Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

Visualization analysis and clinical interpretability evaluation of artificial intelligence-assisted tongue diagnosis

Lihui Liu[a, b, c], Kaiwen Hu[a, b, c, d*], Yana Zhou[b, c, e, f*]

a. School of Traditional Chinese Medicine, Hubei University of Chinese Medicine, Wuhan, Hubei 430061, China

b. Oncology Department, Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan, Hubei 430074, China

c. Hubei Key Laboratory of Theory and Application Research of Liver and Kidney in Traditional Chinese Medicine, Affiliated Hospital of Hubei University of Chinese Medicine, Wuhan, Hubei 430074, China

d. Oncology Department, Dongfang Hospital, Beijing University of Chinese Medicine, Beijing 100078, China

e. Hubei Province Academy of Traditional Chinese Medicine, Wuhan, Hubei 430061, China

f. Hubei Shizhen Laboratory, Wuhan, Hubei 430060, China

A R T I C L E I N F O

Article history
Received 28 February 2026
Accepted 06 April 2026
Available online 25 June 2026

Keywords
Tongue diagnosis
Artificial intelligence
Traditional Chinese medicine
Diagnostic accuracy
Bibliometric analysis
Meta-analysis

A B S T R A C T

Objective To map the research landscape of artificial intelligence (AI)-assisted tongue diagnosis through bibliometric analysis and to quantify its diagnostic accuracy and clinical interpretability through a diagnostic test accuracy (DTA) meta-analysis.

Methods For the bibliometric analysis, the Web of Science Core Collection (WoSCC) was queried for English-language articles and reviews on AI-assisted tongue diagnosis published between January 1, 2014 and December 31, 2025, and analysed using Bibliometrix, VOSviewer, and CiteSpace, with major output dimensions including annual publication output and disciplinary distribution, journal and citation characteristics, country/region and institutional collaboration, author networks, keyword co-occurrence, and keyword burst detection. For the DTA meta-analysis, four databases [Scopus, PubMed, Web of Science, and China National Knowledge Infrastructure (CNKI)] were searched in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA) guidelines. A bivariate random-effects model hierarchical summary receiver operating characteristic (HSROC) was used to pool sensitivity and specificity, with subgroup analyses by disease category, AI model architecture, and sample-size strata. Methodological quality was assessed with the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) tool, and publication bias was evaluated by Deeks' funnel plot asymmetry test.

Results A total of 198 publications met the bibliometric eligibility criteria. Annual output increased 24.5-fold (from 2 in 2014 to 49 in 2025), with the period 2022 – 2025 alone accounting for 65.2% of all publications. China contributed approximately 83.5% of all institutional affiliations, with Shanghai University of Traditional Chinese Medicine and Jiatuo Xu being the most productive institution and author, respectively. Keyword analysis identified four thematic clusters (AI and deep-learning architectures, image processing and segmentation, traditional Chinese medicine (TCM)-specific applications, and disease-specific applications) and a temporal evolution from traditional machine learning to deep learning and transformer-based, explainable, and multimodal AI architectures. Sixteen DTA meta-analysis studies (14 755 participants) covering metabolic and hepatic disorders, oncological and oral lesions,

*Corresponding author: Yana Zhou, E-mail: 1565285090@qq.com. Kaiwen Hu, E-mail: kaiwenh@163.com.

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

DOI: 10.1016/j.dcmed.2026.05.004

Citation: LIU LH, HU KW, ZHOU YN. Visualization analysis and clinical interpretability evaluation of artificial intelligence-assisted tongue diagnosis. Digital Chinese Medicine, 2026, 9(2): 223-240.

cardiovascular risk, diabetes, and other clinical applications were included in the DTA meta-analysis. The pooled sensitivity was 90.3% [95% confidence interval (CI): 86.7% – 93.1%] and the pooled specificity was 93.0% (95% CI: 90.6% – 94.7%); the area under the summary receiver operating characteristic (SROC) curve (AUC) was 0.961. Heterogeneity was substantial (I^2 = 95.8% for sensitivity; I^2 = 92.1% for specificity). Subgroup performance was broadly consistent across disease categories, AI architectures, and sample-size strata, and Deeks' test indicated no significant publication bias (P = 0.258).
Conclusion AI-assisted tongue diagnosis has progressed rapidly and shows pooled diagnostic performance comparable to established screening modalities, supporting its potential as a complementary and easily accessible decision-support tool.

1 Introduction

Tongue diagnosis, referred to as "Shezhen (舌诊)" in traditional Chinese medicine (TCM), is one of the four fundamental diagnostic methods alongside inspection, auscultation and olfaction, inquiry, and palpation [1]. As a key component of inspection, this practice has a documented history of more than two millennia in East Asian medical traditions and is founded on the principle that the tongue mirrors the overall physiological and pathological status of the body through observable attributes such as color, morphology, coating, and moisture [2]. According to TCM theory, the tongue is connected to the internal organs via meridian pathways, and changes in its appearance can signify various pathological states [3]. As articulated in classical TCM texts including the *Huangdi Neijing* (《黄帝内经》, *Inner Canon of Huangdi*) and *Aoshi Shanghan Jinjing Lu* (《敖氏伤寒金镜录》, *Ao's Golden Mirror Records on Cold Damage*), the "tongue is the sprout of the heart" and the "tongue is the external manifestation of the spleen and stomach", reflecting the belief that its visual characteristics correspond to the functional status of the cardiac and digestive systems [4].

These historical theoretical assertions are increasingly supported by contemporary biomedical evidence: variations in tongue color have been linked to altered microcirculation patterns detectable by capillaroscopy, while the microbiota of the tongue coating has been shown to reflect the gut microbial dysbiosis associated with metabolic disorders such as type 2 diabetes mellitus and non-alcoholic fatty liver disease [3]. Epithelial metabolomic analyses have further identified disease-specific tongue profiles [2], and tongue-microcirculation indices correlate with systemic inflammatory biomarkers [4]. These convergences between the classical TCM framework and modern bioinformatics features provide a compelling biological basis for artificial intelligence (AI)-assisted tongue diagnosis and reinforce its potential as an objective, quantitative, and reproducible assessment method.

Traditional tongue diagnosis is predominantly dependent on the subjective judgement of experienced practitioners, resulting in inter-observer variability and standardization challenges [4]. The absence of objective measurement tools has historically limited the broader clinical adoption and scientific validation of this age-old diagnostic method. Over recent decades, however, the accelerated advancement of AI, particularly in deep learning and computer vision, has unlocked new avenues for achieving objective, quantitative, and reproducible analysis of tongue images [5, 6].

The application of AI in tongue diagnosis has progressed considerably in the past decade. Early studies focused on basic image-processing techniques for color normalization and feature extraction, employing traditional machine learning algorithms such as support vector machine (SVM) and random forest (RF) [7, 8]. Subsequently, convolutional neural networks (CNNs) and their variants, including ResNet, visual geometry group network (VGG), and U-Net, have achieved remarkable performance in tongue-image segmentation, feature recognition, and disease classification [9-11]. More recent efforts have turned to transformer-based architectures and multimodal fusion approaches to further improve diagnostic accuracy [12, 13]. Despite this rapid advancement, a comprehensive evaluation that integrates bibliometric mapping of the research landscape with meta-analytic quantification of diagnostic performance and an assessment of clinical interpretability has not been conducted. Bibliometric analysis provides objective insights into publication trends, collaboration patterns, and emerging research themes [14, 15], whereas diagnostic test accuracy (DTA) meta-analysis offers statistically robust pooled estimates of diagnostic accuracy [16].

This study was therefore designed to characterize the current research landscape of AI-assisted tongue diagnosis through a comprehensive bibliometric analysis, and to evaluate its pooled diagnostic accuracy and clinical interpretability through a DTA meta-analysis, in order to inform the technology's clinical translation and to support the modernization of TCM.

2 Data and methods

2.1 Visualization analysis

2.1.1 Data source and search strategy The Web of Science Core Collection (WoSCC) was selected as the primary source for bibliometric analysis based on its comprehensive coverage of high-quality peer-reviewed publications and standardized metadata for citation-based studies [14]. WoSCC does not index Chinese-language journals, which could result in under-representation of Chinese-language literature, so the DTA meta-analysis search strategy also included the China National Knowledge Infrastructure (CNKI). The WoSCC search employed the following query: TS = ((“tongue diagnosis” OR “tongue inspection” OR “tongue image” OR “tongue coating” OR “Shezhen”) AND (“artificial intelligence” OR “machine learning” OR “deep learning” OR “computer vision” OR “image processing” OR “neural network*” OR “CNN”) AND (“diagnose*” OR “classific*” OR “predict*” OR “screen*”)).

2.1.2 Inclusion and exclusion criteria Records were eligible for inclusion if they were: (i) indexed in the WoSCC; (ii) articles or reviews; (iii) published in English; (iv) published between January 1, 2014 and December 31, 2025; and (v) topically relevant to AI-assisted tongue diagnosis as defined by the search query. Records were excluded if they were: (i) conference papers, letters, editorials, or book chapters, due to potential duplication with subsequently published articles, insufficient methodological detail, and inconsistent indexing in WoSCC; (ii) duplicate records identified during database export; or (iii) judged irrelevant after title and abstract screening.

2.1.3 Data processing and standardization Eligible records were exported from WoSCC in plain-text format containing complete bibliographic and cited-reference fields. Duplicate records were removed using the WoSCC built-in deduplication function, followed by manual verification. Standardization procedures were applied: (i) author-name disambiguation—variant spellings of the same author (e.g., “JT Xu” and “Jiatuo Xu”) were merged by cross-referencing institutional affiliations and Open Researcher and Contributor ID (ORCID); (ii) country and institutional normalization—inconsistent designations (e.g., “People’s Republic of China” “PR China” and “China”) were unified into a single standard form; (iii) keyword normalization—synonymous or abbreviated terms (e.g., “CNN” and “convolutional neural network”) were merged into their canonical forms. The cleaned dataset was imported into VOSviewer (version 1.6.20) and CiteSpace (version 6.2.R4) for network visualization and burst detection.

2.1.4 Bibliometric analyses Complementary bibliometric analyses were performed, including annual publication output, disciplinary distribution, journal and citation analysis, geographic, institutional, and authorship analysis, co-authorship networks at the country, institution, and author levels, keyword co-occurrence analysis (including author keywords that occurred at least three times after search-string terms had been removed), keyword burst (performed in CiteSpace using the Kleinberg burst-detection algorithm with the burst-strength parameter γ set to 0.5) and temporal evolution detection, and highly cited paper profiling. Frequency statistics and temporal trend analyses were performed in Python (version 3.10) using Pandas and Matplotlib.

2.2 Meta-analysis of diagnostic accuracy

2.2.1 Literature search strategy A comprehensive search of four electronic databases [Scopus, PubMed, Web of Science (WoS), and CNKI] was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA) guidelines [17]. The search covered studies published between January 1, 2014 and December 31, 2025, with no language restrictions. Searches were conducted across the title, abstract, and keyword fields of each database. The search strategy integrated terms representing three domains: tongue diagnosis (“tongue diagnosis” OR “tongue image” OR “tongue inspection” OR “Shezhen”) AND AI (“artificial intelligence” OR “machine learning” OR “deep learning” OR “convolutional neural network”) AND diagnostic accuracy (“sensitivity” OR “specificity” OR “diagnostic accuracy” OR “AUC” OR “area under the curve”). Search strings were adapted to the syntax of each database. Reference lists of included studies were manually checked to identify additional eligible records.

2.2.2 Inclusion and exclusion criteria Studies were eligible for inclusion if they: (i) evaluated AI-assisted tongue diagnosis for disease detection or classification; (ii) reported extractable diagnostic-accuracy metrics, including sensitivity, specificity, the area under the curve (AUC), or accuracy; (iii) used a clinical diagnosis or laboratory test as a reference standard; and (iv) were original research articles published in peer-reviewed journals. Studies were excluded if they: (i) lacked extractable diagnostic-accuracy data; (ii) focused on algorithm development without clinical validation; (iii) were duplicate publications or conference abstracts; or (iv) were review articles, case reports, or otherwise outside the scope of the present study.

2.2.3 Data extraction and quality assessment Data extraction was performed independently by two reviewers using a standardized form. The extracted variables comprised first author, publication year, country, study design, sample size, participant characteristics, disease type, tongue-image acquisition method, AI model architecture, reference standard, sensitivity, specificity, and AUC with 95% confidence intervals (CI). Disagreements were resolved through discussion and consensus. The methodological quality of included studies was assessed using the Quality Assessment of Diagnostic Accuracy

Studies version 2 (QUADAS-2) tool [18], which evaluates risk of bias across four domains: patient selection [Domain (D) 1], index test (D2), reference standard (D3), and flow and timing (D4).

2.2.4 Statistical analysis (i) Pooled diagnostic accuracy. A bivariate random-effects model, also known as the hierarchical summary receiver operating characteristic (HSROC) model, was employed to pool sensitivity and specificity, accounting for their intrinsic correlation [19]. From the bivariate estimates, the diagnostic odds ratio (DOR), positive likelihood ratio (+LR), and negative likelihood ratio (–LR) were derived.

(ii) Heterogeneity and threshold effect. Heterogeneity was quantified using the I^2 statistic and Cochran's Q test, with $I^2 > 75\%$ considered to represent substantial heterogeneity [20]. The potential for a threshold effect was evaluated by calculating Spearman's correlation coefficient between the logit of sensitivity and the logit of false-positive rate (1 – specificity) [21], with $|r| > 0.6$ indicating a clinically meaningful threshold effect.

(iii) Summary ROC (SROC) and publication bias. The SROC curve was derived from the HSROC bivariate model, and the area under the SROC curve (e.g., AUC) was calculated from the same model as a numerical summary of overall diagnostic discrimination. Publication bias was evaluated using Deeks' funnel plot asymmetry test, a method specifically designed for DTA meta-analyses [22].

(iv) Subgroup analysis and meta-regression. Subgroup analyses were conducted based on disease category (metabolic/hepatic disorders vs. oncological/oral lesions vs. other clinical applications), AI model architecture (deep learning models vs. traditional or ensemble machine learning models), and sample size, stratified into three categories: small ($n < 500$), medium ($500 \leqslant n \leqslant 1\,000$), and large ($n > 1\,000$). To explore potential sources of heterogeneity, meta-regression was performed using AI model type, sample size ($n > 1\,000$ vs. $n \leqslant 1\,000$), and publication year as covariates.

(v) Sensitivity analysis. To assess the robustness of the pooled estimates, sensitivity analyses were performed by (a) excluding studies with sample sizes < 200 participants; (b) restricting analysis to studies published from January 1, 2019 onward, when standardised imaging protocols became more widely adopted; and (c) comparing the bivariate random-effects model with a univariate random-effects model (DerSimonian-Laird estimator) applied separately to sensitivity and specificity, from which 95% prediction intervals (PI) were calculated to estimate the expected range of diagnostic accuracy in individual clinical settings.

(vi) Software and model fitting. All analyses were performed in Python using the SciPy and NumPy libraries. The bivariate random-effects model was fitted using maximum likelihood estimation with a bivariate normal prior applied to the logit-transformed sensitivity and specificity. Model convergence was verified by confirming that the gradient of the log-likelihood was negligible at the parameter estimates and that the Hessian was positive definite.

3 Results

3.1 Bibliometric landscape of AI-assisted tongue diagnosis

3.1.1 Rapid expansion of publication output across an interdisciplinary field A total of 198 publications meeting the eligibility criteria were retrieved from the WoSCC and constituted the final dataset for bibliometric analysis (Figure 1). Annual publication output increased 24.5-fold over the study period, from 2 in 2014 to 49 in 2025. Three distinct phases of growth were observed: an early exploration phase with no clear upward trend (2014 – 2017, n = 10; mean ≈ 2.5 publications per year), a steady accumulation phase (2018 – 2021, n = 59; mean ≈ 14.8 publications per year), and a rapid expansion phase (2022 – 2025, n = 129; mean ≈ 32.2 publications per year), with the latter phase accounting for 65.2% of all publications (Figure 2A). The publication count in 2024 (n = 19) appeared lower than the peak in 2025 (n = 49); a plausible explanation is that a substantial number of articles posted online in late 2024 received 2025 issue assignments under the WoSCC indexing process. Analysis of WoSCC subject categories confirmed the interdisciplinary nature of this field, with publications spanning computer science, biomedical engineering, medical informatics, and integrative and complementary medicine (Figure 2B).

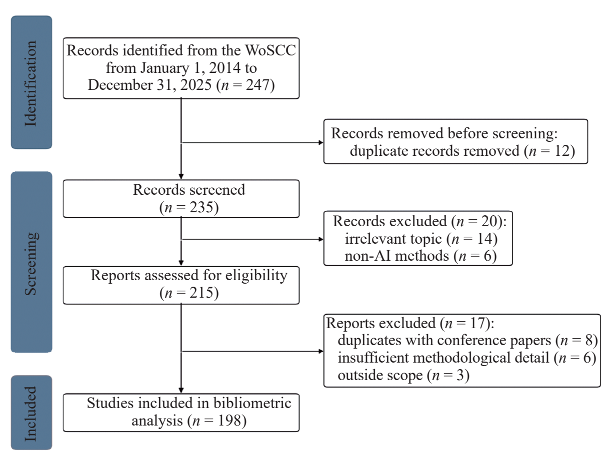

Figure 1 PRISMA-style flow diagram of the bibliometric search and screening process applied to the WoSCC dataset

3.1.2 Geographic, institutional, and authorship landscape A total of 31 countries/regions contributed to AI-assisted tongue-diagnosis research. China was the dominant contributor, accounting for 580 institutional

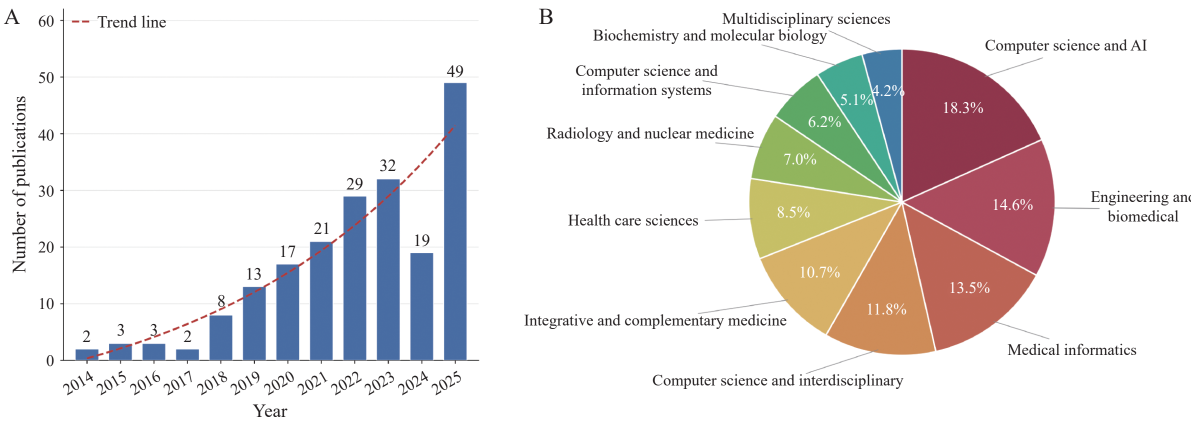

Figure 2 Annual publication output of AI-assisted tongue diagnosis research from 2014 to 2025 (A) and disciplinary distribution by WoSCC subject categories (B)

affiliations, approximately 83.5% of the 695 affiliation occurrences across all 198 publications, reflecting both the cultural significance of tongue diagnosis in TCM and substantial domestic research investment. Other Asian countries (notably India, Japan, and the South Korea) constituted a secondary tier of contributors. Western countries, including the USA, Germany, and the UK, also participated, indicating emerging international interest in this field. Institutional output was concentrated within China's leading universities of Chinese medicine, with Shanghai University of Traditional Chinese Medicine as the most productive contributor, followed by Beijing University of Chinese Medicine, Xiamen University, China Academy of Chinese Medical Sciences, and Tsinghua University (Figure 3).

Authorship analysis was integrated with the geographic and institutional results because country and institutional attributions were derived from author affiliations. The 15 most prolific authors are presented in Figure 4A; Jiatuo Xu (Shanghai University of Traditional Chinese Medicine) was the leading contributor, heading a core group of researchers affiliated with major Chinese universities of Chinese medicine. Author and institutional collaboration networks revealed distinct clusters centred on these universities (Figure 4B and 4C). In the country-collaboration network, the China and USA pair and China and Germany pair appeared as the most prominent international connections, as evidenced by their comparatively thicker co-authorship edges.

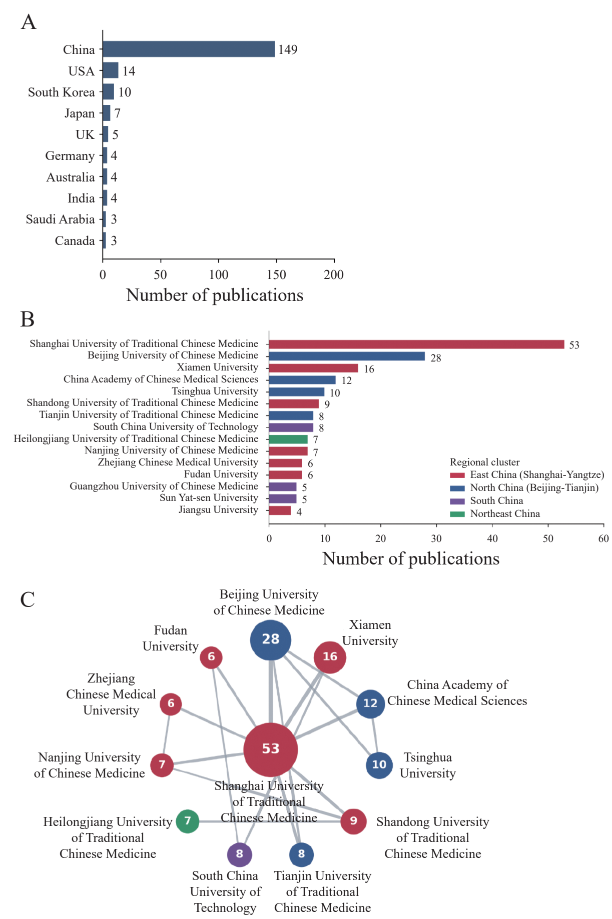

Figure 3 Geographic and institutional landscape of AI-assisted tongue diagnosis research

A, the top 10 contributing countries/regions ranked by publication count. B, the top 15 contributing institutions. C, institutional collaboration network among the top 12 institutions. Node size is proportional to the number of publications by the institution, and edge thickness is proportional to the co-authorship strength between two institutions.

3.1.3 Journal and citation analysis The 198 included publications were distributed across diverse engineering and biomedical journals, predominantly indexed in Science Citation Index (SCI), Science Citation Index Expanded (SCIE), and/or PubMed/MEDLINE. To characterise the publication outlets in greater depth, the top 15 source journals ranked by the number of relevant publications retrieved are presented in Table 1 together with their indexing status (SCI/SCIE/EI), country of publication, 2024 JCR impact factor (IF) and quartile, and the number of relevant publications retrieved. The most frequent outlet was *IEEE Access* (n = 15), followed by *Scientific Reports* (n = 10), *Frontiers in Physiology* (n = 8), *Biomedical Signal Processing and Control* (n = 6), and *Digital Health* (n = 5). Several methodologically rigorous

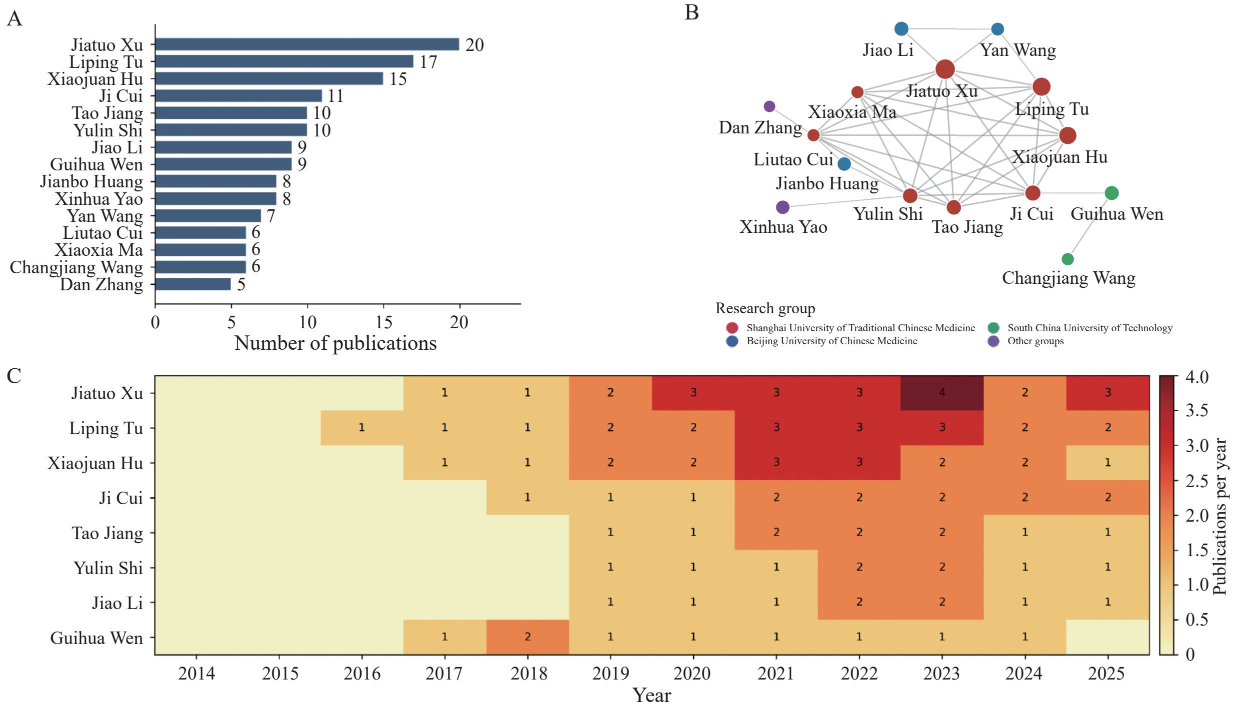

Figure 4 Authorship and collaboration patterns of AI-assisted tongue diagnosis research

A, the top 15 most prolific authors ranked by publication count. B, author co-authorship network. Node size is proportional to the publication count of each author, and node color denotes the research group and edge thickness is proportional to co-authorship strength. C, annual publication heatmap for the top 8 most prolific authors.

Table 1 Top 15 source journals for AI-assisted tongue diagnosis research ranked by the number of relevant publications retrieved with indexing status, country of publication, and 2024 Journal Citation Reports (JCR) metrics (released in 2025)

Rank	Journal	*n*	Indexing	Country/publisher	IF	Quartile
1	*IEEE Access*	15	SCIE, EI	USA/IEEE	3.6	Q2
2	*Scientific Reports*	10	SCIE, PubMed	UK/Nature Publishing Group	3.9	Q1
3	*Frontiers in Physiology*	8	SCIE, PubMed	Switzerland/Frontiers Media	3.4	Q1
4	*Biomedical Signal Processing and Control*	6	SCIE, EI	UK/Elsevier	4.9	Q2
5	*Digital Health*	5	SCIE, PubMed	UK/SAGE Publications	3.3	Q1
6	*Applied Sciences* (*Basel*)	4	SCIE, EI	Switzerland/MDPI	2.5	Q2
7	*Concurrency and Computation: Practice and Experience*	4	SCIE, EI	USA/Wiley	1.5	Q3
8	*Evidence-Based Complementary and Alternative Medicine*	4	Discontinued (SCIE)	UK/Hindawi	—	—
9	*Technology and Health Care*	3	SCIE, PubMed	Netherlands/IOS Press	1.8	Q3
10	*Neural Computing and Applications*	3	SCIE, EI	UK/Springer	6.52	Q2 (CS-AI)
11	*IEEE Journal of Biomedical and Health Informatics*	3	SCI, SCIE, EI, PubMed	USA/IEEE	6.8	Q1
12	*Chinese Medicine*	3	SCIE, PubMed	UK/BioMed Central	5.7	Q1
13	*European Journal of Integrative Medicine*	3	SCIE	Germany/Elsevier	1.7	Q3
14	*BioMed Research International*	3	SCIE, PubMed	UK/Hindawi	2.3	Q3
15	*Journal of Healthcare Engineering*	3	Discontinued (SCIE)	UK/Hindawi	—	—

EI, Engineering Index. —, not applicable (journal either has its IF temporarily withheld in 2024 JCR or has been discontinued from SCIE coverage). *n*, number of relevant publications retrieved. PubMed, indexed in PubMed/MEDLINE. Quartile, Journal Citation Reports quartile based on the journal's primary category. CS-AI, Computer Science and Artificial Intelligence (one of the multiple JCR subject categories under which the journal is indexed; here, the Q2 ranking is taken from this primary category).

Q1-ranked journals also featured prominently, including *IEEE Journal of Biomedical and Health Informatics* (IF = 6.8; Q1) and *Chinese Medicine* (IF = 5.7; Q1), indicating that the field has attracted attention from established methodological and clinical informatics communities.

To complement the journal-level analysis, the 15 most-cited publications are profiled in Table 2 with first author, year, abbreviated title, journal, and citation count (as recorded in WoSCC). The most highly cited paper was XU et al. [10], which introduced a multitask joint learning framework for automated tongue-image segmentation and classification (157 citations). Other highly cited works included the multicentre gastric-cancer cohort by YUAN et al. [23], the tooth-marked-tongue recognition study by WANG et al. [11], and the prediabetes feature-fusion model by LI et al. [7]. Most of the highly cited works originated from major Chinese universities of Chinese medicine and were published in technically rigorous engineering and informatics journals, supporting the methodological maturation of the field.

3.1.4 Four research themes and disease applications identified by keyword co-occurrence Keyword frequency analysis was performed after explicitly removing the search-string terms (e.g., “tongue diagnosis” and “tongue image”). The most frequent residual author keywords were deep learning, machine learning, TCM, AI, image segmentation, convolutional neural network, tongue segmentation, classification, feature extraction, diabetes, transfer learning, image classification, TCM constitution, computer-aided diagnosis, neural network, ResNet, U-Net, non-alcoholic fatty liver disease (NAFLD), attention mechanism, and gastric cancer (Figure 5A).

Co-occurrence network analysis (Figure 5B) revealed four major thematic clusters with distinct hub keywords and characteristic co-occurring terms: AI and deep learning architectures, was anchored by hub keywords “deep learning” “convolutional neural network” and “transformer”; image processing and segmentation, was anchored by “image processing” “image segmentation” and “feature extraction”; TCM-specific applications, was anchored by “tongue inspection” “syndrome differentiation” and “TCM constitution”; disease-specific applications, was anchored by “diabetes mellitus” “type 2 diabetes” and “prediabetes”, with co-occurring terms NAFLD, gastric cancer, hypertension, oral cancer, and thyroid nodules.

3.1.5 Temporal evolution of research hotspots The temporal keyword heatmap (Figure 6A) illustrates a progressive shift in research emphasis from traditional

Table 2 Top 15 most-cited publications on AI-assisted tongue diagnosis from the bibliometric dataset

Rank	First author	Year	Title	Journal	Citation
1	XU Q [10]	2020	Multi-task joint-learning model for segmenting and classifying tongue images using a deep neural network	*IEEE Journal of Biomedical and Health Informatics*	157
2	YUAN L [23]	2023	Development of a tongue image-based machine-learning tool for the diagnosis of gastric cancer: a prospective multicentre clinical cohort study	*eClinicalMedicine*	81
3	WANG X [11]	2020	Artificial intelligence in tongue diagnosis: using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark	*Computational and Structural Biotechnology Journal*	78
4	LI J [7]	2021	A tongue features fusion approach to predicting prediabetes and diabetes with machine learning	*Journal of Biomedical Informatics*	72
5	ZHUO L [24]	2014	An SA–GA–BP neural network-based color correction algorithm for TCM tongue images	*Neurocomputing*	53
6	LI J [25]	2021	Establishment of noninvasive diabetes risk prediction model based on tongue features and machine learning techniques	*International Journal of Medical Informatics*	51
7	TANIA MH [26]	2019	Advances in automated tongue diagnosis techniques	*Integrative Medicine Research*	50
8	ZHOU CG [27]	2019	TongueNet: accurate localisation and segmentation for tongue images using deep neural networks	*IEEE Access*	48
9	MA JJ [28]	2019	Complexity perception classification method for tongue constitution recognition	*Artificial Intelligence in Medicine*	47
10	ZHOU JH [29]	2019	TongueNet: U-Net with a morphological processing layer for fast tongue segmentation	*Applied Sciences (Basel)*	43
11	JIANG T [9]	2021	Tongue image quality assessment based on a deep convolutional neural network	*BMC Medical Informatics and Decision Making*	41
12	ZHUANG QB [30]	2022	Human-computer interaction based health diagnostics using ResNet34 for tongue image classification	*Computer Methods and Programs in Biomedicine*	36
13	HUANG ZH [31]	2022	A novel tongue segmentation method based on improved U-Net	*Neurocomputing*	36
14	JIANG T [32]	2021	Application of computer tongue image analysis technology in the diagnosis of NAFLD	*Computers in Biology and Medicine*	36
15	LI J [33]	2022	A multi-step approach for tongue image classification in patients with diabetes	*Computers in Biology and Medicine*	34

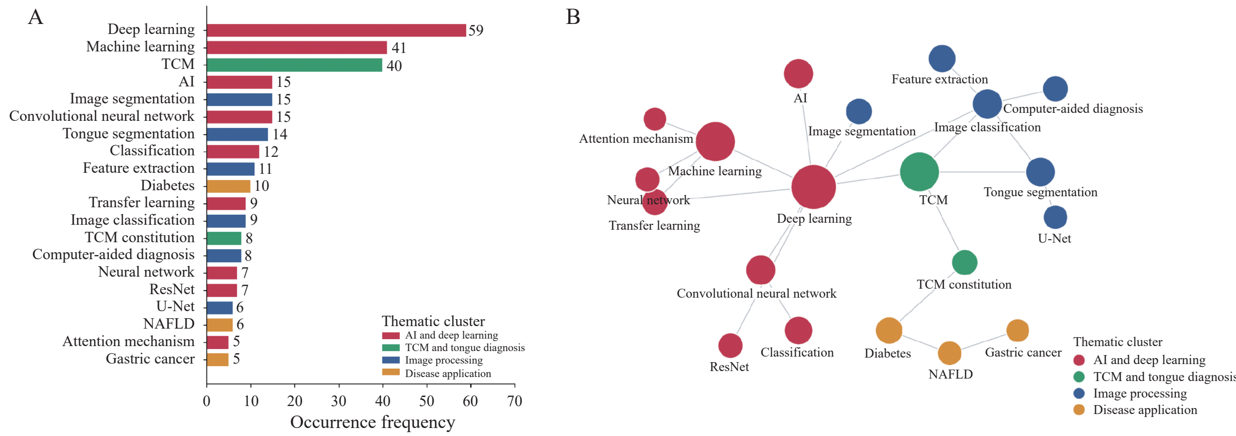

Figure 5 The top 20 keywords by co-occurrence frequency (A) and co-occurrence network (B) of AI-assisted tongue diagnosis research

Node size is proportional to keyword occurrence frequency, and edge thickness is proportional to co-occurrence strength.

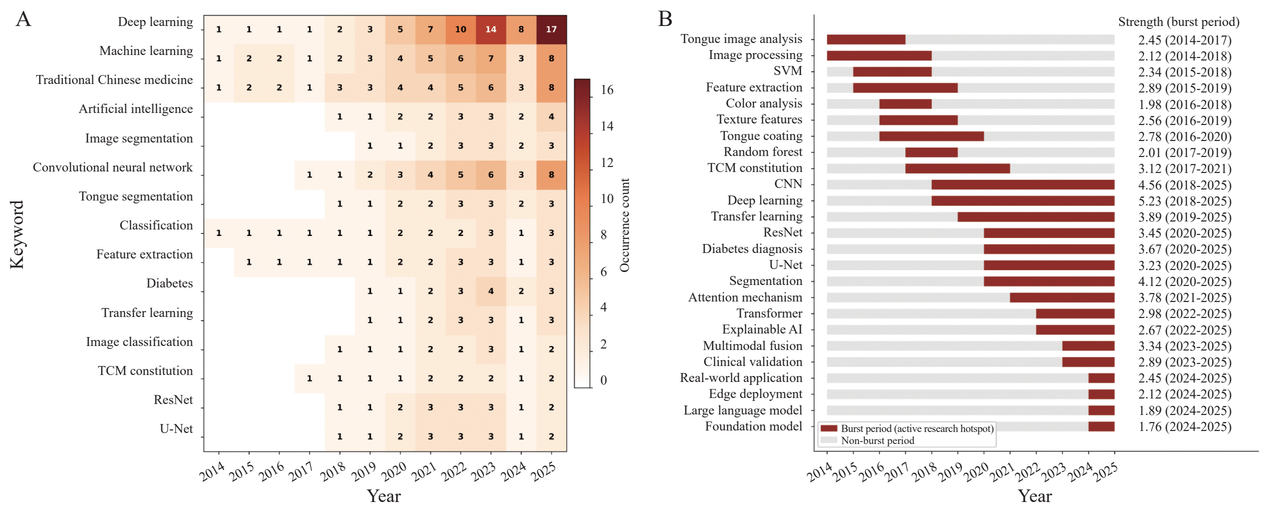

Figure 6 Temporal evolution of the top 15 keywords (A) and top 25 keywords with significant citation bursts (B)

machine-learning methods toward deep-learning approaches across the study period (2014 – 2025). Keyword burst analysis performed with CiteSpace identified the top 25 keywords with significant citation bursts (Figure 6B). Early research hotspots (2014 – 2018) centred on "image processing" "feature extraction" and "SVM". During 2018 – 2021, the focus shifted toward "deep learning" "CNN" and "transfer learning". Current hotspots (2022 – 2025) include "transformer" "attention mechanism" "explainable AI" and "multimodal fusion", indicating the adoption of advanced AI architectures and the rising importance of clinical interpretability. Among disease-specific applications, "diabetes diagnosis" "TCM constitution identification" "liver fibrosis screening" and "oral cancer detection" have all gained increasing research attention since 2022.

3.2 DTA meta-analysis and clinical interpretability

3.2.1 Study characteristics Following the PRISMA-DTA guidelines, the search and screening process yielded 16 eligible studies for inclusion in the DTA meta-analysis (Figure 7), collectively enrolling 14 755 participants. The included studies spanned multiple disease domains, paralleling the multi-disease keyword cluster identified in Section 3.1.4: NAFLD/metabolic dysfunction-associated fatty liver disease (MAFLD)/liver fibrosis ($n = 5$), oral cancer/oral lesion classification ($n = 2$), pulmonary-nodule malignancy ($n = 1$), cardiovascular risk prediction ($n = 1$), tongue radiomics for insomnia degree assessment ($n = 1$), traditional Thai medicine constitution ($n = 1$), spotted-tongue recognition ($n = 1$), general tongue classification ($n = 1$), and primary diabetes diagnosis ($n = 3$). Regarding geographic distribution, 12 studies were from China, with the remaining 4 from Turkey ($n = 1$), Thailand ($n = 1$), and India ($n = 2$). Sample sizes ranged from 166 to 2 895 participants (median = 720; interquartile range: 530 – 1 350) across studies reporting a numerical sample size; four studies did not report a primary sample-size figure in the extracted data and were retained on the basis of reporting both sensitivity and specificity. AI architectures spanned CNN-based models, generative adversarial

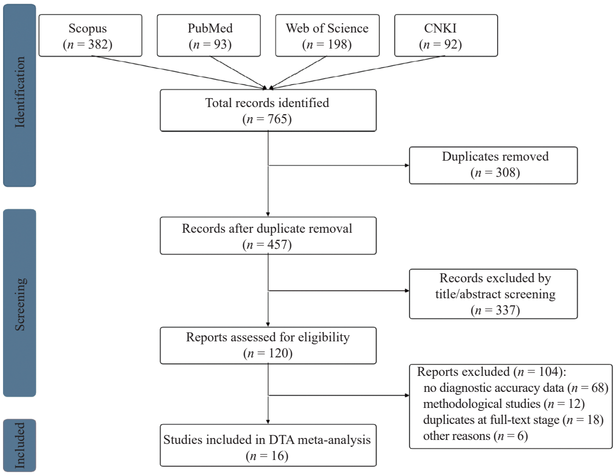

Figure 7 PRISMA-DTA flow diagram of the DTA meta-analysis study selection

networks (GAN), multi-task learning, transfer-learning CNN, and hybrid deep Q-network and CNN-SVM approaches. Reported sensitivity ranged from 0.69 to 0.98 (median = 0.91), and specificity ranged from 0.85 to 0.99 (median = 0.94). Detailed study characteristics are summarised in Table 3.

3.2.2 Pooled diagnostic performance and clinical likelihood The forest plots for sensitivity and specificity are presented in Figure 8A and 8B. The pooled sensitivity was 90.3% (95% CI: 86.7% – 93.1%; I^2 = 95.8%, P < 0.000 1), and the pooled specificity was 93.0% (95% CI: 90.6% – 94.7%; I^2 = 92.1%, P < 0.000 1). The I^2 values for sensitivity and specificity substantially exceeded the conventional 75% threshold for substantial heterogeneity, indicating extreme between-study variability. The pooled point estimates should therefore be interpreted as average

Table 3 Characteristics of studies included in the DTA meta-analysis

No.	First author	Year	Country	*N*	Disease	AI model	Sensitivity	Specificity
1	JIANG T [32]	2021	China	1 778	NAFLD	Multi-feature machine learning (LR, SVM, RF) + serological indices	0.797	0.852
2	DAI S [34]	2024	China	166	MAFLD	GAN + 16S rRNA microbiome features	0.946	0.987
3	LU XZ [35]	2025	China	1 601	Liver fibrosis	Tongue image deep-learning model (multi task)	0.862	0.874
4	WANG RR [36]	2024	China	720	NAFLD	Tongue image features + tongue-coating microbiota; machine learning classifier	0.914	0.852
5	ZHANG JJ [37]	2025	China	1 350	MAFLD with CAD risk	Multimodal fusion (tongue images + clinical indicators)	0.778	0.908
6	PENG CD [38]	2022	China	—	Spotted-tongue (multiple diseases)	Multiscale CNN	0.882	0.942
7	MA Q [39]	2025	China	530	Pulmonary-nodule malignancy	Oral microbiota-based machine-learning prediction model	0.692	0.881
8	BALASUBRAMANIYAN S [40]	2022	India	1 250	Type 2 diabetes mellitus	CNN (ResNet-50) feature extraction + deep radial basis function neural network (RBFNN) classifier	0.873	0.930
9	YUAN CH [41]	2025	China	2 895	Cardiovascular risk	Dynamic weighted ensemble learning	0.949	0.962
10	LIU PJ [42]	2025	China	715	Oral cancer	19-Layer convolutional neural network (smartphone images)	0.957	0.962
11	COŞGUN BAYBARS S [43]	2025	Turkey	623	Coated tongue (oral diagnosis)	CNN	0.882	0.850
12	DAMKLIANG K [44]	2025	Thailand	741	Thai-medicine constitution	Transfer-learning CNN (Tri-Dhat classification)	0.960	0.970
13	YE R [45]	2024	China	411	Insomnia degree (radiomics)	Tongue-imaging-based radiomics	0.780	0.916
14	MATHEW JK [46]	2023	India	—	Diabetes	ExpACVO-optimised hybrid Deep Q-Network	0.960	0.960
15	LIU XY [47]	2025	China	—	General tongue classification	Lightweight CNN-SVM hybrid	0.920	0.950
16	DEEPA SN [48]	2021	India	—	Diabetes	CNN DenseNet + particle-swarm optimisation (PSO)-tuned SVM	0.984	0.967

N, number of participants in the study. —, not reported in the source publication.

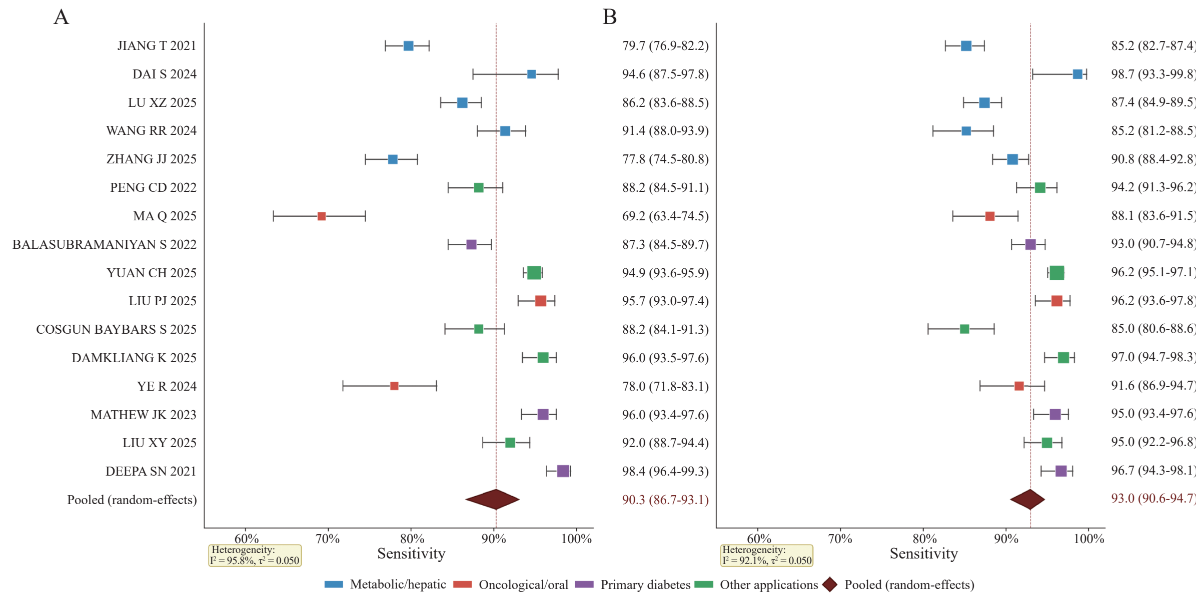

Figure 8 Forest plots of sensitivity (A) and specificity (B) for AI-assisted tongue diagnosis across the 16 included studies

performance across markedly different study conditions and disease domains rather than as stable, generalizable diagnostic parameters. A supplementary univariate random-effects model (DerSimonian-Laird estimator) yielded 95% PI of 66.2% – 97.8% for sensitivity and 78.1% – 98.0% for specificity, indicating that the true diagnostic performance of AI-assisted tongue diagnosis in a given clinical setting may deviate substantially from the pooled estimates.

The pooled diagnostic odds ratio (DOR) was 123.16 (95% CI: 76.42 – 198.46; I^2 = 97.9%), substantially exceeding the conventional threshold of 25. The +LR was 12.82, indicating that a positive test result increases the odds of disease by approximately 13-fold. The –LR was 0.104, indicating that a negative test result reduces the odds of disease by approximately 10-fold. Both likelihood ratios approximately met conventional thresholds for clinically useful diagnostic tests (+LR > 10 and –LR ≈ 0.1). A summary of all pooled diagnostic accuracy estimates is presented in Table 4.

3.2.3 SROC curve, sample-size effect, and publication bias The SROC curve demonstrated excellent overall discriminatory performance, with an AUC of 0.961 (Figure 9A), positioning AI-assisted tongue diagnosis above the 0.90 threshold widely regarded as indicative of excellent discrimination [22]. Deeks' funnel plot asymmetry test indicated no significant publication bias (P = 0.258; Figure 9B). The relationship between sample size and diagnostic accuracy is illustrated in Figure 9C; both sensitivity and specificity exhibited a modest stabilization with increasing sample size, with no systematic trend toward inflated accuracy in smaller studies.

3.2.4 Heterogeneity and subgroup analysis Substantial heterogeneity was observed for both sensitivity (I^2 = 95.8%) and specificity (I^2 = 92.1%). No significant

Table 4 Summary of pooled diagnostic accuracy measures from the DTA meta-analysis

Measure	Estimate	95% CI	I^2	P value
Sensitivity	90.3%	86.7% – 93.1%	95.8%	< 0.000 1
Specificity	93.0%	90.6% – 94.7%	92.1%	< 0.000 1
Diagnostic OR	123.16	76.42 – 198.46	97.9%	< 0.000 1
Positive LR	12.82	—	—	—
Negative LR	0.104	—	—	—
AUC	0.961	—	—	—

OR, odds ratio. —, not applicable.

threshold effect was detected (Spearman's r = 0.18, P = 0.51), indicating that heterogeneity likely arose from non-threshold sources, such as differences in target disease, study populations, AI model architectures, imaging-acquisition protocols, and reference-standard definitions.

Subgroup analyses based on the 16 studies included in the DTA meta-analysis by disease category, AI model architecture, and sample size are summarised in Table 5. (i) Disease category: studies on metabolic and hepatic disorders (NAFLD/MAFLD/liver fibrosis; n = 5) yielded a pooled sensitivity of 86.4% and specificity of 88.3%; oncological and oral-lesion studies (n = 3) yielded a pooled sensitivity of 81.4% and specificity of 91.6%; primary diabetes-diagnosis studies (n = 2) reported a pooled sensitivity of 96.9% and specificity of 96.3% (this subgroup result should be interpreted with caution given the small number of primary diabetes-diagnosis studies, which limits the statistical power and stability of the corresponding pooled estimates); and the remaining studies (n = 6) yielded a pooled sensitivity of 91.6% and specificity of 93.5%. (ii) AI model architecture: deep-learning models (n = 12) achieved marginally higher pooled performance than traditional or ensemble machine-learning approaches (n = 4); meta-regression confirmed that model

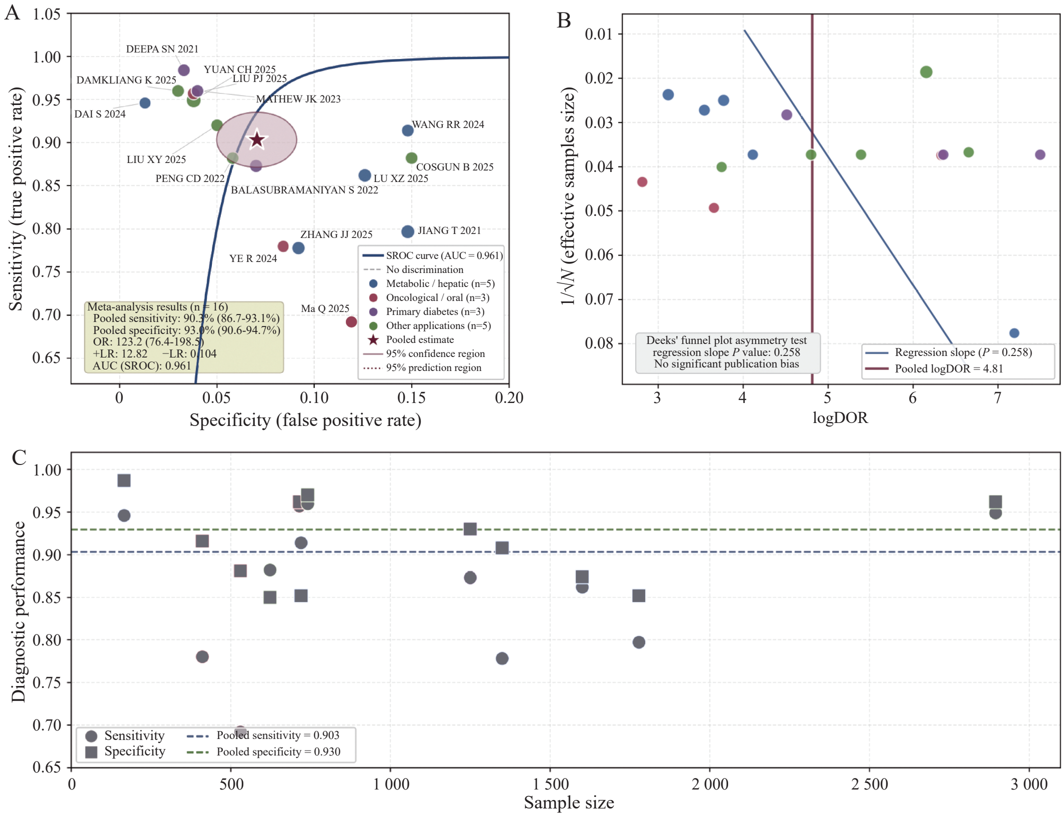

Figure 9 Diagnostic performance of AI-assisted tongue diagnosis

A, SROC curve for AI-assisted tongue diagnosis. B, Deeks' funnel plot for publication bias assessment. C, relationship between sample size and diagnostic accuracy across the included studies.

Table 5 Subgroup analyses of pooled diagnostic accuracy by disease category, AI model architecture, and sample-size strata

Subgroup	Stratum	n	Pooled sensitivity (95% CI)	Pooled specificity (95% CI)	I^2 (sensitivity/ specificity)	Meta-regression P value
Disease category	Metabolic & hepatic disorders (NAFLD/MAFLD/liver fibrosis)	5	86.4% (78.6% – 91.7%)	88.3% (81.7% – 92.7%)	97.0%/92.4%	Reference
	Oncological & oral-lesion applications (oral cancer, pulmonary-nodule malignancy, tongue radiomics)	3	81.4% (69.9% – 89.3%)	91.6% (87.5% – 94.4%)	94.5%/78.0%	$P = 0.310$
	Primary diabetes-diagnosis studies	2	96.9% (90.4% – 99.0%)	96.3% (92.7% – 98.2%)	—	$P = 0.038$
	Other applications (cardiovascular risk, TCM/Thai-medicine constitution, spotted-tongue, coated-tongue, general)	6	91.6% (88.0% – 94.2%)	93.5% (90.6% – 95.6%)	94.8%/89.7%	$P = 0.120$
AI model architecture	Deep-learning models (CNN-based, GAN, multi-task learning, transfer-learning CNN)	12	91.0% (87.4% – 93.7%)	93.4% (90.9% – 95.3%)	97.0%/94.7%	$P > 0.500$
	Traditional/ensemble machine-learning approaches	4	88.3% (78.5% – 94.0%)	91.9% (85.6% – 95.6%)	96.8%/94.0%	—
Sample size	Small ($n < 500$)	4	84.0% (74.5% – 90.5%)	89.5% (81.4% – 94.3%)	95.2%/91.5%	$P = 0.12$ (sensitivity); 0.18 (specificity)
	Medium ($500 \leqslant n \leqslant 1\,000$)	5	87.5% (81.5% – 91.8%)	91.5% (87.3% – 94.5%)	96.2%/93.0%	—
	Large ($n > 1\,000$)	7	92.1% (89.0% – 94.4%)	94.0% (91.5% – 95.8%)	97.4%/95.6%	—
Publication year	Continuous, per year	16	—	—	—	$P = 0.10$ (sensitivity); 0.041 (specificity)

n, number of studies in each subgroup. —, statistic not computed.

architecture was not a significant moderator. (iii) Sample size: across the three sample-size strata, pooled sensitivity was 84.0%, 87.5%, and 92.1%, and pooled specificity was 89.5%, 91.5%, and 94.0%, for small ($n < 500$), medium ($500 \leqslant n \leqslant 1\,000$), and large ($n > 1\,000$) studies, respectively. Publication year showed a positive association with specificity ($P = 0.041$), indicating significantly higher pooled specificity in more recently published studies. However, the corresponding coefficient for sensitivity was not statistically significant ($P = 0.10$). This pattern may be attributable, at least in part, to progressive improvements in the standardization of imaging acquisition protocols. No single covariate fully explained the observed between-study variability.

3.2.5 Methodological quality of included studies by QUADAS-2 tool The QUADAS-2 assessment is visualized as a traffic-light plot at the study level (Figure 10) and as a domain-level summary bar chart (Figure 11). Overall, methodological quality was acceptable: no included study was rated as high risk in any domain. Patient selection (D1) and flow and timing (D4) emerged as the two domains with the most prevalent unclear-risk concerns (31.2% and 25.0% of studies, respectively), primarily because several studies did not explicitly describe consecutive enrolment or did not clearly report the time interval between tongue-image acquisition and ascertainment of the reference standard. Index test (D2) and reference standard (D3) were predominantly judged at low risk (100.0% and 81.2% low risk, respectively).

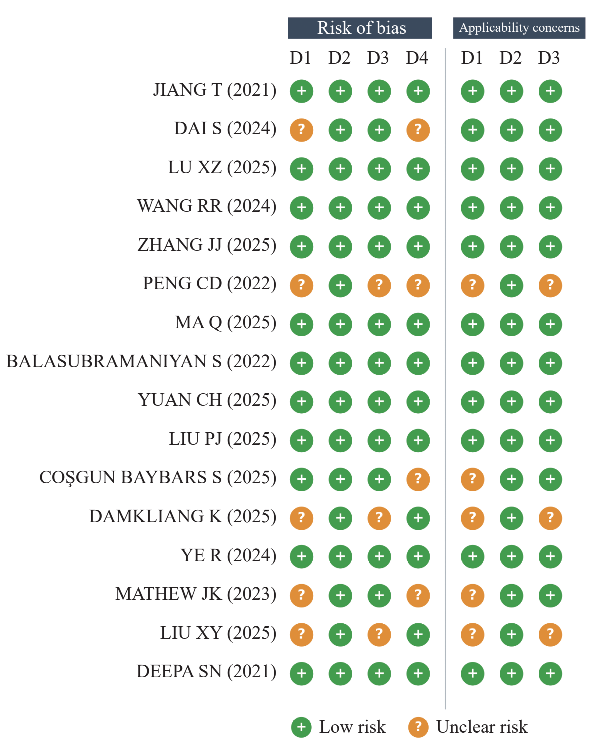

Figure 10 QUADAS-2 traffic-light plot of risk of bias and applicability concerns for the 16 included studies

Figure 11 QUADAS-2 domain-level summary of risk of bias and applicability concerns across all 16 included studies

4 Discussion

4.1 Main findings of bibliometric and DTA meta-analysis

This integrated bibliometric and DTA meta-analytic evaluation maps the research landscape of AI-assisted tongue diagnosis and quantifies its diagnostic performance. The bibliometric analysis showed that the field has experienced a substantial 24.5-fold expansion in publication output between 2014 and 2025, with marked acceleration during 2022 – 2025 and a clear thematic transition from traditional machine-learning approaches to deep-learning, transformer-based, and multimodal-fusion architectures.

The DTA meta-analysis demonstrated promising pooled diagnostic performance (sensitivity = 90.3%, specificity = 93.0%, DOR = 123.16, +LR = 12.82, –LR = 0.104, AUC = 0.961), with both likelihood ratios meeting conventional thresholds for clinically useful diagnostic tests. By integrating these quantitative findings with the bibliometric mapping, the present study renders the clinical interpretability of AI-assisted tongue diagnosis more objectively and comprehensively visible. At the same time, the substantial heterogeneity ($I^2 > 95\%$) and the exclusive reliance on Chinese cohorts circumscribe the generalisability of these estimates, which should therefore be interpreted as indicative rather than definitive, pending validation in internationally diverse, multi-ethnic populations.

4.2 Paradigm shift of AI-assisted tongue diagnosis and global health implications

Our findings highlight a paradigm shift from subjective, experience-dependent assessment to objective, reproducible, and quantifiable evaluation. This shift addresses a longstanding barrier to integrating traditional diagnostic methods into evidence-based healthcare systems [49]. The substantial inter-observer variability inherent to conventional tongue diagnosis, with disagreement rates of approximately 30% reported even among experienced practitioners [50], has historically constrained its scientific credibility and clinical applicability. AI technologies fundamentally overcome this limitation by providing consistent, bias-free assessments that can be standardised across practitioners and clinical settings.

Within this paradigm, the diagnostic accuracy of AI-assisted tongue diagnosis is comparable to that of established screening modalities. Conventional screening for metabolic and hepatic disorders, such as glycated haemoglobin (HbA1c) and fasting plasma glucose for diabetes, or transient elastography for liver fibrosis, requires blood collection or specialised equipment. In contrast, AI-assisted tongue diagnosis delivered comparable accuracy in a rapid and easily accessible manner. Viewed within the broader landscape of AI-assisted medical imaging, its performance is consistent with benchmarks reported for deep learning in digital pathology and other diagnostic imaging domains [51-54]. This evidence supports its clinical viability as a complementary first-line screening approach, particularly in primary-care and community-based settings where easily accessible, painless decision support is needed.

The biological plausibility of AI-assisted tongue diagnosis is further substantiated by converging multi-omics evidence. Tongue colour correlates with microcirculatory status; tongue-coating microbiome composition mirrors gut microbial dysbiosis in metabolic disorders [3]; and epithelial metabolomics reveals disease-specific shifts in glycolytic and amino-acid profiles [2]. Moreover, a “tongue-microcirculation axis” links sublingual vessel morphology to systemic microvascular health [3, 4]. Together, these mechanistic pathways support a translational framework in which TCM tongue diagnosis, AI-driven feature extraction, and multi-omics profiling for disease characterization converge.

When integrated with global health priorities, the demonstrated diagnostic performance suggests that AI-assisted tongue diagnosis could meaningfully contribute to addressing the global burden of undiagnosed metabolic disorders, oral and oropharyngeal cancers, and thyroid abnormalities, which together account for a substantial proportion of preventable morbidity worldwide [42, 55, 56]. The smartphone-deployable, point-of-care nature of AI-assisted tongue diagnosis aligns with World Health Organization (WHO) priorities for scalable, accessible screening in resource-limited settings [57, 58].

4.3 Clinical implications and implementation considerations

The demonstrated diagnostic performance supports the implementation of AI-assisted tongue diagnosis as a complementary, easily accessible decision-support tool. The high pooled sensitivity (90.3%) renders it suitable for population-level screening, while the high pooled specificity (93.0%) helps to limit unnecessary follow-up testing. As supported by the keyword co-occurrence pattern, current evidence is concentrated on metabolic–hepatic disorders, oncological and oral-lesion applications, and TCM constitution identification; the clinical uptake of AI-assisted tongue diagnosis should therefore be positioned within these contexts. This technology should complement, not replace, clinical judgement and should be deployed as a decision-support tool consistent with Food and Drug Administration (FDA) regulatory frameworks for AI-based Software as a Medical Device (SaMD) [59, 60]. Future integration of multimodal data (including vital signs, medical history, and biomarkers) within foundation-model architectures is expected to further improve diagnostic performance [61, 62]; in line with this trend, recent study has demonstrated AI-driven multimodal fusion of tongue imaging with adjunct clinical biomarkers for stratifying coronary artery disease risk in patients with metabolic-dysfunction-associated fatty liver disease [37]. Realizing the global health potential of this technology will additionally require deliberate efforts to ensure algorithmic fairness across diverse populations, given that AI models trained predominantly on a single demographic group may exhibit reduced performance when applied to other populations [63, 64].

4.4 The imperative in AI-assisted clinical diagnosis interpretability

A further key insight that emerges from the integrative bibliometric and meta-analytic perspective is that the clinical interpretability of AI-assisted tongue diagnosis can now be more objectively and comprehensively mapped, which is precisely the most-anticipated value of AI-assisted diagnosis in clinical applications. Keyword burst analysis indicated that “explainable AI” emerged as a research hotspot from 2022 onward, signifying the rising importance of, but still developmental status of, clinical interpretability in this field. The inherent “black-box” nature of several deep-learning models presents a substantial barrier to clinical adoption, as healthcare professionals require transparent decision pathways to trust and effectively incorporate AI-driven insights [65].

The interpretability problem in AI-assisted tongue diagnosis presents unique complexities and opportunities. Unlike conventional medical imaging, where diagnostic features often correspond to established anatomical or pathological markers, tongue diagnosis operates within a TCM-specific theoretical framework, incorporating constructs such as Qi stagnation, blood stasis, and Yin-Yang imbalance, which lack direct equivalents in Western biomedical ontology [66]. Developing explainable AI approaches that bridge these conceptual systems represents both a substantial methodological problem and an unprecedented opportunity for cross-cultural medical knowledge synthesis. Recent advances in attention mechanisms, gradient-weighted class activation mapping (Grad-CAM), and concept-based explainability methods offer promising avenues toward interpretable tongue diagnosis [67].

4.5 Future research trends

Our integrated bibliometric and meta-analytic approach maps the clinical interpretability of AI-assisted tongue diagnosis in a more objective and comprehensive manner. The same analysis also clarifies several open challenges that merit prioritisation in future work. First, although the DTA evidence base spans multiple disease domains (including NAFLD/MAFLD, oral cancer, pulmonary-nodule malignancy, cardiovascular risk, and diabetes) the number of studies per disease remains small, and the 16 studies are heterogeneous in target condition and reference standard. Notably, the second-most-cited paper in the bibliometric dataset by YUAN et al. [23] evaluated AI-assisted tongue diagnosis for the early detection of gastric cancer in a multicentre prospective cohort; in parallel, more recent 2024 – 2025 work has further extended AI-assisted tongue diagnosis to a wide range of indications, including microbiome-coupled diagnosis of MAFLD [34], non-invasive diagnostic feature extraction in NAFLD [36], oral microbiota–based prediction of pulmonary-nodule malignancy [39], smartphone-based detection of oral cancer [42], TCM constitution identification [68], home-based liver fibrosis monitoring [35], and coronary artery disease risk stratification in MAFLD [37]. Systematic validation across this broader range of diseases and TCM-specific health states is essential for establishing the true clinical utility of AI-assisted tongue diagnosis. Second, all 16 included DTA studies were conducted in or led by Chinese research groups, with limited representation from other countries, which constrains the generalisability of the findings; multi-centre, international validation studies in populations with different backgrounds and disease-prevalence patterns are clearly needed [69]. Third, the substantial heterogeneity observed (I^2 > 95%) reflects significant variability in image-acquisition protocols, preprocessing methods, and AI architectures; the development of standardised imaging protocols, including lighting, camera positioning, color calibration, and tongue posture, would enhance reproducibility and facilitate more robust meta-analytic comparisons [70]. Fourth, the current evidence base is dominated by cross-sectional designs; prospective cohort studies that assess the prognostic value of tongue-based AI assessment, particularly for monitoring disease progression and treatment response, are needed to strengthen the evidence for clinical implementation [71]. Fifth, as AI-assisted tongue diagnosis moves closer to clinical deployment, the establishment of clear regulatory frameworks becomes essential; while the SaMD pathway provides a general template, specific guidance for TCM-derived AI applications remains underdeveloped [72]. An important future direction is the application of AI-assisted tongue diagnosis to TCM syndrome differentiation; current studies focus almost exclusively on biomedical endpoints, and extending validation to syndrome patterns such as Qi deficiency, blood stasis, and damp-heat would better evaluate the technology within its original clinical context.

4.6 Strengths and limitations of the current evaluation

This study has several notable strengths. First, this work represents a comprehensive integrated evaluation that combines bibliometric mapping with a DTA meta-analysis of AI-assisted tongue diagnosis, providing both a landscape overview and a quantitative performance assessment of clinical interpretability. Second, the bibliometric analysis employed multiple validated tools (Bibliometrix, VOSviewer, and CiteSpace), enabling a multifaceted examination of research evolution, collaboration patterns, and emerging trends. Third, the DTA meta-analysis adhered to PRISMA-DTA guidelines and used appropriate bivariate random-effects models that account for the intrinsic correlation between sensitivity and specificity in diagnostic studies [19]. Fourth, the comprehensive search strategy spanned multiple databases in both English and Chinese, enhancing the completeness of the evidence retrieved.

Several limitations of this evaluation should be acknowledged. First, the bibliometric analysis was restricted to WoSCC, which may have excluded relevant publications in Chinese-language journals or databases not covered by this index, although this limitation was partly mitigated by the inclusion of CNKI in the DTA search strategy. Second, despite a non-significant Deeks' test, publication bias cannot be definitively ruled out, as studies with positive results tend to be more frequently published. Third, the relatively small number of studies included in the DTA meta-analysis and the heterogeneous distribution across disease domains limit the statistical power of the disease-specific subgroup analyses. Fourth, while QUADAS-2 was used for quality assessment, the application of the more recently developed QUADAS-AI tool [72] and adherence to Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) + AI reporting guidelines would have further enhanced the methodological rigour and transparency specific to AI-based diagnostic-test evaluation.

5 Conclusion

This integrated bibliometric and DTA meta-analytic evaluation indicates that AI-assisted tongue diagnosis has rapidly evolved from early image-processing approaches to advanced deep-learning, transformer-based, and multimodal-fusion architectures, with pooled diagnostic performance that is comparable to established screening modalities across multiple disease domains. The integrative perspective renders the clinical interpretability of AI-assisted tongue diagnosis more objectively and comprehensively visible, supporting its potential as a complementary, easily accessible, and rapid decision-support tool for population-level screening within the framework of TCM modernisation. Substantial heterogeneity and the dominance of Chinese cohorts indicate that international, multi-ethnic external validation, standardised image-acquisition protocols, and the integration of explainable AI methods are necessary before broader clinical translation.

Fundings

Hubei Provincial Science and Technology Plan Project (2025CCB018).

Author contributions

Lihui Liu: conceptualization, data curation, formal analysis, visualization, and writing – original draft. Yana Zhou: conceptualization, funding acquisition, project administration, supervision, and writing – review and editing. Kaiwen Hu: methodology, formal analysis, supervision, and writing – review and editing. All authors approved the submission and take responsibility for this manuscript.

Competing interests

The authors declare no competing interests.

References

[1] JIANG M, LU C, ZHANG C, et al. Syndrome differentiation in modern research of traditional Chinese medicine. Journal of Ethnopharmacology, 2012, 140(3): 634–642.
[2] ZHAO Y, NIE C, WANG L, et al. Application of metabolomics in tongue manifestation of traditional Chinese medicine. Evidence-Based Complementary and Alternative Medicine, 2013, 2013: 204908.
[3] LI Y, JIANG B, CHEN G, et al. Quantitative analysis of tongue color and microcirculation in patients with cardiovascular disease. Frontiers in Cardiovascular Medicine, 2021, 8: 730203.
[4] LIN H, ZHOU Y, HU K, et al. Tongue coating microbiota and systemic inflammatory biomarkers: a multi-omics review on the tongue-microcirculation axis. Chinese Medicine, 2025, 20(1): 162.
[5] RAJPURKAR P, CHEN E, BANERJEE O, et al. AI in health and medicine. Nature Medicine, 2022, 28(1): 31–38.
[6] TOPOL EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 2019, 25(1): 44–56.
[7] LI J, YUAN P, HU XJ, et al. A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. Journal of Biomedical Informatics, 2021, 115: 103693.
[8] ZHOU SK, GREENSPAN H, DAVATZIKOS C, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE, 2021, 109(5): 820–838.
[9] JIANG T, HU XJ, YAO XH, et al. Tongue image quality assessment based on a deep convolutional neural network. BMC Medical Informatics and Decision Making, 2021, 21(1): 147.
[10] XU Q, ZENG Y, TANG WJ, et al. Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network. IEEE Journal of Biomedical and Health Informatics, 2020, 24(9): 2481–2489.
[11] WANG X, LIU JW, WU CY, et al. Artificial intelligence in tongue diagnosis: using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark. Computational and Structural Biotechnology Journal, 2020, 18: 973–980.
[12] WAGNER SJ, REISENBÜCHLER D, WEST NP, et al. Transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study. Cancer Cell, 2023, 41(9): 1650–1661.
[13] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge. Nature, 2023, 620(7972): 172–180.
[14] DONTHU N, KUMAR S, MUKHERJEE D, et al. How to conduct a bibliometric analysis: an overview and guidelines. Journal of Business Research, 2021, 133: 285–296.
[15] ARIA M, CUCCURULLO C. Bibliometrix: an R-tool for comprehensive science mapping analysis. Journal of Informetrics, 2017, 11(4): 959–975.
[16] LEEFLANG MMG, DEEKS JJ, GATSONIS C, et al. Systematic reviews of diagnostic test accuracy. Annals of Internal Medicine, 2008, 149(12): 889–897.
[17] MCINNES MDF, MOHER D, THOMBS BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA, 2018, 319(4): 388.
[18] WHITING PF, RUTJES AWS, WESTWOOD ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 2011, 155(8): 529–536.
[19] REITSMA JB, GLAS AS, RUTJES AWS, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology, 2005, 58(10): 982–990.
[20] HIGGINS JPT, THOMPSON SG, DEEKS JJ, et al. Measuring inconsistency in meta-analyses. BMJ, 2003, 327(7414): 557–560.
[21] MOSES LE, SHAPIRO D, LITTENBERG B. Combining independent studies of a diagnostic test into a summary ROC curve:

data-analytic approaches and some additional considerations. Statistics in Medicine, 1993, 12(14): 1293–1316.

[22] DEEKS JJ, MACASKILL P, IRWIG L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. Journal of Clinical Epidemiology, 2005, 58(9): 882–893.

[23] YUAN L, YANG L, ZHANG SC, et al. Development of a tongue image-based machine learning tool for the diagnosis of gastric cancer: a prospective multicentre clinical cohort study. eClinicalMedicine, 2023, 57: 101834.

[24] ZHUO L, ZHANG J, DONG P, et al. An SA–GA–BP neural network-based color correction algorithm for TCM tongue images. Neurocomputing, 2014, 134: 111–116.

[25] LI J, CHEN QG, HU XJ, et al. Establishment of noninvasive diabetes risk prediction model based on tongue features and machine learning techniques. International Journal of Medical Informatics, 2021, 149: 104429.

[26] TANIA MH, LWIN K, HOSSAIN MA. Advances in automated tongue diagnosis techniques. Integrative Medicine Research, 2019, 8(1): 42–56.

[27] ZHOU CG, FAN HY, LI ZY. Tonguenet: accurate localization and segmentation for tongue images using deep neural networks. IEEE Access, 2019, 7: 148779–148789.

[28] MA JJ, WEN GH, WANG CJ, et al. Complexity perception classification method for tongue constitution recognition. Artificial Intelligence in Medicine, 2019, 96: 123–133.

[29] ZHOU JH, ZHANG Q, ZHANG B, et al. TongueNet: a precise and fast tongue segmentation system using U-Net with a morphological processing layer. Applied Sciences, 2019, 9(15): 3128.

[30] ZHUANG QB, GAN SZ, ZHANG LY. Human-computer interaction based health diagnostics using ResNet34 for tongue image classification. Computer Methods and Programs in Biomedicine, 2022, 226: 107096.

[31] HUANG ZH, MIAO JQ, SONG HB, et al. A novel tongue segmentation method based on improved U-Net. Neurocomputing, 2022, 500: 73–89.

[32] JIANG T, GUO XJ, TU LP, et al. Application of computer tongue image analysis technology in the diagnosis of NAFLD. Computers in Biology and Medicine, 2021, 135: 104622.

[33] LI J, HUANG JB, JIANG T, et al. A multi-step approach for tongue image classification in patients with diabetes. Computers in Biology and Medicine, 2022, 149: 105935.

[34] DAI SX, GUO XJ, LIU S, et al. Application of intelligent tongue image analysis in conjunction with microbiomes in the diagnosis of MAFLD. Heliyon, 2024, 10(7): e29269.

[35] LU XZ, LIU S, LIN XX, et al. An AI-powered tongue image model for home-based monitoring of liver fibrosis. NPJ Digital Medicine, 2026, 9: 67.

[36] WANG RR, CHEN JL, DUAN SJ, et al. Noninvasive diagnostic technique for nonalcoholic fatty liver disease based on features of tongue images. Chinese Journal of Integrative Medicine, 2024, 30(3): 203–212.

[37] ZHANG JJ, FENG S, XUE J, et al. AI-driven multimodal fusion of tongue images and clinical indicators for identifying MAFLD patients at risk of coronary artery disease: an exploratory study. iLIVER, 2025, 4(3): 100181.

[38] PENG CD, WANG L, JIANG DM, et al. Establishing and validating a spotted tongue recognition and extraction model based on multiscale convolutional neural network. Digital Chinese Medicine, 2022, 5(1): 49–58.

[39] MA Q, ZHANG L, WU H, et al. Oral microbiota as a biomarker for predicting the risk of malignancy in indeterminate pulmonary nodules: a prospective multicenter study. International Journal of Surgery, 2025, 111(2): 2055–2071.

[40] BALASUBRAMANIYAN S, JEYAKUMAR V, NACHIMUTHU DS. Panoramic tongue imaging and deep convolutional machine learning model for diabetes diagnosis in humans. Scientific Reports, 2022, 12: 186.

[41] YUAN CH, LIU ZY, LI XY, et al. A dynamic weighted ensemble learning framework for cardiovascular risk prediction in type 2 diabetes: a comparative study with SHAP-based interpretability. Scientific Reports, 2025, 15: 45029.

[42] LIU PJ, BAGI K. A tailored deep learning approach for early detection of oral cancer using a 19-layer CNN on clinical lip and tongue images. Scientific Reports, 2025, 15: 23851.

[43] COŞGUN BAYBARS S, TALU MH, DANACı Ç, et al. Artificial intelligence in oral diagnosis: detecting coated tongue with convolutional neural networks. Diagnostics, 2025, 15(8): 1024.

[44] DAMKLIANG K, SUDKHAW T, YINGTAWEE T, et al. Leveraging transfer learning for Tri-Dhat classification of tongue images in traditional Thai medicine. ECTI Transactions on Computer and Information Technology (ECTI-CIT), 2025, 19(3): 442–457.

[45] YE R, JIANG ZK, SHAO R, et al. Development and validation of tongue imaging-based radiomics tool for the diagnosis of insomnia degree: a two-center study. Medical Data Mining, 2024, 7(1): 4.

[46] MATHEW JK, GOPALAKRISHNAN K, SARANYA G, et al. ExpACVO: exponential anti corona virus optimization enabled hybrid deep learning model for diabetic tongue image classification. Proceedings of the 2023 IEEE International Conference on Artificial Intelligence and Innovation in Healthcare Industries (ICAIIHI), 2023: 1–6.

[47] LIU XY, WANG SW, LV HY, et al. Lightweight CNN-SVM based hybrid model for tongue image classification. 2025 5th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS). IEEE, 2025. doi: 10.1109/acctcs66275.2025.00087.

[48] DEEPA SN, BANERJEE A. Intelligent decision support model using tongue image features for healthcare monitoring of diabetes diagnosis and classification. Network Modeling Analysis in Health Informatics and Bioinformatics, 2021, 10(1): 41.

[49] World Health Organization. WHO Traditional Medicine Strategy 2014 – 2023. Geneva: World Health Organization, 2013.

[50] LO LC, CHEN YF, CHEN WJ, et al. The study on the agreement between automatic tongue diagnosis system and traditional Chinese medicine practitioners. Evidence-Based Complementary and Alternative Medicine, 2012, 2012: 505063.

[51] LIU XX, FAES L, KALE AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health, 2019, 1(6): e271–e297.

[52] AGGARWAL R, SOUNDERAJAH V, MARTIN G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digital Medicine, 2021, 4(1): 65.

[53] ZHANG ZH, LI RR, CHEN Y, et al. Integration of traditional, complementary, and alternative medicine with modern biomedicine: the scientization, evidence, and challenges for integration of traditional Chinese medicine. Acupuncture and Herbal Medicine, 2024, 4(1): 68–78.

[54] MCGENITY C, CLARKE EL, JENNINGS C, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. NPJ Digital Medicine, 2024, 7(1): 114.

[55] Non-Communicable Disease Risk Factor Collaboration. Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. The Lancet, 2024, 404(10467): 2077–2093.

[56] CHAN JCN, LIM LL, WAREHAM NJ, et al. The Lancet Commission on diabetes: using data to transform diabetes care and patient lives. The Lancet, 2020, 396(10267): 2019–2082.

[57] FLOOD D, SEIGLIE JA, DUNN M, et al. The state of diabetes treatment coverage in 55 low-income and middle-income countries: a cross-sectional study of nationally representative, individual-level data in 680 102 adults. The Lancet Healthy Longevity, 2021, 2(6): e340–e351.

[58] American Diabetes Association Professional Practice Committee. Classification and diagnosis of diabetes: Standards of Care in Diabetes – 2024. Diabetes Care, 2024, 47(Suppl 1): S20–S42.

[59] HE JX, BAXTER SL, XU J, et al. The practical implementation of artificial intelligence technologies in medicine. Nature Medicine, 2019, 25(1): 30–36.

[60] MOOR M, BANERJEE O, ABAD ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature, 2023, 616(7956): 259–265.

[61] THIRUNAVUKARASU AJ, TING DSJ, ELANGOVAN K, et al. Large language models in medicine. Nature Medicine, 2023, 29(8): 1930–1940.

[62] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1(5): 206–215.

[63] GICHOYA JW, BANERJEE I, BHIMIREDDY AR, et al. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health, 2022, 4(6): e406–e414.

[64] VAIDYA A, CHEN RJ, WILLIAMSON DFK, et al. Demographic bias in misdiagnosis by computational pathology models. Nature Medicine, 2024, 30(4): 1174–1190.

[65] BIENEFELD N, BOSS JM, LÜTHY R, et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. NPJ Digital Medicine, 2023, 6(1): 94.

[66] YANG J, DUNG NT, THACH PN, et al. Generalizability assessment of AI models across hospitals in a low-middle and high income country. Nature Communications, 2024, 15(1): 8270.

[67] VAN DER VELDEN BHM, KUIJF HJ, GILHUIJS KGA, et al. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 2022, 79: 102470.

[68] LIU YY, FAN LM, ZHAO M, et al. Study on a traditional Chinese medicine constitution recognition model using tongue image characteristics and deep learning: a prospective dual-center investigation. Chinese Medicine, 2025, 20(1): 84.

[69] HAN R, ACOSTA JN, SHAKERI Z, et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. The Lancet Digital Health, 2024, 6(5): e367–e373.

[70] VOLLMER S, MATEEN BA, BOHNER G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. British Medical Journal, 2020, 368: l6927.

[71] JI H, ZHAO X, CHEN X, et al. Jinlida for diabetes prevention in impaired glucose tolerance and multiple metabolic abnormalities: the focus randomised clinical trial. JAMA Internal Medicine, 2024, 184(7): 727–735.

[72] SOUNDERAJAH V, ASHRAFIAN H, ROSE S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nature Medicine, 2021, 27(10): 1663–1665.

(Editor-in-Charge Siyi Wei)

人工智能辅助舌诊的可视化分析与临床可解释性评价

刘梨会[a, b, c], 胡凯文[a, b, c, d*], 周亚娜[b, c, e, f*]

a. 湖北中医药大学中医学院, 湖北 武汉 430061, 中国
b. 湖北省中医院肿瘤科, 湖北 武汉 430074, 中国
c. 湖北中医药大学附属医院中医肝肾研究及应用湖北省重点实验室, 湖北 武汉 430074, 中国
d. 北京中医药大学东方医院肿瘤科, 北京 100078, 中国
e. 湖北省中医药研究院, 湖北 武汉 430061, 中国
f. 湖北时珍实验室, 湖北 武汉 430060, 中国

【摘要】目的 通过文献计量学分析勾勒人工智能（AI）辅助舌诊的研究全景，并通过诊断准确性试验的 meta 分析定量评价其诊断准确性与临床可解释性。**方法** 文献计量学分析通过检索 Web of Science 核心合集（WoSCC）中 2014 年 1 月 1 日至 2025 年 12 月 31 日发表的 AI 辅助舌诊相关英文论文与综述，采用 Bibliometrix、VOSviewer 与 CiteSpace，从年度发文量与学科分布、期刊与引文特征、国家/地区与机构合作、作者网络、关键词共现以及关键词突现检测等多个维度进行综合分析。诊断准确性试验 meta 分析按照诊断准确性试验的系统评价/meta 分析报告规范（PRISMA-DTA），系统检索 Scopus、PubMed、Web of Science 与中国知网（CNKI）四个数据库。采用双变量随机效应模型的层次汇总受试者工作特征曲线（HSROC）合并敏感度与特异度，并按疾病类别、AI 模型架构及样本量分层进行亚组分析。方法学质量采用诊断试验准确性研究质量评价工具第 2 版（QUADAS-2）进行评估，发表偏倚通过 Deeks 漏斗图不对称性检验评估。**结果** 文献计量学分析共纳入 198 篇文献。2014 – 2025 年该领域年度发文量增长 24.5 倍（由 2014 年 2 篇增至 2025 年 49 篇），2022 – 2025 年发文量占全部文献的 65.2%。中国贡献了约 83.5% 的机构归属，其中上海中医药大学发文最多，许家佗为发文最多的作者。关键词分析识别出 AI 与深度学习架构、图像处理与分割、中医特异性应用、疾病特异性应用四个主题集群，并呈现出从传统机器学习向深度学习、transformer 架构、可解释性 AI 及多模态融合架构演进的时序特征。诊断准确性 meta 分析共纳入 16 项研究（14 755 名受试者），覆盖代谢与肝脏疾病、肿瘤与口腔病变、心血管风险、糖尿病等多个领域。合并敏感度为 90.3% [95% 置信区间（CI）：86.7% – 93.1%]，合并特异度为 93.0%（95% CI：90.6% – 94.7%），汇总受试者工作特征（SROC）曲线下面积（AUC）为 0.961；异质性显著（敏感度 I^2 = 95.8%；特异度 I^2 = 92.1%）。亚组分析显示，不同疾病类别、AI 架构与样本量分层之间性能总体一致，Deeks 检验未提示显著的发表偏倚（P = 0.258）。**结论** AI 辅助舌诊已快速发展，其合并诊断性能与既有筛查方式相当，提示其有望作为一种互补、便捷可及的辅助决策工具应用于临床。

【关键词】舌诊；人工智能；中医；诊断准确性；文献计量学；meta 分析