-aided method on tracking disease progression and evaluating pharmacological effect from a bioinformatics perspective.

Yet, there are several important topics for future researches. First, the specificity is slightly higher than sensitivity in most of our experiments. It may be caused by the inequality of positive and negative samples. Further researches are needed to investigate the influence of the number of tooth-marked tongue and non-tooth-marked tongue images. Second, our results demonstrate that the ResNet34 (the deepest CNN in our study) outperformed the shallower architectures (VGG16, Sun's 7 layer model) for all the metrics, including accuracy, sensitivity, and specificity. However, the CNN with deeper layer is usually more computationally intensive. Therefore, it's necessary to find better balance between the model performance and computation cost. Third,

we found that the overall accuracy of the models on tongue region image dataset is slightly higher than that of raw tongue image dataset, suggesting the importance of developing advanced tongue segmentation algorithms in the future [24]. And our framework may provide an ideal platform for evaluating these algorithms. Fourth, the validation of the manually created dataset is essential for developing algorithm. Future studies should carefully control the manual classification process (e.g., angle of the view and distance between the eyes and monitor screen), and evaluate the inter- and intro-observer reliability. Finally, there are different grades of tooth-mark appearances. Differentiating tongue images into more groups and constructing multiple classification models may increase the clinical applicability.

**Author Contributions**

JXC conceived and designed the study. XW, JWL, and JHL analyzed data; XW wrote the manuscript; JWL, CYW, QQL, XRW, YFC, XLC, XHP, BLC, and JYL collected the data; QQD, ZHL, SFZ, YL, and DBZ reviewed the methods and the results. All authors have reviewed the manuscript.

**Funding**

This work is supported by the National Key Research and Development Program (2017YFC1700106).

**Conflict of Interest**

The authors declare that they have no conflict of interest.

**Figure legends**

**Figure 1.** Overview of the datasets construction and main processing procedures. (A) The illustration of tongue image capturing with standard equipment. (B) Construction of the raw tongue image dataset and exemplar of tooth-marked and non-tooth-marked tongue. (C) Construction of the tongue region image dataset and exemplar of tooth-marked and non-tooth-marked tongue. (D) The training, testing, and validating

of the convolutional neural network model. (E) The testing of the models in a new dataset of tongue images captured by ordinary camera.

**Figure 2.** Visualization of the ResNet34 model structure. Conv and pool stand for convolutional and pooling, respectively. The pooling or stride size is 2 (denoted by "/2"). "7 × 7 conv, 64" means that the convolutional kernel size is 7 × 7 and number of filters is 64. Solid lines indicate the input and output have identical dimensions, dashed lines indicate the input and output have different dimensions.

**Figure 3.** Comparison with other tooth-marked tongue recognition methods. Our models with ResNet34 and VGG16 architectures can increase the accuracy of tooth-marked tongue classification by about 20%.

**Figure 4.** Grad-CAM visualization examplars for the tongue image with tooth-mark. The upper panel shows the tongue region images and lower panel shows the heatmap of indicative regions by Grad-CAM overlapped on the tongue region images.

**References**

[1] Zhang D, Zhang H, Zhang B. Tongue image analysis. New York, NY: Springer Berlin Heidelberg; 2017.

[2] Li F, Dong C. Diagnostics of Traditional Chinese Medicine. Beijing: Science Press; 2017.

[3] Li X, Zhang Y, Cui Q, Yi X, Zhang Y. Tooth-Marked Tongue Recognition Using Multiple Instance Learning and CNN Features. IEEE Trans Cybern 2019;49:380–7. https://doi.org/10.1109/TCYB.2017.2772289.

[4] Hsu Y, Chen Y, Lo L, Chiang JY. Automatic tongue feature extraction. 2010 Int. Comput. Symp., 2010, p. 936–41. https://doi.org/10.1109/COMPSYM.2010.5685377.

[5] Lo LC, Chen YF, Chen WJ, Cheng TL, Chiang JY. The Study on the Agreement between Automatic Tongue Diagnosis System and Traditional Chinese Medicine Practitioners. Evidence-Based Complement Altern Med 2012. https://doi.org/Artn 50506310.1155/2012/505063.

[6] Shao Q, Li XQ, Fu ZC. Recognition of Teeth-Marked Tongue Based on Gradient of Concave Region. 2014 Int Conf Audio, Lang Image Process (Icalip), Vols 1-2 2014:968–72.

[7] Sun Y, Dai SM, Li JD, Zhang Y, Li XQ. Tooth-Marked Tongue Recognition Using Gradient-Weighted Class Activation Maps. Futur Internet 2019;11. https://doi.org/ARTN 4510.3390/fi11020045.

[8] Wang H, Zhang XF, Cai YH. Research on Teeth Marks Recognition in Tongue Image. 2014 Int Conf Med Biometrics (Icmb 2014) 2014:80–4. https://doi.org/10.1109/Icmb.2014.21.

[9] He KM, Zhang XY, Ren SQ, Sun J. Deep Residual Learning for Image Recognition. 2016 Ieee Conf Comput Vis Pattern Recognit 2016:770–8. https://doi.org/10.1109/Cvpr.2016.90.

[10] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv E-Prints 2014:arXiv:1409.1556.

[11] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 25, Lake Tahoe, NV, USA: 2012, p. 1097–105. https://doi.org/10.1145/3065386.

[12] Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair. Proc. 27th Int. Conf. Mach. Learn. (ICML-10), June 21-24, 2010, Haifa, Isr., 2010.

[13] He KM, Zhang XY, Ren SQ, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015 Ieee Int Conf Comput Vis 2015:1026–34. https://doi.org/10.1109/Iccv.2015.123.

[14] Gorur K, Bozkurt MR, Bascil MS, Temurtas F. GKP signal processing using deep CNN and SVM for tongue-machine interface. Trait Du Signal 2019;36:319–29. https://doi.org/10.18280/ts.360404.

[15] Gorur K, Bozkurt MR, Bascil MS, Temurtas F. Glossokinetic potential based tongue–machine interface for 1-D extraction using neural networks. Biocybern Biomed Eng 2018;38:745–59. https://doi.org/10.1016/j.bbe.2018.06.004.

[16] Bascil MS. Convolutional neural network to extract the best treatment way of warts based on data mining. Rev d'Intelligence Artif 2019;33:165–70. https://doi.org/10.18280/ria.330301.

[17] Bascil MS. A New Approach on HCI Extracting Conscious Jaw Movements Based on EEG Signals Using Machine Learnings. J Med Syst 2018;42:169. https://doi.org/10.1007/s10916-018-1027-1.

[18] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2017 Ieee Int Conf Comput Vis 2017:618–26. https://doi.org/10.1109/Iccv.2017.74.

[19] Ding J, Cao GT, Meng D. Classification of Tongue Images Based on Doublet SVM. 2016 Int Symp Syst Softw Reliab 2016:77–81. https://doi.org/10.1109/Isssr.2016.24.

[20] Wang XZ, Zhang B, Yang ZM, Wang HQ, Zhang D. Statistical Analysis of Tongue Images for Feature Extraction and Diagnostics. Ieee Trans Image Process 2013;22:5336–47. https://doi.org/10.1109/Tip.2013.2284070.

[21] Hsu PC, Wu HK, Huang YC, Chang HH, Lee TC, Chen YP, et al. The tongue features associated with type 2 diabetes mellitus. Medicine (Baltimore) 2019;98.

https://doi.org/ARTN e1556710.1097/MD.0000000000015567.
[22] Lo LC, Cheng TL, Chen YJ, Natsagdorj S, Chiang JY. TCM tongue diagnosis index of early-stage breast cancer. Complement Ther Med 2015;23:705–13. https://doi.org/10.1016/j.ctim.2015.07.001.
[23] Dan M, Guitao C, Duan Y, Minghua Z, Liping T, Jiatuo X, et al. A deep tongue image features analysis model for medical application. 2016 IEEE Int. Conf. Bioinforma. Biomed., 2016, p. 1918–22. https://doi.org/10.1109/BIBM.2016.7822815.
[24] Zhou C, Fan H, Li Z. Tonguenet: Accurate Localization and Segmentation for Tongue Images Using Deep Neural Networks. IEEE Access 2019;7:148779–89. https://doi.org/10.1109/ACCESS.2019.2946681.

**Author Statement**

**Xu Wang**: Methodology, Formal analysis, Software, Writing - Original Draft & Review & Editing. **Jingwei Liu**: Investigation, Data Curation, Writing - Original Draft. **Chaoyong Wu**: Investigation, Data Curation. **Junhong Liu**: Software. **Qianqian Li**: Investigation. **Yufeng Chen**: Investigation. **Xinrong Wang**: Investigation. **Xinli Chen**: Investigation. **Xiaohan Pang**: Investigation. **Binglong Chang**: Investigation. **Jiaying Lin**: Investigation. **Shifeng Zhao**: Software. **Zhihong Li**: Resources. **Qingqiong Deng**: Software. **Yi Lu**: Writing - Review. **Dongbin Zhao**: Writing - Review. **Jianxin Chen**: Conceptualization, Funding acquisition, Supervision, Writing - Review.

**Conflict of Interest**

The authors declare that they have no conflict of interest.

Highlights

1. Large datasets for tooth-marked tongue recognition are constructed.
2. The deep CNN in artificial intelligence gain classification accuracy over 90%.
3. The models can be successfully generalized to images with different illuminations.

4. Isolation of the tongue region can enhance tongue diagnosis performance.