RESEARCH ARTICLE

# A Tongue Image Dataset with Pathological Annotations for AI-assisted Diagnosis in Traditional Chinese Medicine

**Longfei Gao**[1] **and Xuebo Jin**[1,*]

[1] School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China

## Abstract

**Tongue diagnosis is a core component of Traditional Chinese Medicine (TCM) with important clinical application value, yet its standardization is severely hampered by the subjectivity of manual interpretation and the lack of unified imaging acquisition protocols. Worse still, the scarcity of large-scale annotated datasets has become a key bottleneck restricting the development of artificial intelligence (AI)-assisted TCM tongue diagnosis technology. To address these critical issues, this study constructs a high-quality standardized dataset dedicated to AI-driven TCM tongue diagnosis research. The dataset contains 6,719 high-resolution tongue images collected under strictly standardized conditions, and all images are annotated with 20 pathological symptom categories in line with TCM theoretical systems. Each image is attached with an average of 2.54 clinical labels, all of which have been double-verified and confirmed by licensed TCM practitioners to ensure clinical authenticity and annotation accuracy. In order to facilitate academic research and industrial applications, we have used three mainstream annotation formats (COCO, TXT, XML) to annotate the data, making the dataset compatible and universal. To verify the practical value and effectiveness of the dataset for AI model training, we conducted a comprehensive benchmark test on it using twelve classic deep learning detection models, including multiple variants of YOLOv5/v7/v8 as well as SSD and MobileNetV2. The experimental results fully demonstrate that the dataset can effectively support the training and performance evaluation of AI models for tongue diagnosis. As a high-quality public data resource, this dataset lays a solid and critical foundation for developing reliable computational analysis tools in the field of TCM, alleviating the long-term data shortage problem that hinders the digital development of TCM tongue diagnosis, and promoting the deep integration of AI technology with TCM research and clinical practice through standardized and high-quality diagnostic image data.**

**Keywords**: tongue diagnosis, traditional Chinese medicine, medical image dataset, deep learning, object detection.

**Submitted:** 14 July 2025
**Accepted:** 09 March 2026
**Published:** 10 March 2026

**Vol.** 2, **No.** 1, 2026.
10.62762/BISH.2026.303296

***Corresponding author:**
✉ Xuebo Jin
jinxuebo@btbu.edu.cn

**Citation**
Gao, L., & Jin, X. (2026). A Tongue Image Dataset with Pathological Annotations for AI-assisted Diagnosis in Traditional Chinese Medicine. *Biomedical Informatics and Smart Healthcare*, 2(1), 5–19.

## 1 Introduction

Traditional Chinese Medicine (TCM), with its millennia-old history, is rooted in a holistic philosophy that emphasizes the harmony between nature and the human body, integrating ecological, psychological, cultural, and scientific perspectives into a unified system of diagnosis and treatment. Unlike modern Western medicine, TCM relies heavily on observational techniques, among which tongue diagnosis is a cornerstone of the "Four Diagnostic Methods" [1, 2]. This process traditionally depends on a practitioner's subjective interpretation of visual features such as color, texture, and coating—a method inherently limited by human perception and variability.

Recent advances in artificial intelligence (AI), particularly deep learning-based image analysis, offer unprecedented opportunities to modernize TCM diagnostics. By automating tongue image assessment, AI can enhance objectivity, consistency, and scalability in TCM practice.

However, the integration of AI into tongue diagnosis is confronted with three critical challenges that impede its progress. First, there is the issue of limited data availability, as tongue diagnosis data is typically collected by individual practitioners and rarely made publicly accessible, in stark contrast to standardized medical imaging datasets such as X-rays or MRIs. This scarcity results in insufficient training data for developing robust AI models. Second, the lack of standardized acquisition protocols introduces significant inconsistencies, with variations in lighting, camera settings, and patient conditions undermining the reliability and robustness of AI algorithms. Finally, non-uniform labeling practices further complicate the matter, as TCM diagnostics rely heavily on subjective interpretations—such as associating a "pale tongue with white coating" with Qi deficiency. These subjective annotations vary widely among practitioners, creating a misalignment between traditional TCM concepts and the structured frameworks required for modern AI tasks like object detection or segmentation.

To date, no large-scale, publicly available tongue image dataset meets these requirements. Existing resources either lack TCM-compliant annotations or sufficient sample diversity for deep learning. Addressing this gap necessitates a rigorously collected dataset with: (1) standardized imaging protocols, (2) clinically validated labels from experienced TCM practitioners, and (3) compatibility with mainstream AI tools. This paper introduces the first such dataset, designed to bridge TCM diagnostics and AI research while preserving the theoretical integrity of traditional medicine.

This study presents a novel, expert-annotated dataset designed to bridge the gap between TCM tongue diagnosis and modern deep learning methodologies. Addressing key challenges—such as diagnostic subjectivity, environmental variability, and data scarcity—we present a meticulously curated collection of 6,719 labeled tongue images, spanning 20 distinct symptom categories that reflect TCM theory. With an average of 2.54 diagnostic labels per image, this dataset captures the granularity and clinical nuance essential for accurate TCM assessment, offering a rare resource for AI-driven research in traditional medicine.

Unlike conventional medical imaging datasets, our work fills a critical void by digitizing a centuries-old diagnostic tradition. Each annotation was rigorously validated by authoritative TCM practitioners, ensuring adherence to classical symptom differentiation while meeting the technical demands of deep learning frameworks. By standardizing tongue image acquisition and labeling, we mitigate longstanding issues of inconsistency, enabling robust model training for tasks such as feature segmentation, symptom classification, and automated diagnosis.

Beyond facilitating AI applications in TCM, this dataset invites broader exploration of cross-cultural medical AI. Traditional diagnostics—often marginalized in digital health initiatives—present unique challenges for computer vision, including subtle feature variations (e.g., coating thickness, color gradations) and context-dependent interpretations. Here, we demonstrate how AI can adapt to these complexities, extending the versatility of image-based diagnostics beyond Western biomedicine. The implications are twofold:

1. **Advancing TCM Digitization:** By providing a high-quality benchmark dataset, we accelerate the integration of AI into TCM practice, enhancing objectivity and scalability while preserving its holistic principles.

2. **Expanding AI's Medical Scope:** This work will extend the dominance of Western-centric medical AI, showcasing how deep learning can embrace diverse diagnostic traditions. Future applications may include personalized TCM

diagnostics, telemedicine platforms, and hybrid AI-human decision systems.

In summary, our dataset not only pioneers the computational modernization of TCM but also underscores AI's potential to unify traditional and contemporary medical paradigms. By leveraging machine learning to decode ancient diagnostic wisdom, we open new pathways for interdisciplinary innovation in global healthcare.

## 2 Related Image Datasets Summary

The proliferation of image detection tasks across diverse application domains has necessitated the development of comprehensive and representative image datasets. These datasets serve as benchmarks for evaluating and enhancing the performance of image detection algorithms.

The UNISA2020 dataset leads source camera identification with real images from identical digital cameras, essential for evaluating Source Camera Identification algorithms [3]. In medical imaging, particularly laparoscopic hysterectomy, AutoLaparo addresses the need for high-quality, multi-task labeled data in computer-assisted surgery [4]. The BAID dataset of 3000 backlit images aids in learning resilient enhancers for backlit photos [5]. The Whole Abdominal Organ Dataset (WORD) meets the demand for medical segmentation data, with 150 CT volumes annotating 16 abdominal organs, aiding research and clinical work in organ segmentation [6]. For floorplan analysis, the CubiCasa5K dataset enhances capabilities with 5000 samples annotated into 80+ categories [7]. In ultrasound video analysis for breast lesion detection, a dataset of 188 annotated videos supports training advanced models like CVA-Net [8]. The SegPC-2021 challenge released 775 images of Multiple Myeloma plasma cells, advancing digital pathology tools and cell segmentation [9].

These datasets represent significant progress across various facets of image detection, each addressing unique challenges within its respective field. Developing these datasets enhances both research and practical application in image detection by advancing the capabilities of machine learning models, by providing comprehensive and diverse training datasets. Although these medical imaging datasets have contributed to the discipline, they often exhibit a concentrated utility, primarily serving specific medical specializations. While this specificity is advantageous for specialized research, it may limit the applicability of the models developed for broader medical or traditional tactile fields. Furthermore, the inherent challenges associated with labeling, including the requirement for specialized medical expertise, can hinder the expansion of the dataset and the implementation of more generalized image detection models.

Several tongue image datasets have been proposed in prior research, though with notable limitations. The dataset hosted on ScienceDB [10] contains 1,194 images but requires restricted access approval from the depositor. The BioHit/TongeImageDataset [11] repository provides compressed files that we found to be corrupted upon download, rendering them unusable. Chang et al. [12] recently released a dataset focusing on four tongue features (fissures, tooth marks, and coating thickness), yet it remains non-public due to institutional privacy policies.

To the best of our knowledge, there is currently no formally published and publicly available AI-ready dataset specifically designed for research in TCM tongue diagnosis. The lack of standardized, large-scale tongue image datasets with comprehensive annotations remains a significant gap in the field of AI-assisted TCM diagnostics. This limitation hinders the development and benchmarking of robust machine learning models for automated tongue analysis in clinical applications.

## 3 Methodology

To maximize the research utility of this dataset in AI-augmented TCM tongue diagnosis, our study adopts a dual-pronged methodological approach:

- Standardized Image Acquisition Protocol: We designed a dedicated hardware system for tongue image capture, ensuring consistent lighting conditions, angle alignment, and color calibration to minimize environmental variability. Further, hospital practitioners underwent rigorous training protocols to standardize data collection procedures, guaranteeing uniformity across all samples. This mitigates inter-operator discrepancies and enhances dataset reliability for downstream AI applications.
- Expert-Annotated, AI-Ready Labelling Framework: To ensure both clinical authenticity and technical applicability, the diagnostic labels were meticulously curated under the guidance of renowned TCM physicians, achieving a dual alignment that captures essential TCM-specific

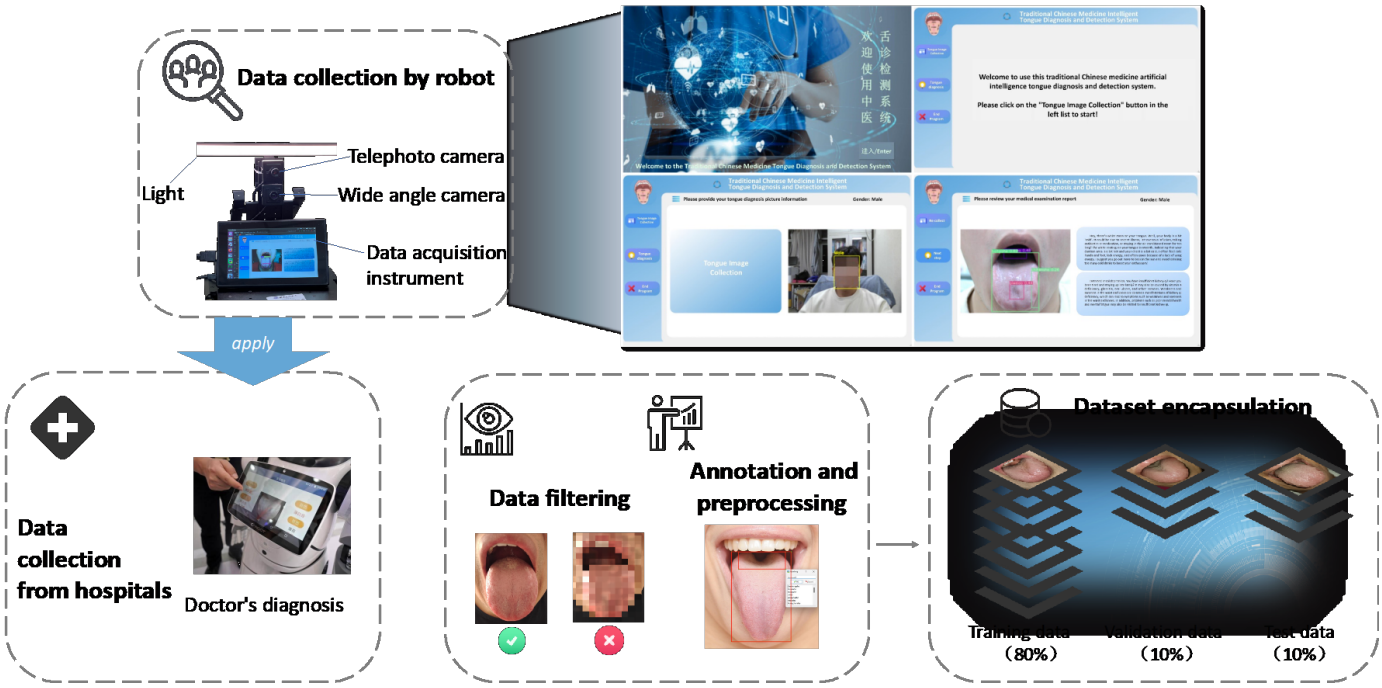


**Figure 1.** Standardized workflow for tongue image acquisition and dataset construction. The protocol initiates with participant registration and informed consent procedures, followed by automated positioning guidance using our specialized imaging device. The integrated system concurrently performs three key functions: (1) demographic profiling via facial recognition technology, (2) standardized image acquisition under controlled lighting conditions (D65 standard illuminant), and (3) systematic archiving of acquired tongue images. The complete dataset was partitioned into training (80%), validation (10%), and test (10%) subsets using stratified random sampling to maintain diagnostic category distributions.

nuances while maintaining deep learning compatibility. The annotations preserve classical TCM diagnostic markers—including tongue coating texture, color gradations, and fissure patterns—rooted in traditional theory, while simultaneously being structured as multi-label classifications and segmentation masks optimized for training modern neural networks such as YOLOs and vision transformers. This approach effectively bridges TCM's holistic diagnostic framework with the computational requirements of AI, enabling the development of models that are both clinically meaningful for traditional medicine and technically robust for machine learning applications.

By integrating domain expertise with technical standardization, our dataset not only supports advanced AI research but also upholds the integrity of TCM's diagnostic traditions. All research data were collected with participants' informed consent, including explicit permission for public sharing. The dataset has been anonymized and shared in compliance with ethical approval protocols to safeguard participant privacy.

### 3.1 Implementation of the Tongue Diagnosis Capture System

To ensure rigorous standardization in data acquisition, we developed a purpose-built tongue imaging system (Figure 1) based on the PyQt framework. The system integrates a synchronized dual-camera array with precision-calibrated illumination modules to achieve consistent imaging conditions. The primary wide-angle imaging unit features intelligent facial proximity detection within an operational range of 30-50cm. At the same time, its embedded ResNet-50 convolutional neural network performs real-time demographic analysis through advanced facial recognition. This initial profiling ensures subject eligibility and triggers subsequent imaging protocols.

Upon successful facial verification, the system activates its secondary telephoto imaging component through an event-triggered mechanism. This high-resolution module features a 5x optical zoom with focus stacking capability, enabling sub-100 $\mu$m resolution for detailed reconstruction of the tongue surface topography. To optimize subject positioning, an adaptive voice guidance protocol provides six-degree-of-freedom adjustment instructions, while real-time quality

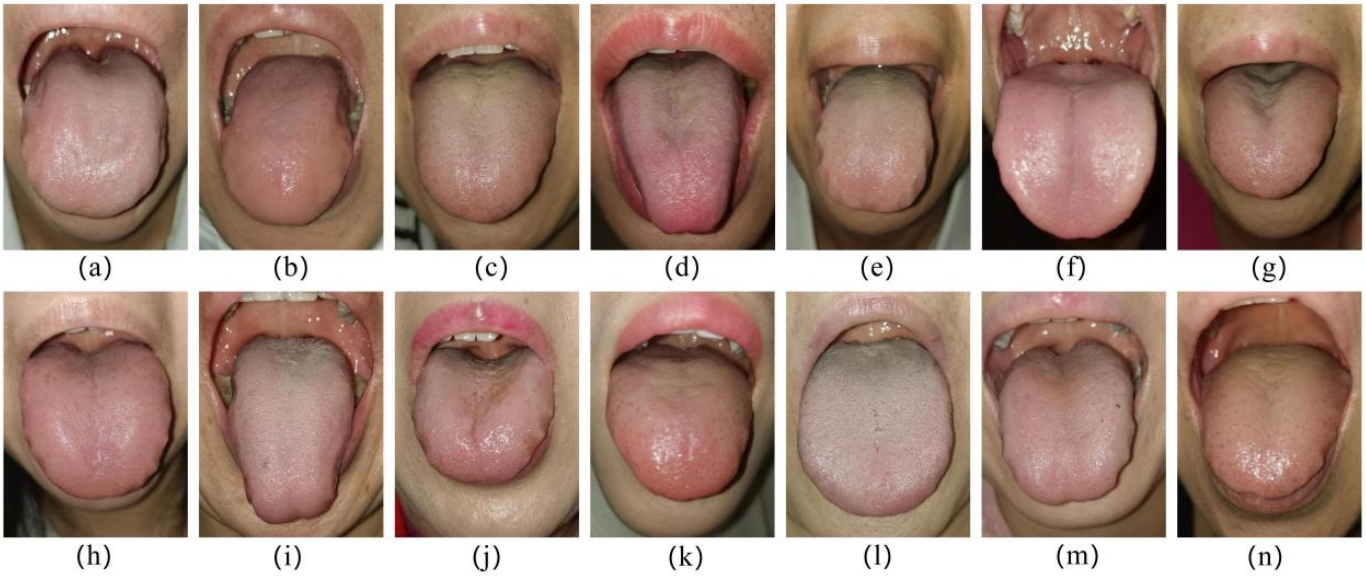


**Figure 2.** Images captured by the tongue diagnosis capture system.

assessment monitors focus peaking, tongue coverage, and motion artifacts. Cross-polarized, spectrally calibrated LED arrays (D65 standard illuminant) automatically adjust intensity between 500-1500 lux to minimize specular reflections and maintain consistent lighting across captures.

For resource-limited research settings, we provide a cost-effective and feasible equipment alternative that significantly reduces costs while ensuring data quality. For imaging, a mirrorless camera that supports RAW or a mid- to high-end smartphone with a professional mode can be used. Key parameters are fixed focal length (50mm equivalent), manual white balance (calibrated using an 18% gray card), and low ISO ($\leq$400). The lighting system can utilize a 5000-5500K LED ring light or a 5600K photography bulb in conjunction with a softbox, using a mobile phone light meter app to ensure uniform lighting. The collection environment can be constructed using a black background cloth, a custom-made adjustable chin rest (3D printed or made of wood), and a simple blackout cloth, maintaining a standard shooting distance of 35cm. For quality control and post-acquisition alignment with the dataset's standardized characteristics (e.g., D65 illuminant and consistent color fidelity), it is recommended to use a mini color chart for color calibration and open-source image processing tools—such as RawTherapee (for non-destructive RAW development and batch white balance correction) or OpenCV (for programmatic batch processing including white balance adjustment, tongue region cropping, and size normalization)—to perform basic preprocessing steps like white balance correction, cropping to the tongue area, and resizing.

The complete acquisition cycle is completed within 3-8 seconds, with all imaging parameters automatically logged in DICOM-compatible metadata. This integrated approach not only ensures clinical-grade image quality but also guarantees research-grade reproducibility, making the dataset robust for downstream AI analysis. By combining automated demographic screening, adaptive positioning guidance, and high-precision optical capture, the system bridges clinical workflow efficiency with the technical demands of machine learning applications. Figure 2 shows examples of images captured by the tongue diagnosis capture system.

### 3.2 Label Selection and Annotation

The images acquired through the standardized tongue diagnosis capture system will subsequently undergo multi-label annotation by TCM practitioners. Our label selection approach is fundamentally rooted in three core principles of TCM: (1) the preventive orientation that focuses on detecting subclinical pathological patterns before they manifest as overt disease; (2) the holistic assessment framework that evaluates tongue features not as isolated markers but as interconnected systems reflecting the body's dynamic equilibrium; and (3) the semantic continuity that deliberately preserves TCM's original diagnostic taxonomy to maintain conceptual integrity with

**Table 1.** Clinically validated tongue feature categories for TCM pattern differentiation.

| No. | Label Name | Category | Label Type | Detailed Explanation |
|---|---|---|---|---|
| 1 | jiankangshe | Healthy Tongue | Global Label | Presents as light red with a moist texture, moderate size, and a thin white coating. Characterized by the absence of stickiness/greasiness or morphological abnormalities. Represents a balanced qi-blood status in healthy individuals. |
| 2 | botaishe | Tongue with Peeling Coating | Local Label | Exhibits partial/complete coating loss ("map tongue"), typically white-greasy. Pathognomonic for spleen-stomach deficiency (qi decline) and yin-fluid depletion. |
| 3 | hongshe | Red Tongue | Global Label | Erythema indicates progression of heat syndrome (excess fire/qi). Hue intensity correlates with pathological heat severity (e.g., crimson suggests extreme heat). |
| 4 | zishe | Purple Tongue | Global Label | Reflects cold/heat syndromes or blood stasis. Modern pathophysiology involves microcirculatory disturbances, hypoxia, or metabolic disorders. |
| 5 | pangdashe | Chubby Tongue | Global Label | Macroglossia with potential protrusion. Signifies cold-damp retention from spleen/kidney yang deficiency (water-dampness internal accumulation). |
| 6 | shoushe | Thin Tongue | Global Label | Atrophic morphology with reduced volume. Indicates blood/yin deficiency syndromes (e.g., spleen deficiency, essence depletion). |
| 7 | hongdianshe | Red Dot Tongue | Local Label | Fungiform papillary hyperemia presenting as punctate lesions. Associated with heat-toxin accumulation or blood stasis. |
| 8 | liewenshe | Cracked Tongue | Local Label | Variably patterned fissures/grooves. Signifies essence-blood depletion or chronic yin deficiency. |
| 9 | chihenshe | Dentate Tongue | Local Label | Lingual scalloping from dental pressure. Pathognomonic for spleen qi/yang deficiency with dampness retention. |
| 10 | baitaishe | White Coating Tongue | Global Label | A white coating suggests cold-damp obstruction or food stagnation. |
| 11 | huangtaishe | Yellow Coating Tongue | Global Label | Transition from white indicates heat progression (e.g., stomach heat, inflammatory processes). |
| 12 | heitaishe | Black Coating Tongue | Global Label | Critical sign of extreme heat/cold. Requires urgent intervention (e.g., heat exhaustion or yang collapse). |
| 13 | huataishe | Smooth Coating Tongue | Local Label | Hyperhydrated surface with drooling tendency. Indicates yang deficiency failing to transform dampness (common in spleen-kidney yang deficiency). |
| 14 | shenquao | Tongue with Sunken Kidney Area | Local Label | Depression at the tongue root. Reflects kidney qi deficiency with associated lumbago/asthenia. May correlate with adrenal fatigue in integrative medicine. |
| 15 | shenqutu | Tongue with Protruding Kidney | Local Label | Bulging lesions suggest kidney heat/toxin accumulation. Differential includes chronic nephropathies or metabolic disorders. |
| 16 | gandanao | Tongue with Concave Liver and Gallbladder Area | Local Label | Lateral tongue depressions. Signifies liver-blood deficiency or qi stagnation (e.g., effects of chronic stress or sleep deprivation). |
| 17 | gandantu | Tongue with Protruding Liver and Gallbladder Area | Local Label | Nodular elevations indicate liver qi stasis or biliary disorders (e.g., gallstones, fatty liver disease). |
| 18 | piweiao | Tongue with Sunken Spleen and Stomach Area | Local Label | Mid-tongue concavity. Pathognomonic for spleen-stomach qi/yin deficiency with possible malabsorption. |
| 19 | xinfeiao | Tongue with Sunken Heart and Lung Area | Local Label | Anterior depression suggests cardiopulmonary qi deficiency or chronic hypoxia (e.g., sleep apnea effects). |
| 20 | xinfeitu | Tongue with Protruding Heart and Lung Area | Local Label | Tip elevation reflects lung yin deficiency or heart fire. Associated with chronic respiratory/circulatory stress. |

centuries of clinical observation and theoretical development. These principles work synergistically to ensure our computational models capture the essence of TCM diagnosis - where early intervention potential, systemic correlation interpretation, and terminological precision together form an inseparable foundation for accurate pattern differentiation and clinical decision-making.

Specifically, we selected 20 TCM tongue diagnosis feature annotations based on classical literature and clinical consensus from clinical practitioners [13, 14]. This system maintains precise alignment with original TCM pathophysiological concepts, enabling accurate pattern differentiation in accordance with TCM's syndrome classification framework. By adhering to these principles, the annotation system ensures that computational models derived from it uphold the theoretical integrity of TCM while translating traditionally qualitative observations into quantifiable data. The label set encompasses all major diagnostic dimensions of tongue examination, including surface characteristics such as coating and moisture, morphological changes like teeth marks and cracks, color variations ranging from pale to crimson, and regional abnormalities corresponding to specific organ systems.

As listed in Table 1, the label systematically categorizes both global tongue characteristics (e.g., color, texture, and morphology) and localized pathological markers (e.g., organ-specific depressions or protrusions). The dataset includes fundamental diagnostic indicators such as healthy tongue (jiankangshe), color variations (e.g., hongshe, zishe), texture abnormalities (e.g., liewenshe, chihenshe), and region-specific features (e.g., shenquao, gandantu). Each label is numerically indexed (0-19) to facilitate machine learning applications in automated tongue image analysis. This structured dataset serves as a valuable resource for advancing research in digital TCM diagnostics and AI-assisted pattern recognition.

The annotation approach comprises three key elements: (1) label category classification, (2) center coordinates of the bounding box, and (3) dimensional parameters (width and height). A comprehensive quality control system spans the entire workflow from annotation to release. Initial annotations by trained technicians undergo review by TCM practitioners with 5+ years of experience, with critical cases requiring consensus from expert panels. Generated using the industry-standard LabelImg annotation

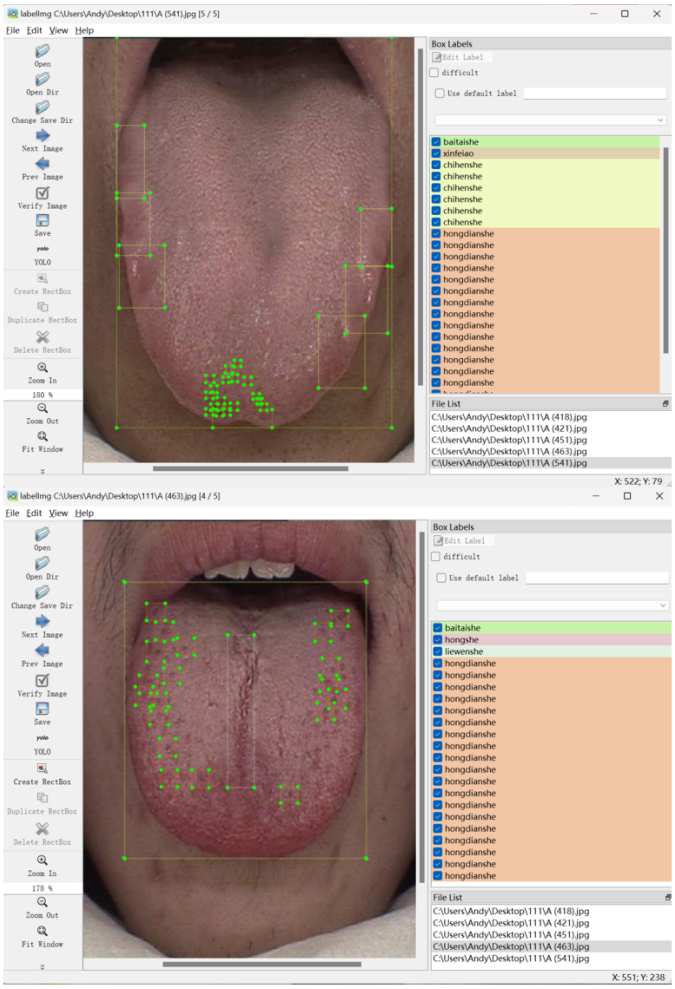


**Figure 3.** Workflow of multi-level annotation for computerized tongue image analysis.

tool, the dataset provides dual-format annotation files: PASCAL VOC XML files that preserve hierarchical metadata of diagnostic features, and YOLO TXT files with normalized coordinates for real-time detection requirements.

As illustrated in Figure 3, the labeling system employs a dual-category approach: Global Labels, which capture the tongue's holistic characteristics, and Local Labels, which focus on specific regional features. This hierarchical annotation framework enables comprehensive analysis by simultaneously representing macroscopic conditions and microscopic pathological details.

Figure 4 visually demonstrates the annotation scheme, distinguishing between global and local labels through a standardized color-coding system. Global labels, marked by a red box for white-coated tongue (baitaishe) and an orange-yellow box for red tongue (hongshe), capture the tongue's overall appearance, with both labels occupying the same

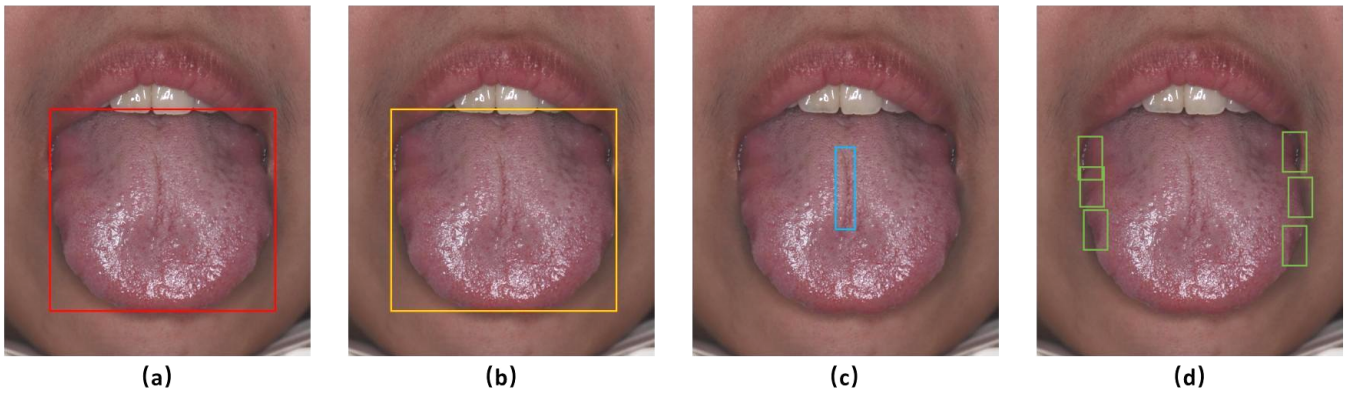


**Figure 4.** Color-coded annotation scheme for tongue features: (a) red box - white-coated tongue (global labels), (b) orange-yellow box - red tongue (global labels, same position as white-coated label); (c) blue box - cracked tongue (local labels), (d) green box - dentate tongue (local labels).

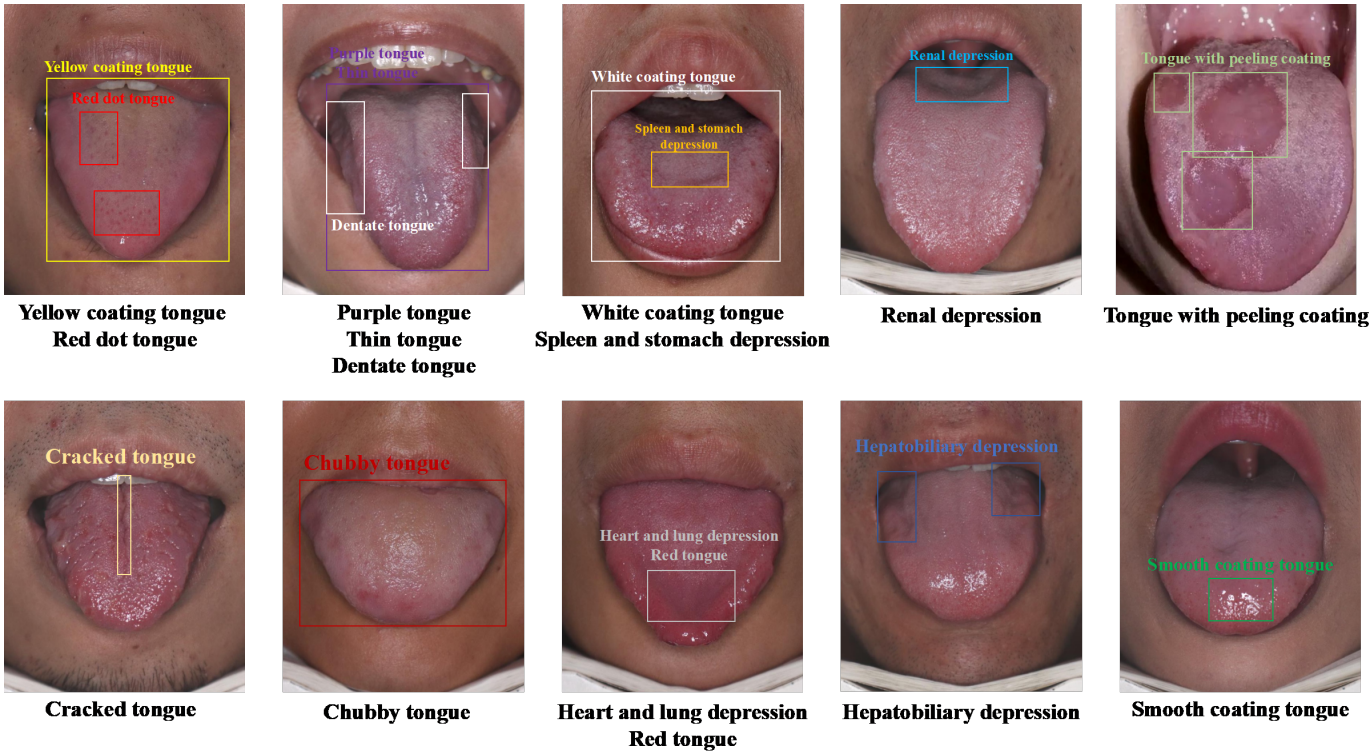


**Figure 5.** Typical examples of an image dataset for tongue diagnosis detection.

spatial region to reflect coexisting conditions. Local labels, highlighted in blue for cracked tongue (liewenshe) and green for dentate tongue (chihenshe), pinpoint specific pathological features, enabling fine-grained analysis. This dual-level annotation approach ensures comprehensive representation of both macroscopic and microscopic diagnostic indicators while maintaining visual clarity for automated detection.

### 3.3 Data Records

We have stored the data in the Dryad Digital Repository [15], including high-resolution source images, multi-format annotation files, detailed technical documentation, and preprocessing scripts. To facilitate replication studies, sample code for PyTorch and TensorFlow implementations is provided, covering complete workflows from data loading and augmentation to evaluation.

### 3.4 Data Overview

Figure 5 illustrates the labeling situation on the tongue image. We find that, unlike other image target detection, the labels for the target not only encompass the overall condition of the entire surface of the tongue but also emphasize the specific local areas, reflecting

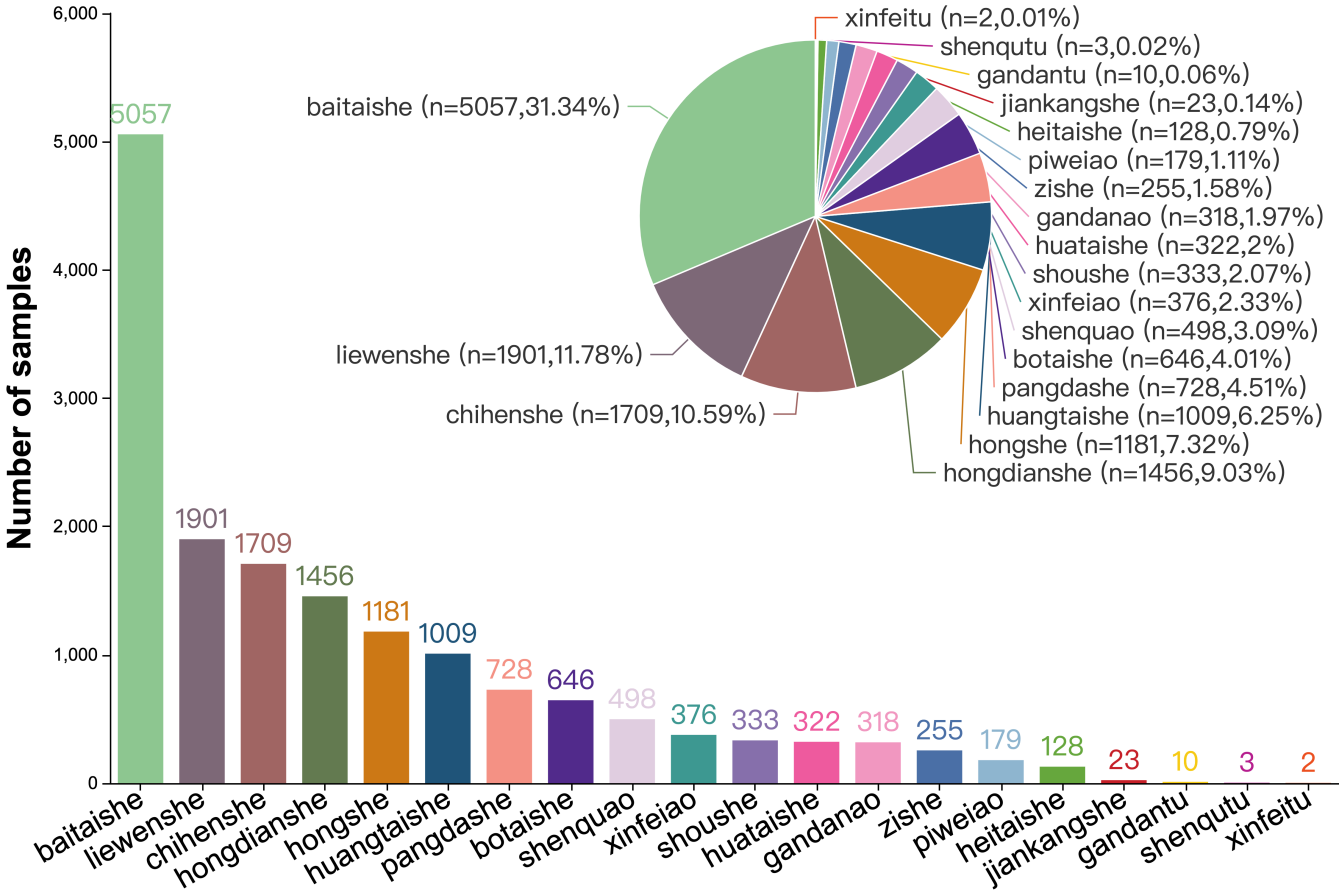


**Figure 6.** Distribution of labels in the tongue diagnosis dataset.

the observational characteristics of the tongue in TCM. Figure 6 illustrates the distribution of image quantities for each label, showing an average of 2.54 labels per tongue image.

Comprising 6,719 clinically validated tongue images, the dataset adheres to machine learning best practices in its division, with 5,594 images (82.3%) allocated for training, 572 images (8.4%) for validation, and 553 images (8.1%) for final evaluation. A specialized subset containing 10% "challenging cases" represents borderline tongue manifestations in TCM differential diagnosis, providing valuable test scenarios for model robustness.

## 4 Experiments

### 4.1 Experimental Setup

To comprehensively evaluate the dataset, all the code used for training and testing models in this dataset was run in a Linux environment, Ubuntu 20.04, CUDA version 11.4, and the experimental framework used was PyTorch. The processor model is Intel Core i7-6800K 3.40GHz, and the graphics card model is three Nvidia GeForce GTX 1080 Ti, with 11GB of graphics memory and 32GB of memory.

### 4.2 Evaluation Metrics

The selected algorithm is also a classic detection algorithm in both one-stage and two-stage detection networks. The precision, recall, and mAP are used as evaluation indicators [16], and their calculation formulas are shown in Equations (1)–(5). To address potential ambiguity in the multi-class evaluation protocol, we clarify that all reported metrics (precision, recall, and mAP) follow the standard COCO object detection evaluation protocol, which is widely adopted in modern benchmarks and implemented in the YOLO frameworks used for our experiments. Although Equations (1)–(3) are presented in binary form for clarity, the multi-class (20-category) extension employs a one-vs-rest strategy with bounding-box Intersection over Union (IoU [17])-based matching, in

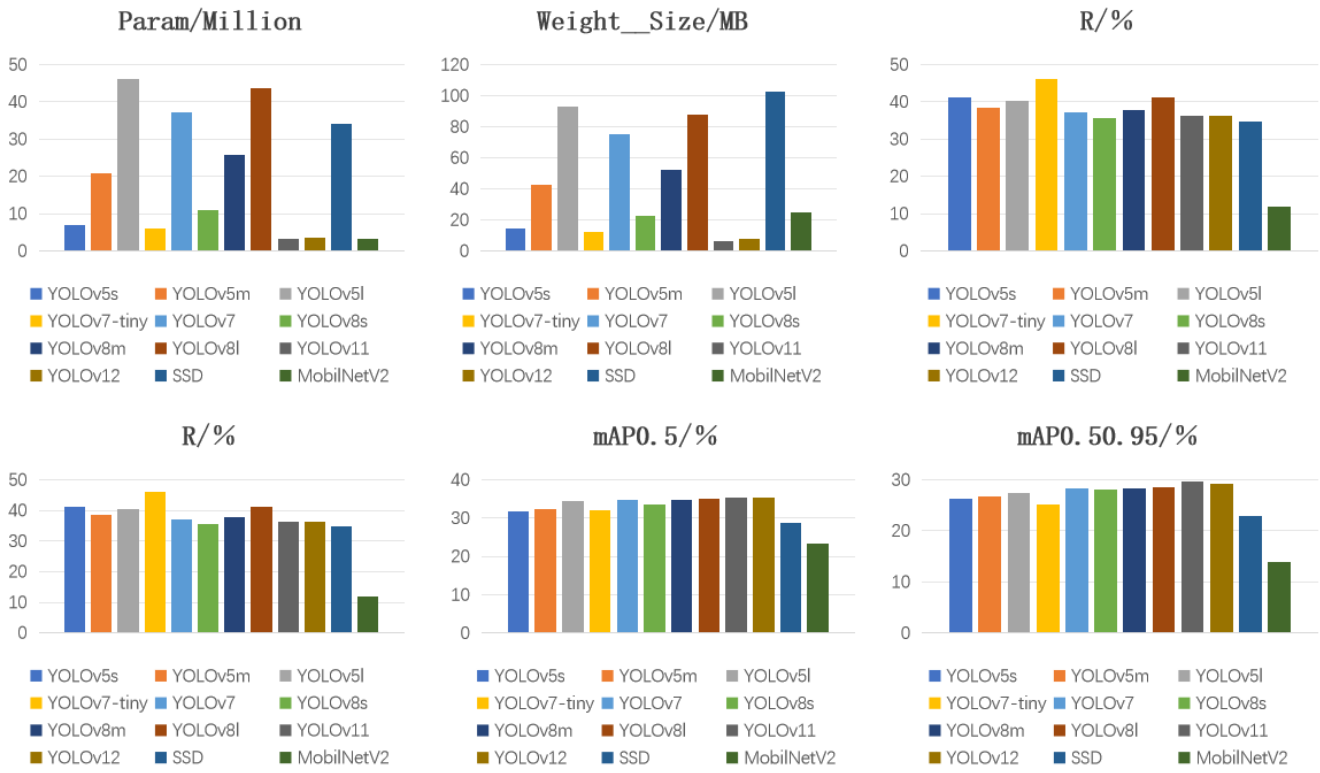


**Figure 7.** Bar chart visualization: performance of tongue diagnosis detectors.

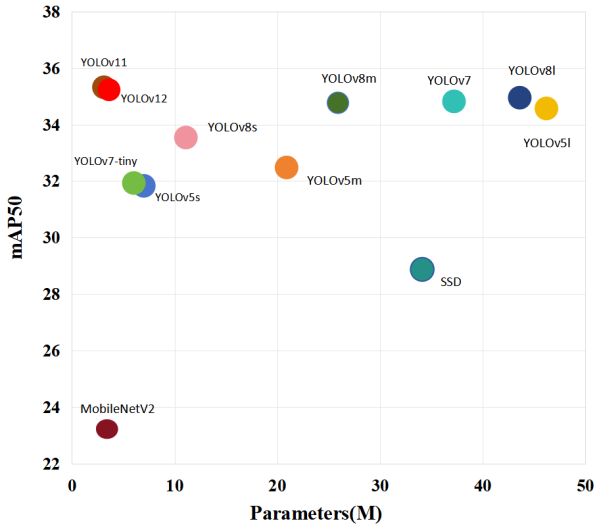


**Figure 8.** Joint visualization of mAP50 accuracy and parameter scale for tongue image analysis.

which detections of each category are independently treated as positives (all others as negatives). A prediction is classified as a True Positive (TP [16]) if its IoU with an unmatched ground-truth box of the same category exceeds the threshold (default 0.5 for mAP@0.5) and it has the highest confidence among overlapping predictions; unmatched predictions of the target category are False Positives (FP [16]), and unmatched ground-truth boxes are False Negatives (FN [16]). Precision and recall are then computed per category using Equations (1)–(2), with Average Precision (AP) obtained via 101-point interpolation of the precision-recall curve Equation (3); the reported mAP@0.5 is the mean AP across all 20 categories at IoU=0.5, while mAP@0.5:0.95 is the mean AP averaged over 10 IoU thresholds from 0.5 to 0.95 (step 0.05), as defined in the COCO protocol Equations (4)–(5). This protocol applies identically to both global labels (e.g., overall tongue color/coating) and local labels (e.g., fissures, tooth marks, organ-specific regions), with IoU computed using the standard axis-aligned rectangular bounding box formula $\text{IoU} = |A \cap B|/|A \cup B|$, where $A$ and $B$ are the predicted and ground-truth rectangles, respectively. All benchmarks were performed using official Ultralytics evaluation scripts, ensuring full compliance with the COCO metrics:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$AP = \int_0^1 \text{Precision}\, d(\text{Recall}) \tag{3}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{4}$$

**Table 2.** Performance comparison of deep learning-based detectors on the tongue diagnosis dataset.

| Algorithm | Param /Million | Weight_Size /MB | P/% | R/% | mAP0.5 % | mAP0.5 0.95/% |
|---|---|---|---|---|---|---|
| YOLOv5s [18, 19] | 7.0 | 14.5 | 39.61 | 41.26 | 31.83 | 26.31 |
| YOLOv5m [20] | 20.9 | 42.3 | 42.47 | 38.54 | 32.48 | 26.65 |
| YOLOv5l [21] | 46.2 | 92.9 | 43.76 | 40.29 | 34.57 | 27.33 |
| YOLOv7-tiny [22] | 6.03 | 12.3 | 40.77 | 46.18 | 31.93 | 25.09 |
| YOLOv7 [23, 24] | 37.2 | 74.9 | 47.23 | 37.12 | 34.82 | 28.38 |
| YOLOv8s [25, 26] | 11.1 | 22.5 | 40.75 | 35.52 | 33.54 | 27.98 |
| YOLOv8m [27] | 25.9 | 52.1 | 40.36 | 37.84 | 34.77 | 28.20 |
| YOLOv8l [28] | 43.6 | 87.7 | 38.19 | 41.07 | 34.95 | 28.59 |
| YOLOv11 [29] | 3.12 | 6.21 | 48.12 | 36.25 | 35.32 | 29.63 |
| YOLOv12 [30] | 3.62 | 7.63 | 46.52 | 36.39 | 35.23 | 29.12 |
| SSD [31, 32] | 34 | 102.7 | 38.72 | 34.75 | 28.86 | 22.90 |
| MobileNetV2 [33] | 3.4 | 24.5 | 34.07 | 11.75 | 23.20 | 13.90 |

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{5}$$

where mAP@0.5 represents the mAP value at an IoU threshold of 0.5, and mAP@0.5:0.95 represents the average of mAP values computed at IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. $TP$ denotes the number of true positive samples, $FP$ denotes the number of false positive samples, $FN$ denotes the number of false negative samples, and $C$ denotes the number of categories.

The study evaluates several deep learning-based object detection models, including various versions of YOLOv5, YOLOv7, YOLOv8, YOLOv11, YOLOv12, SSD, and MobileNetV2, on a tongue diagnosis dataset to determine the most suitable architecture for medical image analysis. Table 2 summarizes the performance comparison among deep learning-based detectors on the tongue diagnosis dataset, while Figure 7 provides a bar chart visualization of these results. Contrary to expectations, increasing model depth within the same family (e.g., from YOLOv5s to YOLOv5l) does not consistently improve accuracy or recall, with performance instead showing fluctuations or even degradation. This phenomenon likely stems from the tongue dataset being significantly smaller than standard benchmarks, such as COCO, causing deeper models with more parameters to overfit. Their redundant parameters also hinder generalization.

Among lightweight models, YOLOv8s emerges as particularly effective, achieving superior mAP0.5 (33.54%) and mAP0.5-0.95 (27.98%) compared to similar-sized models, such as YOLOv7-tiny and YOLOv5s, while maintaining a low parameter count of just 11.1 million. For applications requiring higher accuracy, YOLOv7 and the medium- and large-scale YOLOv8 variants deliver the best performance, with YOLOv8l achieving 34.95% mAP@0.5. Notably, YOLOv7 achieves competitive accuracy (34.82% mAP@0.5) with fewer parameters (37.2M) than YOLOv8l (43.6M), thanks to its efficient E-ELAN module, which optimizes gradient flow for improved feature learning. While YOLOv8 shows slight mAP improvements through its innovative C2F structure and Anchor-Free detection strategy, the practical advantages of YOLOv7's parameter efficiency and faster inference make it particularly suitable for tongue diagnosis applications. The more advanced YOLOv11 and YOLOv12 have more powerful performance, not only with fewer parameters and faster inference speed, but also with higher accuracy than previous YOLO versions, and mAP values far higher than other models, making them more suitable for the diagnosis and recognition of tongue diagnosis data. The poorer performance of SSD (28.86% mAP0.5) and MobileNetV2 (23.20% mAP0.5) highlights the limitations of older architectures and extremely lightweight designs for this complex medical imaging task. These findings suggest that for tongue diagnosis and similar specialized medical imaging tasks with limited datasets, mid-sized models like YOLOv7 or YOLOv8m offer the optimal balance between accuracy and computational efficiency, avoiding both the underperformance of overly simplistic models and the diminishing returns of excessively large architectures.

As we know, mAP50-95 is an indicator obtained by averaging 10 mAP values with an interval of 5%, ranging from a mAP threshold of 50% to a

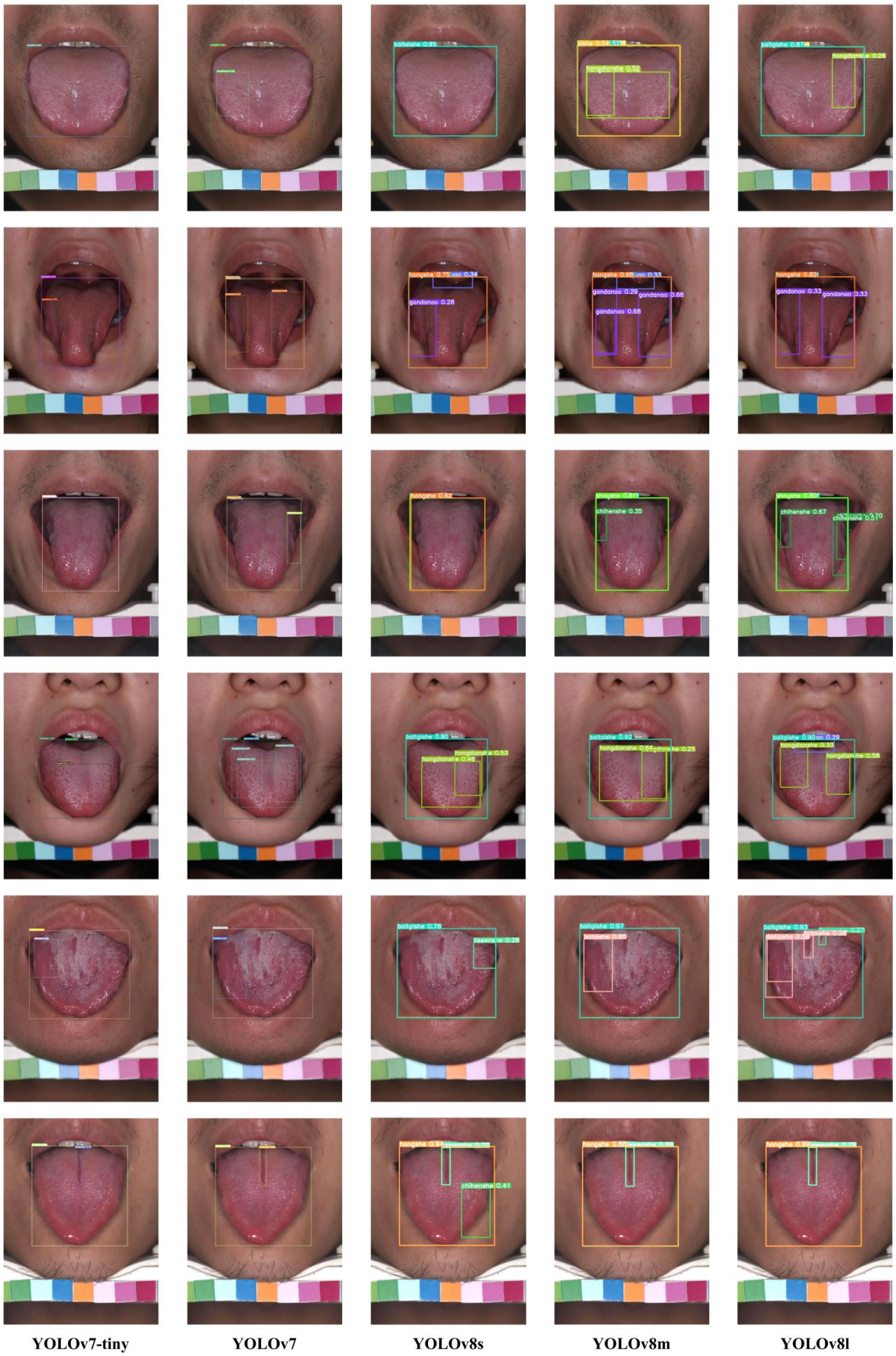


**Figure 9.** Benchmarking object detection performance on tongue diagnosis dataset.

mAP threshold of 95%. Compared to mAP50, it is more stringent and requires higher network detection capabilities. As shown in Figure 8, we compared the performance of non-lightweight networks on the tongue diagnosis dataset, with an evaluation metric of mAP50-95. It can be seen that the performance of YOLOv11 and YOLOv12 is very impressive, followed by YOLOv8. Considering the application methods of intelligent tongue diagnosis and the particularity of tongue diagnosis datasets, a balance between computing power and accuracy should be considered, because in most scenarios, the computing power of intelligent tongue diagnosis robots is not sufficient to support fast inference and testing, and relatively lightweight networks are more feasible. Further, it is necessary to ensure the accuracy of the detector to support the scientific rigor of tongue diagnosis results.

In Figure 9, we compared the results of the detection network trained on the tongue diagnosis dataset and tested on test cases. It is not difficult to see that YOLOv8l, as the model with the largest number of parameters, has more accurate and comprehensive detection and classification results. However, YOLOv7-tiny and YOLOv8s, with smaller parameter quantities, can reduce detection time and GPU memory consumption. Nevertheless, they often result in false positives and missed detections, which are typically caused by incomplete feature extraction and insufficient network fitting. Meanwhile, in the face of similar color features, such as purple and red tongues, the classification performance of the above detectors still needs to be improved. Improving the detection performance of detectors is also crucial for irregular features, including cracked tongues and dentate tongues.

This dataset supports both object detection and classification tasks. In addition to the aforementioned YOLO series [34], compatible networks also include SD [35], Faster R-CNN [36], and Mask R-CNN [37] models. The data comes pre-processed and can be directly loaded using Python (OpenCV/PIL) or deep learning frameworks (PyTorch/TensorFlow). Refer to the "Label Selection and Annotation" sections for detailed labeling criteria. The dataset is ready for immediate use - simply download and start training without additional preprocessing.

Our work is currently open source and will continue to be revised and updated in the future. Currently, it has been updated to version 3.0, including the coco annotation format, the xml annotation format, and the YOLO annotation format. txt. These annotation data are all provided for free, and we also provide the code for YOLO (https://github.com/m28805746-max/Intelligent-tongue-diagnosis-detection-dataset). Our dataset is provided through cloud storage, and the demo includes some simple instances, along with configuration files for the relevant models. We have also added a brief explanation of the dataset to the README.md.

## 5 Conclusion

This study focuses on the core data bottleneck of the development of traditional Chinese medicine tongue diagnosis AI. The core contribution is the construction of a publicly available standardized AI specific dataset for traditional Chinese medicine tongue diagnosis. This dataset fills the gap in the field of lacking large-scale high-quality annotated data through strict standardized collection, clinically validated pathological labeling, and multi format compatible output design; The simultaneous development of tongue image collection and annotation standards has effectively adapted traditional Chinese medicine theory to AI frameworks, and the benchmark testing of 12 deep learning models has provided direct model selection references for subsequent research. In addition, we also provide download links for the dataset, usage instructions, and ways to download the validation code. The open source and release of this dataset not only lays the data foundation for the development of reliable computational analysis tools in the field of traditional Chinese medicine tongue diagnosis, alleviates the core problems that restrict its digital development, but also promotes the deep integration of AI technology with traditional Chinese medicine research and clinical practice, providing a replicable research paradigm for the intelligent transformation of traditional medicine. In addition, this study also found that the current model still has shortcomings in detecting similar tongue color and irregular tongue image features. Subsequent research will further expand rare tongue image samples, optimize the model's feature extraction ability, improve the accuracy and generalization of AI tongue diagnosis detection, and promote its wider clinical application.

## Data Availability Statement

The tongue diagnosis dataset and all associated code supporting the findings of this study are openly available. The dataset can be accessed via our GitHub repository at https://github.com/m28805746-max/Intellige

nt-tongue-diagnosis-detection-dataset. The YOLO-based detection code is also provided in the same repository. Our work is open source and will continue to be revised and updated in the future. All annotation data are provided free of charge for academic use.

## Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62433002, Grant 62476014, Grant 62473008, Grant 62173007, and Grant 62203020; in part by the Beijing Nova Program under Grant 20240484710; in part by the Project of Beijing Municipal University Teacher Team Construction Sup-port Plan under Grant BPHR20220104; in part by the Beijing Scholars Program under Grant No.099.

## Conflicts of Interest

The authors declare no conflicts of interest.

## AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

## Ethical Approval and Consent to Participate

The study involves non-invasive tongue image acquisition and does not collect identifiable personal information. All participants provided informed consent for image collection and public release of anonymized data. According to institutional guidelines, formal ethical approval was not required.

## References

[1] Wang, W. Y., Zhou, H., Wang, Y. F., Sang, B. S., & Liu, L. (2021). Current policies and measures on the development of traditional Chinese medicine in China. *Pharmacological research, 163*, 105187. [CrossRef]

[2] Kim, M., Cobbin, D., & Zaslawski, C. (2008). Traditional Chinese medicine tongue inspection: an examination of the inter-and intrapractitioner reliability for specific tongue characteristics. *The Journal of Alternative and Complementary Medicine: Paradigm, Practice, and Policy Advancing Integrative Health, 14*(5), 527-536. [CrossRef]

[3] Bruno, A., Capasso, P., Cattaneo, G., Petrillo, U. F., & Improta, R. (2023). A novel image dataset for source camera identification and image based recognition systems. *Multimedia Tools and Applications, 82*(8), 11221-11237. [CrossRef]

[4] Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T. H., Dou, Q., & Liu, Y. (2022, September). Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 486-496). Cham: Springer Nature Switzerland. [CrossRef]

[5] Lv, X., Zhang, S., Liu, Q., Xie, H., Zhong, B., & Zhou, H. (2022). BacklitNet: A dataset and network for backlit image enhancement. *Computer Vision and Image Understanding, 218*, 103403. [CrossRef]

[6] Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., ... & Zhang, S. (2022). WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. *Medical Image Analysis, 82*, 102642. [CrossRef]

[7] Kalervo, A., Ylioinas, J., Häikiö, M., Karhu, A., & Kannala, J. (2019, May). Cubicasa5k: A dataset and an improved multi-task model for floorplan image analysis. In *Scandinavian Conference on Image Analysis* (pp. 28-40). Cham: Springer International Publishing. [CrossRef]

[8] Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., & Wang, L. (2022, September). A new dataset and a baseline model for breast lesion detection in ultrasound videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 614-623). Cham: Springer Nature Switzerland. [CrossRef]

[9] Gupta, A., Gehlot, S., Goswami, S., Motwani, S., Gupta, R., Faura, Á. G., ... & Ye, J. (2023). Segpc-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images. *Medical Image Analysis, 83*, 102677. [CrossRef]

[10] ScienceDB. (2013). *Tongue Image Dataset* [Data set]. Retrieved from https://www.scidb.cn/en/detail?dataSetId=8417299de5ef4f3db5ec62e01a969d54

[11] BioHit. (n.d.). *TongeImageDataset* [Source code]. GitHub. Retrieved from https://github.com/BioHit/TongeImageDataset

[12] Chang, W. H., Chen, C. C., Wu, H. K., Hsu, P. C., Lo, L. C., Chu, H. T., & Chang, H. H. (2024). Tongue feature dataset construction and real-time detection. *PLoS ONE, 19*(3), e0296070. [CrossRef]

[13] Fu, J., & Yang, M. (2019). *Yellow Emperor's Classic of Medicine, The-Essential Questions: Translation Of Huangdi Neijing Suwen*. World Scientific.

[14] Maciocia, G. (1987). *Tongue Diagnosis in Chinese Medicine*. Eastland Press.

[15] Jin, X., Gao, L., Tong, A., Chen, Z., Kong, J., Sun, N., ... & Su, T. (2025). TCM-Tongue: A Standardized Tongue Image Dataset with Pathological Annotations for AI-Assisted TCM Diagnosis. *arXiv preprint arXiv:2507.18288*.

[16] Casas, E., Ramos, L., Bendek, E., & Rivas-Echeverría, F. (2023). Assessing the Effectiveness of YOLO Architectures for Smoke and Wildfire Detection. *IEEE Access, 11*, 96554–96583. [CrossRef]

[17] Su, K., Cao, L., Zhao, B., Li, N., Wu, D., & Han, X. (2024). N-IoU: better IoU-based bounding box regression loss for object detection. *Neural Computing and Applications, 36*(6), 3049-3063. [CrossRef]

[18] Jiang, T., Li, C., Yang, M., & Wang, Z. (2022). An improved YOLOv5s algorithm for object detection with an attention mechanism. *Electronics, 11*(16), 2494. [CrossRef]

[19] Feng, J., Yu, C., Shi, X., Zheng, Z., Yang, L., & Hu, Y. (2023). Research on winter jujube object detection based on optimized Yolov5s. *Agronomy, 13*(3), 810. [CrossRef]

[20] Tsai, F. T., Nguyen, V. T., Duong, T. P., Phan, Q. H., & Lien, C. H. (2023). Tomato fruit detection using modified yolov5m model with convolutional neural networks. *Plants, 12*(17), 3067. [CrossRef]

[21] Liu, H. W., Zheng, Y. L., Zhong, C. J., Liao, K. Y., Sun, B. Y., Zhao, H. X., ... & Xie, B. (2024). Defect detection of printed matter based on improved YOLOv5l. *Laser & Optoelectronics Progress, 61*(10), 1012002.

[22] Wang, S., Yao, L., Xu, L., Hu, D., Zhou, J., & Chen, Y. (2024). An improved YOLOv7-Tiny method for the segmentation of images of vegetable fields. *Agriculture, 14*(6), 856. [CrossRef]

[23] Yu, M., & Jia, Y. (2023, October). Improved YOLOv7 Small Object Detection Algorithm for Seaside Aerial Images. In *International Symposium on Artificial Intelligence and Robotics* (pp. 483-491). Singapore: Springer Nature Singapore. [CrossRef]

[24] Ge, Z., Zhang, Y., Jiang, Y., Ge, H., Wu, X., Jia, Z., ... & Jia, K. (2024). Lightweight YOLOv7 algorithm for multi-object recognition on contrabands in terahertz images. *Applied Sciences, 14*(4), 1398. [CrossRef]

[25] Zhu, D., Huang, Z., Yung, K., & Ip, A. W. (2024). Drug recognition detection based on deep learning and improved YOLOv8. *Journal of Organizational and End User Computing (JOEUC), 36*(1), 1-21. [CrossRef]

[26] Huang, Y., Jiang, X., Zhou, C., Zhuo, X., Xiong, J., & Zhang, M. (2025). Study on mango ripeness detection on production line based on improved YOLOv8s. *Journal of Food Measurement and Characterization, 19*(1), 768-780. [CrossRef]

[27] Wang, Z., Yuan, G., Zhou, H., Ma, Y., & Ma, Y. (2023). Foreign-object detection in high-voltage transmission line based on improved YOLOv8m. *Applied Sciences, 13*(23), 12775. [CrossRef]

[28] Alaqeb, A. M. A., Rashid, M. M., Zaki, H. F. M., & Embong, A. H. (2024, August). An estimation algorithm for improved maritime obstacle detection. In *2024 9th International Conference on Mechatronics Engineering (ICOM)* (pp. 459-465). IEEE. [CrossRef]

[29] He, L. H., Zhou, Y. Z., Liu, L., Cao, W., & Ma, J. H. (2025). Research on object detection and recognition in remote sensing images based on YOLOv11. *Scientific Reports, 15*(1), 14032. [CrossRef]

[30] Ge, T., Ning, B., & Xie, Y. (2025). YOLO-AFR: an improved YOLOv12-based model for accurate and real-time dangerous driving behavior detection. *Applied Sciences, 15*(11), 6090. [CrossRef]

[31] Chen, Z., Wu, K., Li, Y., Wang, M., & Li, W. (2019). SSD-MSN: An Improved Multi-Scale Object Detection Network Based on SSD. *IEEE Access, 7*, 80622–80632. [CrossRef]

[32] Jiang, Y., Peng, T., & Tan, N. (2019). Cp-ssd: Context information scene perception object detection based on ssd. *Applied Sciences, 9*(14), 2785. [CrossRef]

[33] Gulzar, Y. (2023). Fruit image classification model based on MobileNetV2 with deep transfer learning technique. *Sustainability, 15*(3), 1906. [CrossRef]

[34] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788). [CrossRef]

[35] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, September). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Cham: Springer International Publishing. [CrossRef]

[36] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence, 39*(6), 1137-1149. [CrossRef]

[37] He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017, October). Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2980-2988). IEEE Computer Society. [CrossRef]

**Longfei Gao** received the B.S. electrical engineering and automation from University of Jinan, Shandong, China, in 2022. Currently pursuing a Master's degree in Control Engineering at Beijing Technology and Business University, main research interests focus on image detection and pattern recognition. (Email: gaolongfei@st.btbu.edu.cn)

**Xuebo Jin** received the B.S. and M.S. degrees in control theory and control engineering from Jilin University, Changchun, China, in 1994 and 1997, and the Ph.D. degree in control theory and control engineering from the University of Zhejiang, Zhejiang, China, in 2004.,She was a Senior Visiting Scholar with the University of Illinois at Chicago, Chicago, IL, USA, in2007. From 2009 to 2012, she was an Assistant Professor with Zhejiang Sci-tech University. Since 2012, she has been a Professor with Beijing Technology and Business University, Beijing, China. Her research includes a variety of areas in information fusion, bigdata analysis, condition estimation, and video tracking. (Email: jinxuebo@btbu.edu.cn)