



Deep learning for subtype recognition of Yang deficiency tongue images in traditional Chinese medicine

Tongbin Zhang^a, Haoran Xu^a, Ziyi Wang^a, Chuanjun Pan^a, Zheng Wang^b, Lei Wang^{a*}

a. School of Mechanical and Automotive Engineering, Anhui Polytechnic University, Wuhu, Anhui 241000, China

b. Department of Thyroid and Breast Surgery, The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital), Wuhu, Anhui 241001, China

ARTICLE INFO ABSTRACT

Article history

Received 11 March 2026

Accepted 21 April 2026

Available online 25 June 2026

Keywords

Yang deficiency syndrome

Tongue image analysis

YD-YOLO

Feature extraction

Syndrome differentiation

Objective To address the lack of fine-grained clinical recognition for specific Yang deficiency syndrome subtypes and the limitations of conventional object detection models in extracting irregular, low-contrast tongue phenotypes. This study aims to develop an objective subtype recognition framework based on an improved You Only Look Once nano (YOLO11n) architecture, using a standardized visual phenotype matrix to translate macroscopic traditional Chinese medicine (TCM) descriptions into quantifiable clinical targets.

Methods This cross-sectional diagnostic study consecutively enrolled adult inpatients admitted to the Department of Thyroid and Breast Surgery, The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital), between September 1, 2024 and June 1, 2025, who were suspected of having Yang deficiency constitution based on initial TCM consultation. Clinical tongue image data were collected for analysis. Based on an Expert Visual Phenotype Annotation Matrix, a five-category recognition system was established, including the following TCM syndrome subtypes: spleen-dampness exuberance syndrome, mild kidney Yang deficiency syndrome, upper heat and lower cold syndrome, simultaneous Yin-Yang deficiency syndrome, and Yin deficiency and fluid depletion syndrome (negative control). The proposed Yang deficiency YOLO (YD-YOLO) model, built upon the YOLO11n baseline, integrates the Cross Stage Partial with kernel size 2 (C3k2)-GhostBottleneck-Dynamic Convolution (GBDC) module into the backbone to adaptively extract low-contrast features, and embeds the multi-path aggregation coordinate attention (MACA) mechanism into the neck to suppress background interference through multi-scale spatial coordination. Gradient-weighted class activation mapping (Grad-CAM) was used to visualize feature attribution and evaluate the biological plausibility of the model's focus. Model performance was evaluated through ablation and comparative experiments using mean average precision (mAP), precision, recall, F1 score, inference speed (frames per second, FPS), overall accuracy, Cohen's kappa, and the area under the receiver operating characteristic (ROC) curve (AUC).

Results Based on the final inclusion of 1 186 clinical cases, the YD-YOLO model had an overall accuracy of 91.5%, a Cohen's kappa of 0.912, and an mAP@50 of 0.731 [higher than the YOLO11n baseline (0.681)], with AUC ranging from 0.91 to 0.97 across all TCM syndrome subtypes. Among the TCM syndrome subtypes, the mild kidney Yang deficiency syndrome had the highest mAP@50 (0.900), and the inference speed reached 89.00 FPS. Grad-CAM analysis

*Corresponding author: Lei Wang, E-mail: wangdalei2000@126.com.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmcd.2026.05.006

Citation: ZHANG TB, XU HR, WANG ZY, et al. Deep learning for subtype recognition of Yang deficiency tongue images in traditional Chinese medicine. Digital Chinese Medicine, 2026, 9(2): 197-210.

Copyright © 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

showed that the model localized activation to key TCM pathological features, such as marginal tooth marks and focal root coatings, while suppressing non-diagnostic oral background noise.

Conclusion The YD-YOLO model demonstrates the feasibility of deep learning for the fine-grained classification of TCM Yang deficiency subtypes. By integrating visual phenotype quantification with model interpretability, the proposed framework provides an objective basis for syndrome differentiation, supporting the development of standardized digital diagnostic systems and the provision of clinical decision support in TCM practice.

1 Introduction

As the core link of the four diagnoses of traditional Chinese medicine (TCM), the theoretical basis of tongue diagnosis can be traced back to the *Huangdi Neijing* (《黄帝内经》, *Inner Canon of Huangdi*). After 2 000 years of clinical verification, it has become a key basis for identifying the cold and heat attributes of the constitution. A study of *Shanghan Shejian* (《伤寒舌鉴》, *Treatise on Febrile Diseases and Tongue Diagnosis*) reported that the systematic classification of Yang-deficiency syndromes provides an important reference for later clinical practice^[1]. In addition, the visceral mapping model of the tongue area laid the theoretical foundation for the objectification of modern tongue diagnosis. Recent large-scale community studies and epidemiological reviews indicate that Yang deficiency subtypes remains one of the most prominent imbalanced constitutions, affecting approximately 9.4% to 16.7% of the general population^[2, 3]. Further differentiation of Yang deficiency subtypes may guide clinical prescriptions. From a theoretical perspective, distinct behavioral factors drive the manifestation of specific tongue phenotypes. A high intake of cold drinks weakens the spleen's transportation function, leading to the thick, greasy coating characteristic of the spleen-dampness exuberance syndrome [typically treated with *Lizhong Tang* (理中汤)]. Sleep deprivation consumes kidney essence, manifesting as a pale, tender tongue with a thick coating at the root [mild kidney Yang deficiency syndrome, treated with *Zhenwu Tang* (真武汤)]. Chronic stress and overwork suppress spleen Yang and exhaust kidney essence, contributing to the dark, cracked appearance of the simultaneous Yin-Yang deficiency syndrome. In addition, excessive exposure to air conditioning introduces pathogenic wind and cold, creating the contradictory sign of a red tip and pale root [upper heat and lower cold syndrome, treated with *Guizhi Longgu Muli Tang* (桂枝龙骨牡蛎汤)]^[4]. However, the fine differentiation of these subtypes in current clinical practice remains largely subjective, lacking quantitative criteria. Therefore, an objective and quantifiable approach to differentiate Yang deficiency subtypes is needed to standardize syndrome evaluation and guide targeted therapy^[5].

In recent years, deep learning has driven the modernization of TCM tongue diagnosis, shifting from subjective observation to objective artificial intelligence (AI) frameworks for quantitative feature extraction^[6-10]. Various convolutional neural networks (CNNs), including Faster Region-based-CNN and You Only Look Once (YOLO) architectures, have been applied to medical object detection and multi-label tongue image analysis^[11-16]. Despite these advances, current research primarily treats complex conditions such as kidney deficiency as broad binary classification tasks (i.e., syndrome versus non-syndrome) based on basic tongue body or coating colors^[17, 18]. This approach overlooks the nuanced clinical reality of the Yang deficiency spectrum, where a lack of fine-grained subtype recognition persists due to a disconnect between generic algorithms and TCM theoretical insights^[19].

Technically, this limitation stems from the limited feature-extraction capacity of conventional object detection models when processing irregular, low-contrast clinical signs. While newer models such as YOLO11 introduce components such as Cross Stage Partial with kernel size 2 (C3k2) to enhance general performance^[20, 21], their fixed convolutional kernels struggle with the highly individualized morphological variations of specific tongue features, such as marginal tooth marks and subtle fissures^[22]. In addition, traditional structures often fail to adapt to low-contrast characteristics, such as the attenuation of the red channel in pale tongues. Although some studies have proposed optimized architectures such as Lightweight Pedestrian (LP)-YOLO^[23], a targeted network architecture capable of adaptively extracting these fine-grained visual phenotypes for Yang deficiency subtyping remains unclear. To address the lack of fine-grained recognition of specific Yang deficiency subtypes and the limited capacity of existing algorithms in extracting low-contrast tongue features, this study aims to develop an objective subtype recognition model. By establishing a standardized visual phenotype matrix to translate macroscopic TCM descriptions into quantifiable targets, and by integrating dynamic convolutions with attention mechanisms into the YOLO11 architecture, this research seeks to provide a reliable clinical decision-support tool for syndrome differentiation and prescription recommendations.

2 Data and methods

2.1 Data collection and preprocessing

This cross-sectional diagnostic study consecutively enrolled adult inpatients admitted to the Department of Thyroid and Breast Surgery, The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital), between September 1, 2024 and June 1, 2025, who were suspected of having Yang deficiency constitution based on initial TCM consultation. Clinical tongue image data were collected for analysis. Inclusion criteria were as follows: (i) aged 18 – 75 years; (ii) admitted for thyroid nodules, differentiated thyroid cancer, early-stage breast cancer with stable post-surgical recovery, or benign breast masses, and in a non-critically ill, cooperative condition suitable for standardized tongue image acquisition; (iii) no antibiotic or glucocorticoid use within the preceding one month; (iv) no history of tongue surgery, oral mucosal disease, or other conditions affecting tongue morphology. Exclusion criteria were as follows: (i) patients with decompensated cardiovascular diseases [e.g., New York Heart Association (NYHA) class III – IV heart failure] or advanced diabetes with severe vascular complications; (ii) pregnant or lactating women; (iii) patients with cognitive impairment or psychiatric disorders unable to cooperate with the acquisition protocol; (iv) patients who had consumed food, beverages, or tobacco within 30 min prior to image acquisition that might stain the tongue coating; (v) patients with acute infection or systemic inflammatory response syndrome. The same equipment (Canon EOS 80D, 50 mm lens, 5 500 K ring flash) was used for all image acquisitions, and acquisition parameters were calibrated in accordance with the International Organization for Standardization (ISO) 20498-1: 2019 and ISO 20498-2: 2017 [24, 25]. This study was conducted in compliance with the Declaration of Helsinki. The research protocol and data collection procedures were reviewed and approved by the Ethics Committee of The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital) (Approval No. 2023132). All participants provided written informed consent before data collection, and all personal identification information was strictly anonymized during image processing and dataset construction.

During the preprocessing phase, bounding box annotations and category assignments were performed using Labellmg. The dataset was split into training, validation, and test sets at an 8 : 1 : 1 ratio.

2.2 Diagnostic criteria and tongue image annotation protocol

The tongue image classification system integrated classical TCM theory with modern constitutional characteristics,

including the root cause of cold-deficiency syndrome [26] and the modern sub-health state [27], together with patterns associated with the transition toward modern lifestyles [28]. Specific behavioral factors were identified as primary drivers of spleen-Yang impairment [29] and predictors of the decline of kidney Yang and the development of concurrent Yin-Yang deficiency syndromes [30].

To construct a dataset suitable for deep learning-based object detection, we conducted data categorization and processing on the highly granular descriptions of Yang deficiency tongue images based on the national standard GB/T 16751.2-1997 [31] and the textbook *Diagnostics of Traditional Chinese Medicine* [32]. Because some traditional subtypes overlap in textual definitions and lack distinct visual differences in images, they cannot be directly used as computer vision labels. Therefore, guided by the objective requirement of object detection models for clear and distinguishable visual features, we extracted descriptions with representative color and morphological characteristics (e.g., tooth marks and white slippery coating) from the standards. By filtering and merging the complex text-based subtypes, we clustered these TCM syndromes into five macro-categories with typical visual differences. These macro-categories served as the clinical diagnostic criteria for this study and were translated into a standardized visual phenotype annotation matrix (Table 1), thereby establishing objective and quantifiable classification standards.

During data labeling, two practitioners independently assessed the intensity of specific visual phenotypes using a 3-point quantitative scale (0 = none, 1 = mild, and 2 = severe); any discrepancies were resolved through discussion to reach a consensus. A syndrome label was assigned only when the cumulative score of its primary visual phenotypes met a predefined threshold τ , which was empirically determined on an independent annotation validation subset by maximizing the Youden index, converting qualitative descriptions into quantifiable parameters for model optimization.

To further illustrate these clinical phenotypes, representative tongue images for the five predefined syndrome subtypes are presented in Figure 1.

2.3 Quantitative criteria for syndrome label assignment

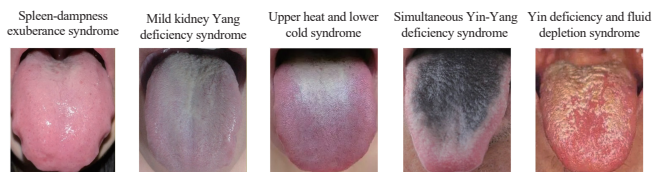
While GB/T 16751.2-1997 provides qualitative definitions for TCM syndromes, quantitative criteria are required for computer vision tasks. To establish these criteria, for any candidate syndrome k , the cumulative visual score $S_{\text{visual}}^{(k)}$ is calculated by summing the intensity scores $P_i \in \{0, 1, 2\}$ of its N primary phenotypes:

$$S_{\text{visual}}^{(k)} = \sum_{i=1}^N P_i^{(k)} \quad (1)$$

A sample is assigned a ground-truth label only if $S_{\text{visual}}^{(k)}$ it meets a strict threshold τ :

Table 1 Expert visual phenotype annotation matrix based on TCM national standards and textbook criteria

Syndrome category	Clinical definition	Primary symptom	Secondary symptom
Spleen-dampness exuberance syndrome	Enlarged tongue body; tooth marks along the margins; thick, white, greasy coating	(i) Marginal geometric indentations (tooth marks); (ii) dense white pixel/texture coverage indicating a thick greasy coating	(i) Expanded transverse tongue area; (ii) pale tongue color
Mild kidney Yang deficiency syndrome	Pale, enlarged, tender tongue; white, watery or slippery coating	(i) Reduced red-channel intensity/attenuation (pale or bluish tone); (ii) high specular reflection/glare consistent with a watery coating	(i) Mild swelling without deep tooth marks; (ii) soft tongue body
Upper heat and lower cold syndrome	Red tongue apex; white, slippery, or thick greasy coating on the root	(i) Localized red apex with focal red spots; (ii) white, thick, or slippery coating concentrated at the root	(i) Relatively normal body color; (ii) front-back color contrast
Simultaneous Yin-Yang deficiency syndrome	Pale or enlarged tongue body; dry tongue with scanty/peeled coating or fissures	(i) Globally pale tongue body color; (ii) scanty/peeled coating with deep fissures	(i) Thin or weak tongue body; (ii) reduced moisture texture
Yin deficiency and fluid depletion syndrome (negative control)	Red or crimson tongue; scanty or absent coating (mirror tongue); fissures	(i) Globally crimson/red tongue color; (ii) absence of coating or deep fissures	(i) Dry tongue surface; (ii) sharply defined fissure network

**Figure 1** Representative tongue images for the five TCM syndrome subtypes

$$\text{Label}^{(k)} = \begin{cases} 1 & \text{if } S_{\text{visual}}^{(k)} \geq \tau \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

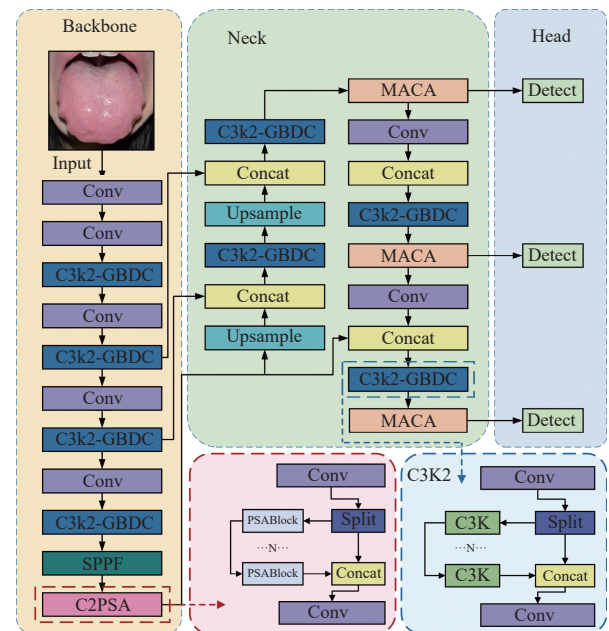
Each sample was assigned a single ground-truth label corresponding to the candidate syndrome with the highest cumulative visual phenotype score among those meeting the threshold τ . The classification task of this study was therefore a single-label five-category problem.

2.4 Proposed network architecture and design

The network architecture is designed to extract specific visual features. For example, a pale tongue is expressed as an attenuation of the red channel in the red, green, and blue (RGB) color space. Dynamic Convolution (Dynamic-Conv) is introduced to extract these low-contrast features through adaptive weight adjustments. A greasy coating is expressed as high-frequency texture density, which is captured by the multipath aggregation coordinate attention (MACA) attention mechanism through multi-scale position association. The proposed Yang deficiency YOLO (YD-YOLO) model integrates the C3k2-GhostBottleneck-Dynamic Convolution (C3k2-GBDC) module into the backbone network and the MACA module into the neck to suppress background noise (Figure 2).

2.5 MACA module

Integrating an attention mechanism at the end of the backbone is effective for distinguishing tongue features

**Figure 2** Improved YOLO11 network structure

Conv, convolution. BN, batch normalization. Concat, concatenation.

from complex backgrounds. Commonly used channel attention mechanisms, such as Squeeze-and-Excitation (SE) [33] and Convolutional Block Attention Module (CBAM) [34], have limitations in capturing long-range spatial dependencies. Therefore, this study introduces the MACA mechanism, which considers multi-scale position information to reduce background interference (Figure 3).

Average pooling was applied along the height and width dimensions to generate direction-aware feature maps, preserving the spatial positions of tongue textures. Global average pooling was then used to strengthen semantic channel dimensions, such as tongue color. Finally, direction-aware features were integrated with global semantic features through feature interaction.

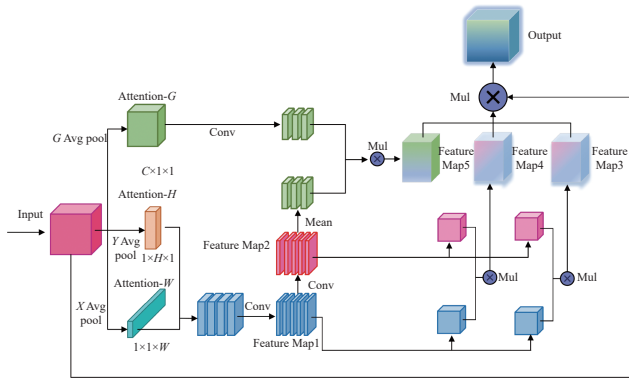


Figure 3 Block diagram of the MACA module

GAP, global average pooling. GMP, global max pooling. FC, fully connected layer. Mul, element-wise multiplication. G Avg pool, global average pooling.

2.6 C3k2-GBDC module

GhostBottleneck and DynamicConv were integrated into the C3k2 module to create the C3k2-GBDC module. This module adaptively adjusts convolutional kernel weights based on input data across morphological variations. Compared with static convolutions, DynamicConv (Figure 4A) enhances feature representation as follows [35]:

$$y = g(\tilde{W}^T x + \tilde{b}), \quad \tilde{W} = \sum_{k=1}^K p_k(x) \tilde{W}_k \quad (3)$$

Here, the input feature x generates dynamic kernel parameters via an attention mechanism (including global average pooling, fully connected mapping, and softmax normalization) [36]. The Ghost module generates feature maps via linear transformations (Figure 4B), and the

GhostBottleneck leverages these mechanisms for feature alignment [37] (Figure 4C). The combined C3k2-GBDC module is shown in Figure 4D.

2.7 Experimental setup

The specific configurations of the experimental environment and the key hyperparameter settings for model training are detailed in Table 2 and 3, respectively.

To address subgroup imbalance and prevent data leakage, data augmentation was applied exclusively to the training set after partitioning. Methods such as MixUp and random rotation were excluded because they could generate nonphysiological artifacts and disrupt the spatial topology of tongue visceral mapping. Instead, Gaussian blur and translation were adopted to increase data diversity while maintaining spatial semantics. In addition, color jitter (adjusting brightness, contrast, and saturation) was applied to simulate varying clinical lighting conditions. To evaluate the biological plausibility of the YD-YOLO model's focus, gradient-weighted class activation mapping (Grad-CAM) was used. By computing the gradients of the target syndrome category with respect to the feature maps in the final neck layer, the system generated heatmaps that visualized the contribution of specific pixel regions to the diagnostic outcome. This methodological setup allowed for a direct comparison between the model's geometric attention and the diagnostic focus areas defined by senior TCM practitioners. To evaluate the contributions of each component, DynamicConv was placed in the backbone to extract low-level textural features, while MACA was embedded in the neck to aggregate semantic and spatial position information.

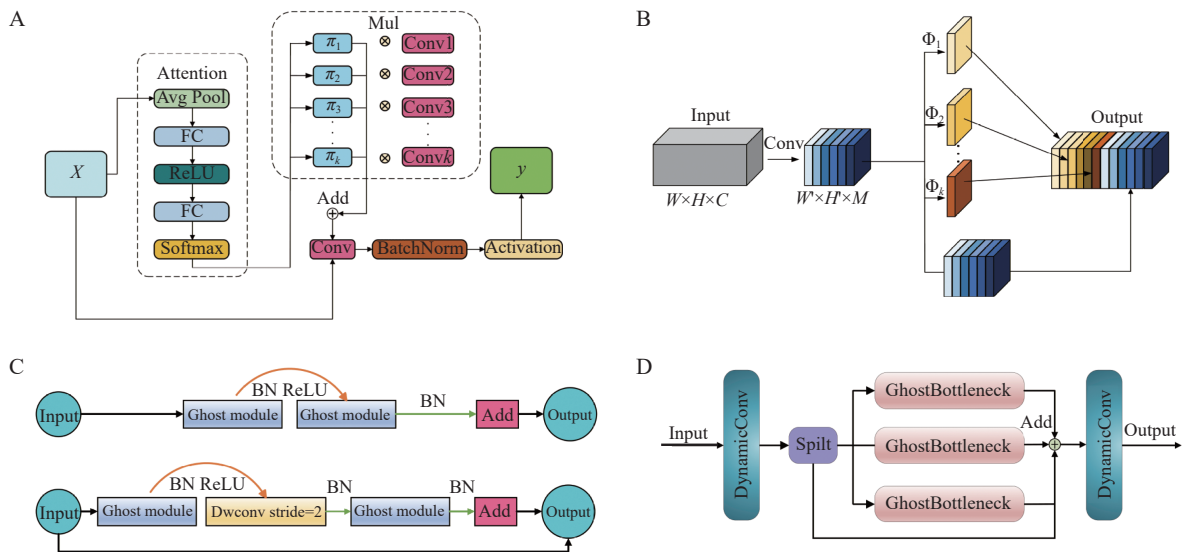


Figure 4 Architecture of the C3k2-GBDC module and its components

A, block diagram of the DynamicConv module. B, block diagram of the Ghost module. C, architecture of the GhostBottleneck module (upper: stride = 1; lower: stride = 2). D, architecture of the C3k2-GBDC module. DWConv, depthwise convolution. PWConv, pointwise convolution. Add, element-wise addition. BN, batch normalization.

Table 2 Experimental environment configuration of the deep learning models

Configuration item	Detail
Operating system	Windows 10
Central processing unit (CPU)	Intel Core i5-12600KF
Random access memory (RAM)	32.0 GB
Deep learning framework	PyTorch 2.0.0 (CUDA 12.2)
Programming language	Python 3.9.19
Graphics processing unit (GPU)	NVIDIA GeForce RTX 3060 (12 GB VRAM)

Table 3 Training parameter settings of the deep learning models

Hyperparameter	Value
Data augmentation strategy	Color jitter, Gaussian blur, translation
Learning rate schedule	Cosine annealing (initial 0.01 $\rightarrow$ 0.0001)
Optimization algorithm	Stochastic gradient descent (SGD)
Total training epochs	300
Batch size	32
Input resolution	640 $\times$ 640 pixels

2.8 Evaluation metrics and statistical analysis

All statistical analyses were performed using Python 3.9 and SPSS 26.0, with a two-sided significance level set at $\alpha = 0.05$. Continuous variables were expressed as mean $\pm$ standard deviation (SD) and were compared across the five syndrome groups using one-way analysis of variance (ANOVA). Categorical variables, including sex distribution and the prevalence rates of individual visual phenotypes, were presented as frequencies and percentages, and group comparisons were conducted using the Chi-square (χ^2) test. The reliability of the expert visual phenotype annotation protocol was evaluated on a randomly selected subset of 200 cases: discriminative ability was assessed via the area under the receiver operating characteristic (ROC) curve (AUC), the optimal threshold (τ) for label assignment was determined by maximizing the Youden index, and inter-rater agreement was quantified with Cohen's kappa coefficient.

Model performance was evaluated on two distinct levels. Detection-level performance—used for the ablation study, the attention-mechanism comparison, and the benchmarking experiments—was assessed using intersection over union (IoU), precision, recall, F1 score, and mean average precision (mAP). Classification-level performance—used to evaluate the model's overall diagnostic reliability on the test set—was assessed using overall accuracy, AUC, and Cohen's kappa (κ). Inference speed was recorded in frames per second (FPS).

The detection metrics were calculated based on the standard definitions of true positive (TP), false positive

(FP), false negative (FN), and true negative (TN). True positive rate (TPR; also termed recall) is the proportion of actual positives correctly identified, and false positive rate (FPR) is the proportion of actual negatives incorrectly classified as positive:

$$\text{IoU} = \frac{\text{Area}(P) \cap \text{Area}(G)}{\text{Area}(P) \cup \text{Area}(G)} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

$$\text{mAP} = \frac{1}{n} \sum_{i=1}^n \text{AP}_i \quad (8)$$

For the classification task, overall accuracy, AUC, and Cohen's kappa (κ) were evaluated:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (9)$$

$$\text{AUC} = \int_0^1 \text{TPR} \, d(\text{FPR}) \quad (10)$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (11)$$

To evaluate the proposed YD-YOLO architecture, three sets of experiments were designed: (i) to conduct an ablation study to quantify the individual and synergistic contributions of the proposed modules evaluating configurations including baseline YOLO11 nano (YOLO11n), YOLO11n + MACA, YOLO11n + C3k2-GBDC, and YOLO11n + C3k2-GBDC + MACA (YD-YOLO); (ii) to assess comparative performance in the medical domain, YD-YOLO was benchmarked against several state-of-the-art detectors including typical YOLO iterations such as YOLOv5, v8, v10, YOLO-Lite, and Swin-Transformer; and (iii) to validate the specificity of MACA in capturing subtle TCM phenotypes, a controlled comparison was conducted against mainstream attention modules (SE and CBAM) under identical hyperparameter settings. In addition, the clinical utility of the YD-YOLO framework was evaluated through an interactive prototype system. The evaluation focused on the system's ability to translate digital confidences into interpretable TCM prescriptions (Li-Fa-Fang-Yao, 理法方药, principles, methods, formulas, and medicines), verifying the practical link between deep learning outputs and clinical decision support.

3 Results

3.1 Dataset distribution and sample characteristics

A total of 1 221 clinical cases were initially enrolled, with one standardized tongue image acquired per patient

(1 221 images in total). Of these, 35 images were excluded primarily due to inadequate image quality (e.g., over-exposure, shadow artifacts, and focus blur), failure to meet the standard 5 500 K illumination criteria, or the presence of mixed syndromic presentations that could not be unequivocally assigned to a single Yang deficiency subtype. A final dataset of 1 186 clinical cases (1 186 high-quality tongue images) was included for model development. The final dataset comprised 320 cases of spleen-dampness exuberance syndrome, 315 cases of mild kidney Yang deficiency syndrome, 280 cases of upper heat and lower cold syndrome, 135 cases of simultaneous Yin-Yang deficiency syndrome, and 136 cases of Yin deficiency and fluid depletion syndrome (negative control group) (Table 4).

3.2 Validation of the annotation protocol

Before model training, the reliability of the annotation protocol was statistically validated on a random subset of 200 cases drawn from the complete dataset. The scoring protocol achieved an AUC of 0.92 (Figure 5A), demonstrating that the cumulative visual phenotype scores could effectively discriminate syndrome-positive cases of Yang deficiency syndrome subtypes from controls. The optimal threshold for label assignment was determined as $\tau = 2$ via the Youden index (Figure 5B), confirming that a minimum cumulative score of 2 across primary visual phenotypes was sufficient to trigger a subtype label. Inter-rater agreement between the two annotators was strong, with Cohen’s kappa of 0.83 (Figure 5C), indicating that the annotation criteria were applied consistently. Furthermore, the score distributions (Figure

5D) showed clear structural separability between Yang deficiency syndrome subtypes and negative controls. Taken together, these results confirmed that the proposed annotation framework produces reliable and reproducible ground-truth labels suitable for model training.

3.3 Ablation analysis of MACA and C3k2-GBDC modules for performance enhancement

The ablation study results (Table 5 and Figure 6) showed that each module independently contributed to performance gains. Adding MACA alone increased mAP@50 from 0.681 (baseline YOLO11n) to 0.714, while adding C3k2-GBDC alone raised it to 0.702. The YD-YOLO achieved the highest mAP@50 of 0.731 and mAP@50–95 of 0.647, representing improvements of 5.0 and 3.5 percentage points over the baseline, respectively, at an inference speed of 89.00 FPS.

The proposed MACA mechanism with the YOLO11n + C3k2-GBDC baseline achieved an mAP@50 of 0.731, outperforming SE (0.711), and CBAM (0.718) under identical baseline conditions, with a minimal increase in parameters (3.37 vs. 3.29 and 3.31 M) and computational cost (8.5 vs. 7.3 and 7.4 GFLOPs) (Table 6). The YD-YOLO model achieved an mAP@50 of 0.731 and an mAP@50–95 of 0.647, and the inference speed was 89 FPS.

Figure 7 further demonstrates that MACA yielded a more balanced performance across five TCM syndrome subtypes, with larger gains particularly in the minority classes (Yin deficiency and fluid depletion syndrome and spleen-dampness exuberance syndrome), indicating improved robustness to class imbalance compared with SE and CBAM.

Table 4 Baseline demographics and core visual phenotypes of the study population

Characteristic		Total	Spleen-dampness exuberance syndrome	Mild kidney Yang deficiency syndrome	Upper heat and lower cold syndrome	Simultaneous Yin-Yang deficiency syndrome	Yin deficiency and fluid depletion syndrome	P value
Demographics	Age (year)	46.5 ± 12.4	42.1 ± 10.5	51.3 ± 11.2	39.8 ± 9.4	58.6 ± 10.1	48.2 ± 11.5	< 0.001
	Sex	Male	545 (46.0%)	158 (49.4%)	138 (43.8%)	132 (47.1%)	65 (48.1%)	0.124
		Female	641 (54.0%)	162 (50.6%)	177 (56.2%)	148 (52.9%)	70 (51.9%)	
	Body mass index (kg/m ²)	23.8 ± 3.5	26.2 ± 3.1	24.1 ± 2.8	22.5 ± 2.6	21.8 ± 2.4	20.4 ± 2.1	< 0.001
Core visual phenotypes	Marginal tooth marks	548 (46.2%)	305 (95.3%)	182 (57.8%)	45 (16.1%)	16 (11.9%)	0 (0.0%)	< 0.001
	Thick/greasy coating	552 (46.5%)	298 (93.1%)	42 (13.3%)	205 (73.2%)	5 (3.7%)	2 (1.5%)	< 0.001
	Pale color	526 (44.3%)	115 (35.9%)	295 (93.7%)	12 (4.3%)	104 (77.0%)	0 (0.0%)	< 0.001
	Red apex with focal spots	395 (33.3%)	12 (3.8%)	5 (1.6%)	268 (95.7%)	15 (11.1%)	95 (69.9%)	< 0.001
	Scanty coating/fissures	248 (20.9%)	0 (0.0%)	8 (2.5%)	0 (0.0%)	115 (85.2%)	125 (91.9%)	< 0.001

Continuous variables are presented as mean ± SD (one-way ANOVA), and categorical variables as *n* (%) (Chi-square test).

3.4 Benchmarking against state-of-the-art detection models

The comparative performance of the YD-YOLO model against other architectures is detailed in Table 7 and 8.

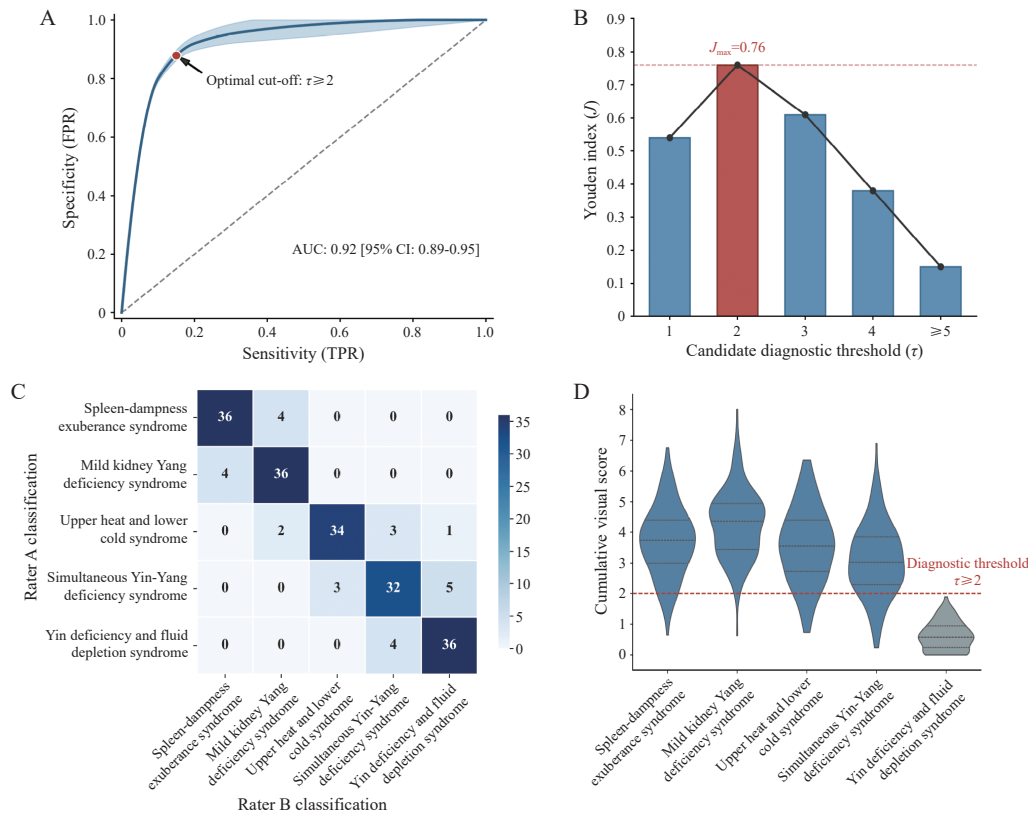


Figure 5 Statistical validation of the visual phenotype annotation protocol
A, ROC curve. B, optimal threshold determination via the Youden index. C, inter-rater reliability (Cohen’s kappa = 0.83, $P < 0.001$). D, score distribution by case status.

Table 5 Ablation experiments of YOLO11n with incremental modules

Model	Params (M)	GFLOPs	mAP@50	mAP@50-95	Speed (FPS)
YOLO11n	2.58	6.3	0.681	0.612	158
YOLO11n + MACA	2.69	7.6	0.714	0.635	123
YOLO11n + C3k2-GBDC	3.26	7.2	0.702	0.629	110
YD-YOLO	3.37	8.5	0.731	0.647	89

GFLOPs, giga floating-point operations.

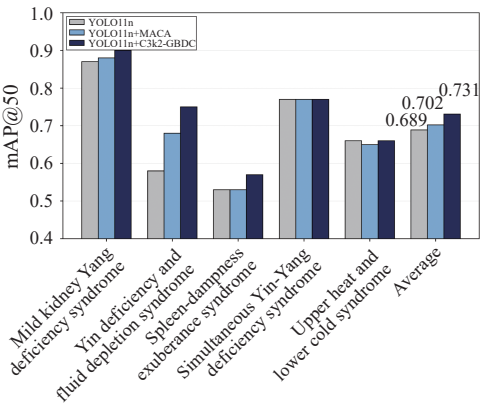


Figure 6 Trend visualization of ablation experiment of YOLO11n with incremental modules independently

Table 6 Ablation study of attention mechanisms on the YOLO11n + C3k2-GBDC baseline

Attention module	Params (M)	GFLOPs	mAP@50	mAP@50-95	Speed (FPS)
YOLO11n + C3k2-GBDC	3.26	7.2	0.702	0.629	110.00
YOLO11n + C3k2-GBDC + SE	3.29	7.3	0.711	0.634	105.00
YOLO11n + C3k2-GBDC + CBAM	3.31	7.4	0.718	0.638	101.00
YOLO11n + C3k2-GBDC + MACA (YD-YOLO)	3.37	8.5	0.731	0.647	89.00

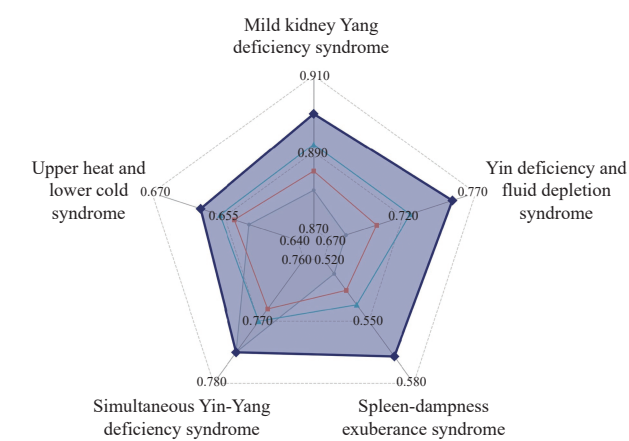


Figure 7 Radar chart comparison of different attention mechanisms

Table 7 Comparative performance analysis of YOLO series models across syndrome subtypes

Model	Metric	Mild kidney Yang deficiency syndrome	Yin deficiency and fluid depletion syndrome	Spleen-dampness exuberance syndrome	Simultaneous Yin-Yang deficiency syndrome	Upper heat and lower cold syndrome	Mean
YOLOv5	Precision	0.671	0.685	0.407	0.469	0.489	0.544
	Recall	0.842	0.625	0.621	0.632	0.700	0.684
	F1 score	0.747	0.654	0.492	0.538	0.576	0.601
	mAP@50	0.846	0.681	0.555	0.703	0.614	0.680
	mAP@50–95	0.782	0.636	0.497	0.626	0.541	0.616
YOLOv8	Precision	0.595	0.632	0.517	0.681	0.323	0.550
	Recall	0.965	0.625	0.664	0.675	0.500	0.686
	F1 score	0.736	0.628	0.581	0.678	0.392	0.603
	mAP@50	0.877	0.572	0.510	0.728	0.651	0.667
	mAP@50–95	0.804	0.518	0.425	0.627	0.575	0.590
YOLO11n	Precision	0.756	0.632	0.385	0.661	0.728	0.632
	Recall	0.825	0.500	0.621	0.737	0.384	0.613
	F1 score	0.789	0.558	0.475	0.697	0.503	0.604
	mAP@50	0.872	0.580	0.527	0.774	0.654	0.681
	mAP@50–95	0.804	0.528	0.461	0.682	0.574	0.610
YD-YOLO	Precision	0.724	0.947	0.545	0.636	0.785	0.738
	Recall	0.825	0.699	0.621	0.632	0.800	0.715
	F1 score	0.771	0.823	0.581	0.634	0.792	0.720
	mAP@50	0.900	0.754	0.567	0.775	0.660	0.731
	mAP@50–95	0.838	0.667	0.501	0.661	0.571	0.647

Table 8 Results of comparative experiments on object detection models

Model	Precision	Recall	F1	mAP@50	mAP@50–95	Size (MB)	GFLOPs	FPS
YOLO-Lite	0.575	0.580	0.577	0.615	0.530	3.5	3.8	190
MobileNetV3-YOLO	0.590	0.605	0.597	0.642	0.565	4.8	4.2	175
YOLOv5	0.544	0.684	0.601	0.679	0.616	14.2	15.8	112
YOLOv8	0.550	0.686	0.603	0.669	0.590	6.2	8.1	140
Faster R-CNN	0.620	0.740	0.675	0.712	0.615	137.0	215.3	15
Detection Transformer (DETR)	0.660	0.760	0.706	0.748	0.665	165.0	152.0	34
Swin-Transformer	0.700	0.790	0.742	0.755	0.682	112.0	245.2	24
YOLO11n	0.632	0.613	0.604	0.681	0.612	5.2	6.3	158
YD-YOLO	0.738	0.715	0.720	0.731	0.647	10.2	8.5	89

89.00 FPS. Compared with large-scale networks, Swin-Transformer had an mAP@50 of 0.755 but requires greater computational resources. The YD-YOLO model balances recognition accuracy and deployment requirements.

3.5 Classification performance and reliability of the YD-YOLO model

The YD-YOLO model achieved an overall accuracy of 91.5% and a Cohen’s kappa coefficient of 0.912 on the test set (Figure 8A). The ROC analysis further confirmed the model’s discriminative performance, with AUC ranging from 0.91 to 0.97 across five syndrome subtypes (Figure 8B).

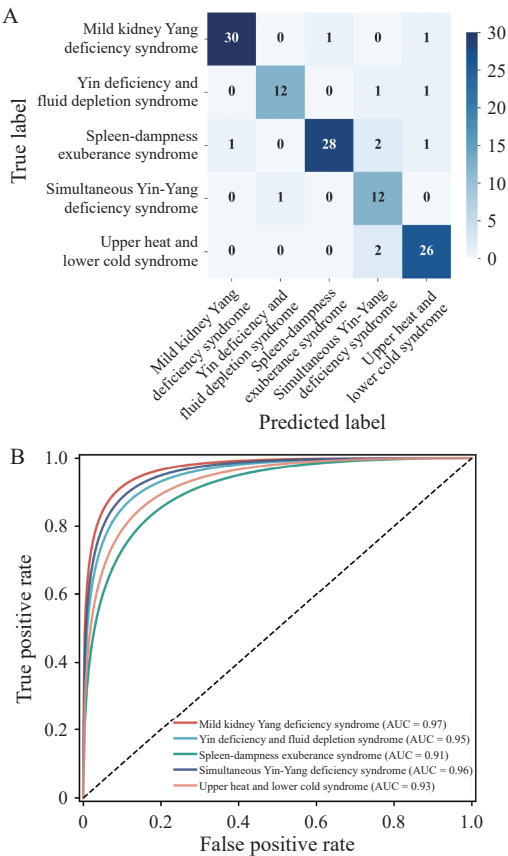


Figure 8 Evaluation of classification performance of the YD-YOLO model
A, confusion matrix. B, ROC curves.

3.6 Feature attribution and visual interpretability using Grad-CAM

Figure 9 comprehensively illustrates the Grad-CAM results for both the ablation study and the comparison of different algorithms. For the spleen-dampness exuberance syndrome subtype, the YOLO11n model activation was distributed across the tongue center, whereas the integration of MACA localized activation to marginal tooth marks. In the mild kidney Yang deficiency syndrome subtype, C3k2-GBDC increased activation in the low-contrast root coating compared with the baseline. In the upper heat and lower cold syndrome subtype, the YD-YOLO model showed activation on both the red apex and the pale root, consistent with the pathology of Xinshen Bu-jiao (心肾不交, heart-kidney non-interaction). Furthermore, when comparing different algorithms, MobileNetV3-YOLO showed activation covering background areas. Swin-Transformer showed broader activation regions. The YD-YOLO model localized activation to specific pathological features while reducing background noise.

4 Discussion

This study establishes a framework for the fine-grained identification of Yang deficiency syndrome subtypes,

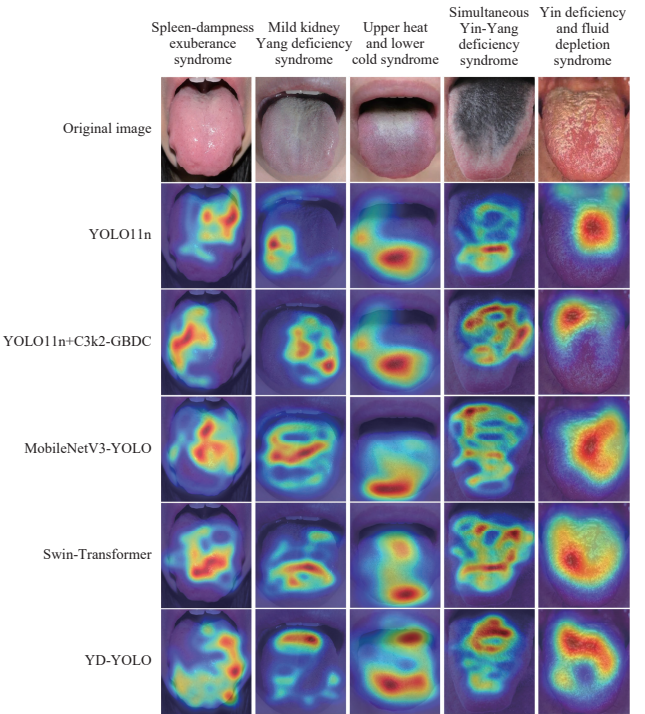


Figure 9 Comparative Grad-CAM visualization across five syndrome subtypes

translating macroscopic clinical descriptions into quantifiable visual targets. While previous studies have examined AI-assisted tongue diagnosis [38, 39], most have focused on binary classification (e.g., healthy and diseased) or basic global features such as overall tongue color. In contrast, our approach addresses the long-standing clinical challenge of differentiating subphenotypes within a specific TCM syndrome. The integration of the MACA and C3k2-GBDC modules enhances feature extraction. Traditional CNNs often struggle with background interference and subtle textural variations in complex medical images [40, 41].

To translate these visual features into clinical utility, a prototype decision-support system was developed based on the TCM principle of Li-Fa-Fang-Yao. The system incorporates two core modules: a visual anchoring module that localizes pathological features using bounding boxes and outputs syndrome subtype probabilities as radar-chart confidence scores (Figure 10), and a targeted prescription association module that maps the model's classification output to corresponding baseline formulas, with fine-grained visual features linked to specific drug additions (Figure 11). For instance, detecting prominent tooth marks in a case classified as spleen-dampness exuberance syndrome triggers a recommendation to add Huangqi (Astragali Radix) and Fuling (Poria). The system is designed strictly as an auxiliary tool, and TCM physicians retain full authority to review, adjust, and finalize all prescription components rather than relying solely on the automated suggestions.

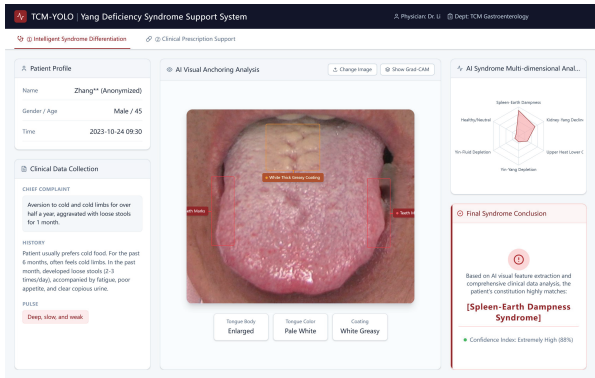


Figure 10 Intelligent discrimination workbench view

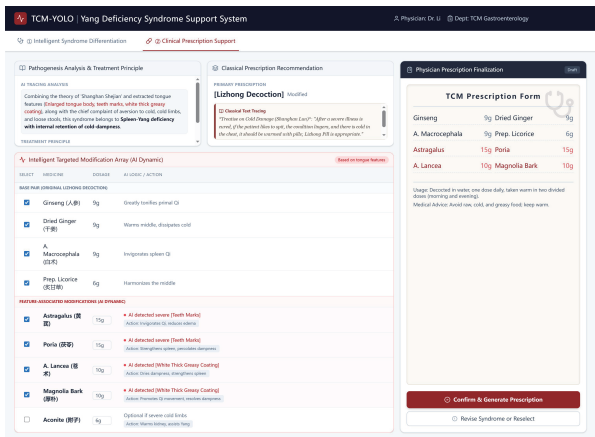


Figure 11 Auxiliary prescription decision-making view

The ablation results and Grad-CAM visualizations provide direct evidence for the contribution of each module. The MACA mechanism suppressed oral background noise by 9.1% in mAP@50 over the baseline (0.714 vs. 0.681), and the corresponding heatmaps confirmed that activation was redirected from non-diagnostic regions to clinically relevant features such as marginal tooth marks and thick coating regions. The C3k2-GBDC module, meanwhile, adaptively enhanced low-contrast signals, specifically the R-channel attenuation in pale tongues characteristic of mild kidney Yang deficiency, yielding an additional 1.7% mAP@50 gain when combined with MACA. Together, these modules enabled the YD-YOLO model to achieve an overall accuracy of 91.5%, with a Cohen’s kappa of 0.912 and AUC values of 0.91 – 0.97 across five syndrome subtypes, indicating both high discriminative power and strong agreement with expert consensus.

Beyond raw performance, the clinical value of this framework lies in its interpretability. Deep learning models are often perceived as “black boxes” [42], which limits their acceptance in clinical settings. Here, the Grad-CAM analysis demonstrated that the model’s attention maps consistently aligned with the anatomical regions prioritized by TCM practitioners during diagnosis (e.g., the tongue margin for tooth marks and the root for coating assessment), providing visual evidence that the model’s

decisions are based on biologically plausible features. This transparency addresses a key barrier to clinical trust. Additionally, by embedding the quantitative visual phenotype criteria into a standardized pipeline (from image acquisition to syndrome classification), the framework reduces inter-practitioner variability and provides a reproducible diagnostic reference that can be deployed across different hospitals, serving also as a structured learning scaffold for junior clinicians.

Despite these results, the applicability and limitations of this study must be acknowledged. First, regarding sample complexity, patients presenting with overlapping phenotypes or complex concurrent syndromes were excluded during this stage to establish a clear baseline. The model’s diagnostic reliability in real-world scenarios with multi-syndrome presentations requires validation. Second, the current evaluation relies on images captured under standard clinical lighting. The model’s robustness to varying illumination conditions is a known limitation [43]. Future research should develop cross-device spectrum offset and color-correction algorithms to enable reliable deployment across diverse mobile devices. Finally, prospective multicenter cohort studies are needed to further validate the clinical effectiveness and long-term diagnostic stability of the proposed system.

5 Conclusion

This study develops a fine-grained recognition framework for Yang deficiency syndrome subtypes by integrating an improved YOLO11 architecture with TCM clinical expertise. Translating the macroscopic descriptions in national standards (GB/T 16751.2-1997) into quantifiable visual phenotype matrices may reduce subjectivity in syndrome differentiation. Using the MACA mechanism and the C3k2-GBDC module, the model achieves an mAP@50 of 0.731 and an overall diagnostic accuracy of 91.5% across five syndrome subtypes. The interactive auxiliary system links deep learning outputs to clinical decision-making, providing interpretable visual evidence and prescription references. This system may offer an objective model for TCM diagnosis, reducing inter-practitioner variability and improving clinical consistency.

Fundings

Project of Outstanding Top-notch Talent Cultivation in Anhui Colleges (gxbjZD2022023), National Training Program of Innovation and Entrepreneurship for Undergraduates (202510363014), and Key Scientific Research Project of Anhui Provincial Department of Education (2024AH051951).

Ethical statement

The research protocol and data collection procedures were ethically reviewed and approved by the Ethics Committee of The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital) (Approval No. 2023132). All participants provided written informed consent prior to enrollment, and the study was conducted in strict accordance with the Declaration of Helsinki.

Author contributions

Tongbin Zhang: conceptualization, methodology, software, validation, visualization, and writing – original draft preparation. Haoran Xu: data curation and formal analysis. Ziyi Wang: data curation and investigation. Chuanjun Pan: formal analysis and validation. Zheng Wang: resources, funding acquisition, supervision, and investigation. Lei Wang: funding acquisition, project administration, supervision, and writing – review and editing. All authors approved the submission and take responsibility for this manuscript.

Competing interests

The authors declare no conflict of interest.

References

- [1] NIE H. Preliminary study on academic achievements and clinical significance of “Shanghan shejian”. *Forum on Traditional Chinese Medicine*, 1989(4): 34–36.
- [2] WU YH, WANG J, BAI MH, et al. Prevalence of traditional Chinese medicine body constitutions in a large community-based study in Hangzhou, China. *Chinese Medicine*, 2025, 20(1): 206.
- [3] LI YS, LI Y. Progress in the study of Yang-deficiency constitution in terms of traditional Chinese medicine: a narrative review. *Journal of Traditional Chinese Medicine*, 2023, 43(2): 409–416.
- [4] ZHANG ZJ. *Synopsis of the golden chamber*. Beijing: People’s Medical Publishing House, 2005.
- [5] WU Z, YANG B. Research on Chen Xiuyuan’s academic thought of “six meridian gasification” and annotation of “treatise on febrile diseases”. *Journal of Traditional Chinese Medical Literature*, 2024, 42(5): 55–58.
- [6] ANDRÉS E, ESCOBAR C, DOI K. Machine learning and artificial intelligence in clinical medicine: trends, impact, and future directions. *Journal of Clinical Medicine*, 2025, 14(22): 8137.
- [7] LIU Q, LI Y, YANG P, et al. A survey of artificial intelligence in tongue image for disease diagnosis and syndrome differentiation. *Digital Health*, 2023. doi: 10.1177/20552076231191044.
- [8] XU JT, JIANG T, LIU S. Research status and prospect of tongue image diagnosis analysis based on machine learning. *Digital Chinese Medicine*, 2024, 7(1): 3–12.
- [9] ZHONG LQ, XIN GJ, PENG QH, et al. Deep learning-based recognition of stained tongue coating images. *Digital Chinese Medicine*, 2024, 7(2): 129–136.
- [10] ZHANG YY, WEI DS, ZHANG Y, et al. Quantitative research on tongue diagnosis in traditional Chinese medicine for obstructive coronary artery disease: a computational analysis based on multimodal feature fusion. *Digital Chinese Medicine*, 2025, 8(4): 443–454.
- [11] RAGAB MG, ABDULKADIR SJ, MUNEEER A, et al. A comprehensive systematic review of YOLO for medical object detection (2018 to 2023). *IEEE Access*, 2024, 12: 57815–57836.
- [12] REDMON J, FARHADI A. YOLO9000: better, faster, stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
- [13] JIANG T, LU Z, HU XJ, et al. Deep learning multi-label tongue image analysis and its application in a population undergoing routine medical checkup. *Evidence-Based Complementary and Alternative Medicine*, 2022, 2022: 3384209.
- [14] GUO ZL, LIU X, LIU WJ. Research on multi-classification of TCM tongue symptoms based on self-supervised learning. 2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE, 2023.
- [15] WU X, XU H, LIN Z. Review of research on deep learning in tongue image classification. *Journal of Computer Engineering and Applications*, 2023, 17(2): 303–323.
- [16] CHEN TN, CHEN YT, ZHOU ZL, et al. Deep learning-based automated tongue analysis system for assisted Chinese medicine diagnosis. *Frontiers in Physiology*, 2025, 16: 1559389.
- [17] LIU HH, ZHANG PQ, HUANG YN, et al. Research on multi-label recognition of tongue features in stroke patients based on deep learning. *Scientific Reports*, 2024, 14: 32144.
- [18] LI KW, QIU ZH, LI JL, et al. Deep learning approach for objective differentiation of kidney deficiency syndrome in reproductive age females: a tongue-face fusion model. *Frontiers in Physiology*, 2025, 16: 1701545.
- [19] NING ZH, WANG YH, WANG L, et al. Research on a vehicle-mounted intelligent TCM syndrome differentiation system based on deep belief network. *IEEE Access*, 2021, 9: 118019–118030.
- [20] KHANAM R, HUSSAIN M. YOLOv11: an overview of the key architectural enhancements. *arXiv*, 2024. doi: 10.48550/arXiv.2410.17725.
- [21] LUI MH, LIU HX, TANG ZC, et al. An adaptive YOLO11 framework for the localisation, tracking, and imaging of small aerial targets using a pan-tilt-zoom camera network. *Engineering*, 2024, 5(4): 3488–3516.
- [22] CHANG WH, CHEN CC, WU HK, et al. Tongue feature dataset construction and real-time detection. *PLoS One*, 2024, 19(3): e0296070.
- [23] QU ZH, LIU HY, KONG WG, et al. LP-YOLO: an improved lightweight pedestrian detection algorithm based on YOLOv11. *Digital Signal Processing*, 2025, 165: 105343.
- [24] International Organization for Standardization. *Traditional Chinese medicine: computerized tongue image analysis system (part 1): general requirements*. Geneva: International Organization for Standardization, 2019.
- [25] International Organization for Standardization. *Traditional*

- Chinese medicine: computerized tongue image analysis system (part 2): light environment. Geneva: International Organization for Standardization, 2017.
- [26] TIAN D. Inner Canon of Huangdi: Plain Questions. Beijing: People's Medical Publishing House, 2017.
- [27] ZHANG J. Complete Works of Jingyue. Beijing: People's Medical Publishing House, 2017.
- [28] WANG Q. Classification and diagnosis basis of nine basic constitutions in chinese medicine. *Journal of Beijing University of Traditional Chinese Medicine*, 2005, 28(4): 1-8.
- [29] HSU MF, TANG PL, PAN TC, et al. Different traditional Chinese medicine constitution is associated with dietary and lifestyle behaviors among adults in Taiwan. *Medicine*, 2022, 101(39): e30692.
- [30] HUANG BZ, ZHOU MJ, SONG SS, et al. Relationship between traditional Chinese medicine body constitution and sleep quality among high-speed railway crew in Beijing, China: a cross-sectional study. *Medicine*, 2025, 104(37): e44563.
- [31] State Bureau of Technical Supervision. Clinic terminology of traditional Chinese medical diagnosis and treatment (part 2): syndromes (GB/T 16751.2-1997). Beijing: Standardization Administration of the People's Republic of China, 1997.
- [32] LI C, WANG Y, ZHANG L. Traditional Chinese Medicine Diagnostics. 10th ed. Beijing: China Press of Traditional Chinese Medicine, 2021.
- [33] HU J, SHEN L, SUN G. Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.
- [34] WOO S, PARK J, LEE JY, et al. CBAM: convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2018: 3-19.
- [35] LIU HR, YU ZZ, XU XW, et al. Global attention augmentation ghost module: more features from lightweight global attention extraction. 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2021.
- [36] YADAV RK, NIGAM N, SINGH DP, et al. Improving vehicle detection using ghost convolution and GhostBottleneck layers in YOLOv5s. 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES). IEEE, 2023.
- [37] HAN K, WANG YH, TIAN Q, et al. GhostNet: more features from cheap operations. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
- [38] MA C, GAO C, HOU S, et al. Development of attention-based robust deep learning model for tongue diagnosis by smart-phone. *bioRxiv*, 2023. doi: 10.1101/2023.02.02.526804.
- [39] LIU B, WANG ZY, YU K, et al. Research on tongue image segmentation and classification methods based on deep learning and machine learning. *Information*, 2025, 16(5): 357.
- [40] MANZARI ON, KALEYBAR JM, SAADAT H, et al. BFUnet: a hybrid CNN-transformer architecture for precise medical image segmentation. *arXiv*, 2024. doi: 10.48550/arXiv.2402.08793.
- [41] ZHANG C, WANG LT, WEI GH, et al. A dual-branch and dual attention transformer and CNN hybrid network for ultrasound image segmentation. *Frontiers in Physiology*, 2024, 15: 1432987.
- [42] MARCHI G. Decoding the "black-box": explainable artificial intelligence towards trustworthy advancement in respiratory medicine. *Breathe*, 2026, 22(1): 250318.
- [43] OUYANG HF, YANG JG, NING XL, et al. Bridging accessibility and precision: evaluating the reliability and validity of a smart-phone-based skin colorimeter. *Clinical, Cosmetic and Investigational Dermatology*, 2026, 19: 1-10.

(Editor-in-Charge Siyi Wei)

基于深度学习的中医阳虚舌象亚型识别

张桐彬^a, 徐浩然^a, 王紫益^a, 潘传军^a, 汪正^b, 王雷^{a*}

a. 安徽工程大学机械与汽车工程学院, 安徽 芜湖 241000, 中国

b. 皖南医科大学第一附属医院(弋矶山医院) 甲乳外科, 安徽 芜湖 241001, 中国

【摘要】目的 针对中医阳虚证亚型临床识别颗粒度不足以及传统目标检测模型在提取不规则、低对比度舌象表型方面的局限性, 本研究旨在构建一种基于改进 YOLO11n 架构的客观识别框架, 通过标准化视觉表型矩阵将宏观中医描述转化为可量化的临床目标。**方法** 本横断面诊断研究连续纳入了 2024 年 9 月 1 日至 2025 年 6 月 1 日期间, 于皖南医科大学第一附属医院(弋矶山医院) 甲乳外科住院、经初步中医会诊为阳虚体质的成年患者, 并收集其临床舌象图像数据用于分析。基于专家视觉表型标注矩阵, 构建了包含以下中医证候亚型的五分类识别体系: 脾虚湿盛证、轻度肾阳虚证、上热下寒证、阴阳两虚证, 以及作为阴性对照的阴虚津亏证。所提出的阳虚识别 YOLO 模型(YD-YOLO)以 YOLO11n 为基线, 在主干网络中集成跨阶段部分卷积与核大小 2 (C3k2)-幽灵瓶颈-动态卷积(GBDC)模块以自适应提取低对比度特征, 并在颈部网络嵌入多路径聚合坐标注意力(MACA)机制, 通过多尺度空间协调抑制背景干扰。采用梯度加权类激活映射(Grad-CAM)实现特征归因可视化, 并评估模型关注区域的生物学合理性。采用消融实验和对比实验评估模型性能, 评价指标包括平均精度均值(mAP)、精确率、召回率、F1 值、推理速度(FPS)、总体准确率、Cohen's kappa 系数及受试者工作特征(ROC)曲线下面积(AUC)。**结果** 研究最终纳入 1186 例临床病例, YD-YOLO 模型总体准确率为 91.5%, Cohen's kappa 系数为 0.912, mAP@50 达 0.731 (优于基线 YOLO11n 模型的 0.681), 各中医证候亚型的 AUC 范围为 0.91~0.97。在各中医证候亚型中, 轻度肾阳虚证的 mAP@50 最高(0.900), 推理速度达到 89.00 FPS。Grad-CAM 分析显示, 模型将激活区域定位于边缘齿痕与舌根部局灶性苔质等中医关键病理特征, 同时抑制非诊断性的口腔背景噪声。**结论** YD-YOLO 模型验证了利用深度学习进行中医阳虚证亚型精细化分类的可行性。将视觉表型量化与模型可解释性相结合, 为中医辨证提供了客观参考, 也为标准化数字化诊断系统的开发和中医临床决策支持提供了助力。

【关键词】 阳虚证; 舌象分析; YD-YOLO; 特征提取; 辨证