A Review of Research on Computer Vision-Based Analysis and Diagnosis of Tongue Patterns in Traditional Chinese Medicine

Qu Xiaoying
Harbin Institute of Technology (Weihai), Weihai, Shandong Province, 264299, China
13426804652@163.com

ABSTRACT

As the key method of TCM syndrome differentiation and treatment, objectification and intelligence of tongue diagnosis are the core of TCM modernization. This paper systematically reviews the technical development in the field of tongue diagnosis analysis and diagnosis of traditional Chinese medicine based on computer vision, and constructs a three-dimensional framework to understand the development of this field from the perspective of paradigm shift. The research combs the research status at home and abroad, and analyzes the progress in tongue image acquisition and preprocessing, feature extraction and analysis, and diagnosis model construction. Secondly, according to the technological generation, the evolution of research methods is divided into three stages: traditional image processing stage, deep learning stage, multimodal fusion stage and cognitive computing stage, and the core methods, technical principles and evolution logic of each stage are deeply analyzed. Based on this, we find out several main problems that the research is facing now: data standardization, whether the model can be understood by people, and clinical verification. Finally, it describes the future direction from three aspects: technical trend, clinical application expansion and standardization development. This research shows that this field is changing from the traditional mode to the cognitive computing mode driven by data and knowledge. Building high-quality multi-modal data sets and understandable dual-drive models is a specific aspect that needs to be paid close attention to in the future.

Keywords: Tongue Diagnosis in Traditional Chinese Medicine; Computer Vision; Deep Learning; Multimodal Fusion; Intelligent Diagnosis.

1. INTRODUCTION

1.1 Research Background

Tongue diagnosis is a core part of visual diagnosis in Traditional Chinese Medicine (TCM). It first appeared in the book Huangdi Neijing, and by the 13th century, there had been a book devoted to tongue diagnosis. It can provide doctors with a lot of diagnostic information, which is very important for judging the development and outcome of the disease. However, before the appearance of modern technical tools, the traditional tongue diagnosis faced some difficulties, such as being too subjective, lacking uniform standards and being difficult to quantify.

Since 1990s, engineering and computer technology have been used to digitize and analyze tongue images, which marks the beginning of objectification research of TCM tongue diagnosis. With the development of technology, the resolution of image acquisition equipment has been continuously improved, and the color reproduction has become more and more accurate. Now we can capture the details in tongue images more clearly.

China has established a relatively complete technical system in tongue image acquisition and processing. From the early portable tongue imaging equipment to the intelligent acquisition equipment in recent years [1]. Li Qingli's team studied the application of hyperspectral imaging system in tongue diagnosis of traditional Chinese medicine, and found that this technology can capture spectral information that traditional RGB images can't present [2]. The “TonguExpert” platform, jointly developed by Wang Sijia, Peng Qianqian and Chen Jianxin of Shanghai Institute of Nutrition and Health, China Academy of Sciences, is an automatic tongue feature analysis system based on deep learning, which can accurately identify and segment the tongue body in the tongue image and extract and analyze the fine tongue features [3].

Internationally, computerized tongue image analysis has made a new breakthrough in the field of disease diagnosis. A team from Tehran University of Medical Sciences in Iran studied the application of AI-driven tongue image analysis as a new non-invasive diagnostic tool in coronary heart disease based on Iranian traditional medical principles. This study analyzed the tongue characteristics of 906 patients who underwent coronary angiography, including tongue color, crack, tongue coating, humidity and erythema. They used Visioface equipment to shoot 12,000 high-definition tongue images to construct a data set, used Segment-Anything model to segment the tongue, and used DeepPlabv3+/Dinov2 architecture to extract features and classify stenosis. The results show that the accuracy of the AI model in predicting left circumflex artery stenosis is as high as 80%, and it is confirmed that tongue color and crack are important predictors of the disease. This study believes that AI-driven tongue diagnosis is a fast and economical auxiliary method, which is suitable for screening coronary artery diseases, especially in places with limited resources. This research method matches the technical framework of computer tongue image analysis, and forms a complete non-invasive screening process, from image acquisition, preprocessing, feature extraction to disease identification, which provides a reference for the internationalization and modernization of tongue diagnosis in traditional Chinese medicine. Many other international studies have also proved the clinical value of tongue image analysis in predicting various diseases [4].

1.2 Significance of the Study

1.2.1 Theoretical Significance

The theoretical importance of this research lies in that it systematically combs how computer vision technology changes step by step in the field of tongue diagnosis in traditional Chinese medicine. It overcomes the shortcomings of previous reviews, which often only focus on a certain technology or a certain usage scenario. Our research has constructed a multi-faceted framework to understand the development of this field from the perspective of macro “paradigm” transformation.

Specifically, this study introduces Kuhn’s paradigm theory and divides the technological evolution into three different paradigms-traditional image processing, deep learning, and multimodal fusion and cognitive computing-thus revealing the internal logic and driving force behind each stage. This framework not only explains the transformation process from manual design features to automatic learning, and then to the dual-drive model of knowledge and data, but also provides theoretical reference for future research on how to combine traditional Chinese medicine theory with artificial intelligence.

1.2.2 Practical Significance

The significance of this research is reflected in three aspects: first, it promotes the objectification and standardization of TCM diagnosis, and provides theoretical reference for standardization by describing the development process of technology; Secondly, it promotes the modernization and intelligent development of traditional Chinese medicine, and points out the direction for the follow-up research; Thirdly, it provides a new path for clinical auxiliary diagnosis, which is helpful to improve the scientificity and universality of TCM diagnosis and treatment.

To understand the significance of these contributions, we must first look at the current situation in this field. In the aspect of diagnostic ability, the intelligent evaluation of diseases and syndromes has become the research focus. Although the traditional image analysis method has been able to digitally represent the color of the tongue, it is still a difficult problem to comprehensively identify and analyze complex morphological features, such as tooth marks, potholes, cracks, thickness changes, decayed moss and peeling moss. In the construction of diagnostic model, domestic research has shifted from predicting a single disease to more complicated TCM syndrome classification. Researchers are using large-scale AI models to integrate multi-source information and promote the development of tongue diagnosis in the direction of intelligent evaluation of diseases and syndromes. For example, a special neural network, such as PMANet, which uses the attention mechanism, can directly identify syndromes on deep features, thus realizing the transformation from helping to diagnose diseases to simulating TCM syndrome differentiation. In clinical application, a large-scale tongue diagnosis model of traditional Chinese medicine can quickly help doctors analyze tongue images, which can improve diagnosis efficiency and reduce misdiagnosis. In telemedicine, patients can take photos of their tongues with smart devices and send them to medical centers for preliminary analysis. The future trend is to build large-scale models, which will combine TCM knowledge, multimodal information and portable devices, so as to make their application in clinical and health management deeper. However, the current research still faces some challenges, such as insufficient data standardization, lack of multi-center data sets, and the model is not easy to understand.

1.3 Research Objectives and Content

Through the comprehensive and in-depth analysis and arrangement of relevant academic literature at home and abroad in the past two decades, this study systematically traces the whole process of this field from technology start, rapid development to frontier exploration now. At the technical application level, the application of large-scale AI model in tongue diagnosis has become more and more mature. These large models can integrate multi-source data-including tongue diagnosis images, patients’ symptoms and medical history-for comprehensive analysis and judgment, and constantly optimize their diagnosis algorithms, so as to realize automatic recognition of tongue images and intelligent diagnosis of diseases.

This paper draws lessons from the scientific paradigm theory put forward by Thomas S. Kuhn, a historian of science. Kuhn to understand the methodological evolution in the field of tongue diagnosis analysis within Traditional Chinese Medicine. Kuhn argued that scientific development is not a linear accumulation, but rather achieves revolutionary breakthroughs through “paradigm shifts”-that is, when the overall framework of theoretical foundations, research methods, and value standards collectively accepted by the scientific community during a given period is replaced by an entirely new paradigm upon encountering anomalous phenomena that cannot be explained. In this study, we define a “paradigm” as the problem-solving approaches, technical pathways and core assumptions collectively followed by researchers within a specific technological phase, encompassing dimensions such as data acquisition methods, feature extraction strategies, model-building logic and diagnostic objectives. Based on this, this paper divides the technical development of tongue diagnosis in traditional Chinese medicine into three different stages: traditional image processing stage, deep learning stage, multimodal fusion and cognitive computing stage, with the aim of revealing the internal logic of each stage and the driving force behind these changes.

The research will not simply list a single technology or application achievement, but clarify the inherent logic and inevitability of the development route: from the beginning to objectification by traditional image processing and machine learning, to the intelligent breakthrough of automatic representation by deep learning, and finally to the stage of multimodal information fusion and cognitive computing that simulates the overall thinking of traditional Chinese medicine.

In order to achieve this goal, our research will be carried out from three closely related aspects: first, make a historical longitudinal analysis, that is, classify important documents according to different generations of technology, and make clear the research paradigm, landmark works and the connecting link between them at each stage. Second, do a horizontal deconstruction of methodology, and deeply analyze the specific methods, principles and their evolution adopted in the core technology path of each stage, such as data acquisition, feature engineering and model construction. Finally, it is a theoretical comprehensive evaluation, discussing how TCM concept of syndrome differentiation and treatment guides the development direction of technology, and providing a forward-looking outlook for future development on the basis of summarizing the existing research bottlenecks.

2. EVOLUTION OF RESEARCH METHODS

2.1 Traditional Image Processing Phase

The traditional image processing stage is from the beginning of the 21st century to around 2015, which is the basic period for studying the objectification of tongue diagnosis in traditional Chinese medicine. The core goal of this stage is to realize the quantitative expression of tongue picture characteristics, so as to provide objective data support for TCM diagnosis. Data preparation is the basis of intelligent analysis of four diagnoses, and this process usually includes data acquisition, preprocessing and labeling. Based on the theory of traditional Chinese medicine and their understanding of tongue images, researchers manually designed image features that can represent diagnostic elements, such as tongue color, fur color and tongue shape. Then they use the classic machine learning algorithm to establish a mapping relationship between these characteristics and diseases or syndromes.

This mode is an inevitable choice under the technical conditions at that time. On the one hand, deep learning technology is still immature, and image analysis depends on the traditional computer vision method; On the other hand, the complexity of TCM theory requires that feature design must have clear diagnostic significance, and manually designed features are very suitable for integrating TCM knowledge into technical processes. Although there are some shortcomings, such as strong subjectivity of feature design and limited generalisation ability, the work at this stage has laid a solid methodological foundation for the follow-up research and accumulated a large number of dataset.

In the traditional image processing stage, the research mainly focuses on the core goal of objectification. On the data acquisition level, the researchers established a preliminary standardized acquisition process by developing a special tongue diagnostic instrument and using a standard color card for color calibration. Color calibration technology mainly adopts Munsell color card or special tongue diagnosis color card, and uses polynomial regression or look-up table method to eliminate equipment differences and illumination effects. Because the quality of tongue image data is directly related to the reliability of subsequent analysis, the acquisition and preprocessing stages have become the focus of early research. The standardized equipment usually adopts D65 standard light source-a simulated sunlight light source with a color temperature of 6500K-which ensures the stability of tongue image color reproduction and provides a reliable basis for subsequent color analysis.

We often use some traditional image processing methods in the key steps of feature extraction and tongue segmentation. Color space analysis is a basic technology; Researchers will convert the tongue picture from RGB color space to a more uniform color space like HSV or Lab, so that the tongue color and tongue coating color can be measured more accurately.

As for tongue segmentation, we will use edge detection (such as Canny operator) to find out the edge of the tongue. Active contour model (such as Snake algorithm) accurately outlines the outline of tongue by minimizing energy; The JSEG region segmentation method will divide the picture into different regions according to the color and texture characteristics [5]. Edge detection technology is used to initially locate the boundary of the tongue, with the Canny operator being a widely used tool due to its excellent signal-to-noise ratio and localisation accuracy. However, simple edge detection struggles to handle complex situations where the tongue's colour is similar to that of the surrounding tissue and the boundaries are blurred. Consequently, active contour models (such as the Snake algorithm) have been introduced into the task of tongue body segmentation. Through an energy minimisation mechanism, this algorithm utilises the combined effects of image forces and internal forces to gradually evolve the initial contour to conform to the true boundaries of the tongue body, thereby achieving a more precise segmentation result. The Snake algorithm is not only applied in tongue diagnosis but is also used in facial diagnosis for the extraction of classification features such as complexion, lustre and facial shape. On the other hand, JSEG region segmentation method will divide the image into uniform regions according to the color and texture characteristics, which provides another technical means for the separation of tongue body and tongue coating. This method has also been used in the segmentation of tongue contour and the identification of tongue coating and tongue color [6].

When studying the shape and color characteristics of tongue, researchers used many different algorithms to help them make preliminary measurements. ZHU et al. used Douglas-Peucker method to calculate the tooth marks on the tongue, and the accuracy rate reached 80.00%. WANG et al. made a new study on a small feature on the tongue-prick. They compared the tongue photos taken under standard white light and pure green light, and found that the pricked area would show special reflection characteristics under green light irradiation. According to this discovery, the researchers successfully extracted the prick feature from the green tongue photo, and the accuracy rate reached 88.47%. Regarding crack extraction, YANG et al. proposed a method based on kernel pseudo-color transformation, and the correct detection rate was 82.00%. This method uses a nonlinear transformation to map the tongue image to a new feature space, thus enhancing the contrast between the crack area and the background, and then combining with morphological operation, the crack can be automatically detected [7].

The thickness, rotting degree, greasy feeling and wetness of tongue coating are the important basis for "pattern differentiation" in traditional Chinese medicine, so it has become a hot spot to study how to measure the texture of tongue coating. In the study of tongue coating texture, LIU and his colleagues used an upgraded "wavelet transform" and "grey-scale co-occurrence matrix" technology to extract texture features from tongue pictures. They also used the "Relief feature selection method" to obtain the "feature vectors" of tongue images, and the final classification accuracy reached 84.50% [8]. Rotten and greasy coating is characterized by rough texture and graininess, which looks like accumulated bean curd residue, and its texture characteristics are very different from normal tongue coating. QU et al. specially studied this kind of rotten moss. They used Gabor wavelet transform to extract roughness features, gray level co-occurrence matrix features and Gabor wavelet transform features, and the accuracy rate of identifying rotten moss exceeded 85.00%. CAO et al. used the optimal linear fusion method and AdaBoost algorithm to study thick moss and thin moss; AdaBoost fusion method based on K-nearest neighbor classifier is more effective in identifying thick moss, and the recognition rate reaches 90.34%.

In diagnostic modelling, classical machine learning algorithms such as logistic regression, support vector machines, decision trees and random forests are widely applied at this stage. A common characteristic of these algorithms is that input features require manual design and extraction; the role of the model is to find the optimal decision boundary in the feature space and perform predictive analysis by establishing a mapping relationship between tongue appearance features and specific diseases [3]. In traditional methods, features are typically designed manually for each category based on knowledge and experience; such features often exhibit relatively poor generalisation capabilities. To compare the strengths and weaknesses of different machine learning algorithms, Shi et al. compared six machine learning algorithms-including decision trees, support vector machines, random forests, neural networks, Naive Bayes and logistic regression-for lung disease prediction models based on tongue diagnosis images. They found that adding objective tongue appearance data to the baseline data could significantly improve the performance of lung disease prediction models [9].

Due to its advantages in handling small sample sizes and high-dimensional data, the Support Vector Machine (SVM) has become one of the most widely used classifiers. Researchers combine extracted features such as colour, texture and geometry into feature vectors, which are then input into the SVM model for the classification and prediction of diseases or syndromes. The logistic regression model, owing to its good interpretability, is used to analyse the contribution weights of each feature to the diagnostic results, thereby revealing, to a certain extent, the strength of the association between tongue features and diseases. Tree-based algorithms, such as decision trees and random forests, are capable of handling non-linear relationships and are insensitive to feature scales, demonstrating good performance in scenarios involving the fusion of multiple features.

Although this research paradigm achieved significant progress given the technological constraints of the time, it also revealed certain inherent limitations. Firstly, the quality of feature design is highly dependent on the researcher's knowledge of Traditional Chinese Medicine and their understanding of image features, making it difficult to standardise and compare feature sets designed by different researchers. Secondly, manually designed features often capture only superficial information from the tongue appearance and have limited capacity to express complex diagnostic relationships. In recent years, feature acquisition in the intelligent application of the Four Examinations has gradually begun to employ machine learning for further feature selection or deep learning for automatic feature extraction. Furthermore, as the various stages are relatively independent (segmentation, feature extraction and classification are carried out sequentially), errors in preceding steps are propagated and accumulated, affecting the final diagnostic performance. These limitations laid the groundwork for the rise of deep learning methods in the subsequent phase.

2.2 The Deep Learning Phase

Around 2015, with the breakthrough progress of deep learning technology in the field of computer vision, research into the analysis of the tongue in traditional Chinese medicine underwent a fundamental transformation, entering a phase characterised by automation and intelligence. The core of the technical approach in this phase is the construction of end-to-end deep learning models. Deep neural networks are capable of automatically learning multi-level feature representations directly from raw tongue image data, without the need for manually designed intermediate features. This paradigm shift has brought about three major technical characteristics: firstly, the automation of feature learning, whereby the model can autonomously discover the most effective combinations of features for diagnostic tasks; Secondly, end-to-end holistic optimisation, allowing tasks such as segmentation, feature extraction and classification to be learned jointly within a unified framework; Thirdly, a leap in representational capability, as deep networks can capture hierarchical information ranging from edge textures to semantic concepts, greatly enhancing the model's expressive power. After extracting effective features, the realisation of intelligent Four Diagnoses requires the construction of models based on these features to enable the classification and differentiation of TCM constitutions and syndromes.

When processing data, steps such as color correction are still needed, but the task of tongue segmentation can now be completed automatically with depth segmentation network, and it is more accurate. The Deep Tongue model developed by LIN is an end-to-end segmentation framework based on ResNet. Its most powerful feature is that it uses residual connection to solve the problem of gradient disappearance in deep network, so that the network can learn more rich hierarchical characteristics. The experimental results show that the MIoU of Deep Tongue model exceeds 94.00%, which is about 10% higher than that of the traditional method, and the segmentation speed is nearly ten times faster. ZHOU and his colleagues further explored the application of semi-supervised learning in tongue segmentation, and proposed a Semi Tongue network model. This model is trained with a small number of labeled data and a large number of unlabeled data. Through a consistent regularization mechanism, it keeps the output of the model stable under various disturbances, so that useful features can be learned from unlabeled data. Compared with fully supervised network models such as U-Net, Deep Tongue and Tongue-Net, when there are few labeled data, the segmentation accuracy of Semi Tongue is higher and more stable, which provides a feasible solution to the problem of insufficient labeled samples in clinical application. The successful application of these depth segmentation networks not only improves the segmentation accuracy, but more importantly, it changes the tongue segmentation from a single preprocessing step into a system component that can be optimized end-to-end, which lays the foundation for subsequent joint learning.

The biggest breakthrough in research methods lies in feature engineering. Instead of designing features manually, researchers use deep convolution neural network to automatically learn and extract multi-level and high-dimensional abstract features directly from the original tongue images. The application of convolutional neural network enables the model to automatically learn layer-by-layer features from edge, texture to semantic information. Very fine features can now be extracted; For example, TonguExpert platform can extract as many as 1,353 phenotypic features from a picture, including color, shape, texture and neural network features.

In the diagnosis model, the research focus has shifted from predicting a single disease to more complex TCM syndrome classification. A special neural network structure has emerged, such as PMANet with attention mechanism, which can directly identify syndromes in deep features, thus realizing the transformation from “assisting disease diagnosis” to “simulating TCM syndrome differentiation”. The hierarchical structure of Convective Neural Networks (CNNS) naturally meets the information processing requirements from local to whole and from concrete to abstract. Shallow layers usually learn the underlying visual features such as edges, corners and color blocks; The middle layer combines these low-level features into middle-level semantic information such as texture and pattern; In the deep layer, semantic concepts highly related to diagnostic tasks will be further abstracted, such as “tooth-mark regions” and “crack morphology”. This automatically constructed hierarchical feature structure can often capture subtle diagnostic clues that are difficult to find through manual design.

The addition of attention mechanism makes feature extraction more powerful. By using small modules such as spatial or channel attention, the model can decide to assign different “importance scores” to different parts of the picture and different color channels, so that it can pay special attention to those areas in the tongue image that are helpful for seeing a doctor. For example, the crack on the tongue usually accounts for only a small part of the picture; The traditional convolution operation may mix these small cracks in the information of the whole picture and become inconspicuous. However, with the attention mechanism, the model can especially enhance the signal in the crack area and reduce the interference of unimportant background, so that features can be captured more accurately. At present, the tongue symptom classification method combined with multi-attention mechanism has become a research hotspot; By combining multi-scale features with attention mechanism, the accuracy of identifying symptoms is greatly improved.

We also introduce a multi-task learning framework, which allows the model to learn multiple related tasks at the same time-such as forecasting tongue colour, coating colour and Tongue Shape-and use the correlation between tasks to improve the overall performance. TonguExpert platform performs well in predicting four kinds of Tongue Manifestations (Tongue Colour, Coating Colour, Tongue Fissures and tooth marks), and the AUC value is between 0.91 and 0.99, which shows the great potential of deep learning in intelligent diagnosis of tongue patterns.

2.3 Intelligent Diagnostic Models: From Disease Prediction to Syndrome Differentiation and Treatment

At the level of diagnosis model, the research focus has shifted from predicting a single disease to more complicated TCM syndrome classification, which marks a deep shift from “assisting disease diagnosis” to “simulating TCM syndrome differentiation”. This change reflects a deeper understanding of the essence of TCM theory: the core of TCM diagnosis is not only to identify specific diseases, but to identify the current “syndrome”-that is, the overall response of the human body to pathogenic factors through four diagnostic methods.

PMANet is a representative work in this field. This network innovatively combines multi-scale feature extraction with attention mechanism: the multi-scale module can capture the overall structure and local details of the tongue appearance at the same time, and the attention mechanism allows the model to focus on the tongue area related to a specific symphony. By doing end-to-end synthesis recognition in the depth feature space, PMANet skips the clear definition of intermediate features and directly learns the mapping relationship from tongue appearance to synthesis, which is much better than those shallow models based on manual features.

The introduction of multi-task learning framework further improves the effect of diagnosis model. Various tasks in TCM diagnosis, such as tongue colour classification, coating colour classification and Tongue Shape Recognition, are not independent of each other, but are internally related. By sharing the underlying feature representation, multi-task learning can optimize multiple related tasks at the same time; On the one hand, this makes use of the correlation between tasks to improve the performance of each task, on the other hand, it can also help the model learn more general feature representation, thus reducing the risk of over-fitting. Under the framework of multi-task learning, TonguExpert platform performs well in predicting four categories of tongue images-Tongue Colour, tongue coating colour, tongue fissures and Tongue Indentations, with AUC values ranging from 0.91 to 0.99 [5], which fully demonstrates the great potential of deep learning in intelligent diagnosis of tongue images.

3. KEY TECHNICAL ISSUES AND CHALLENGES

Although great progress has been made in the field of deep learning, there are still key challenges at the data and model levels.

3.1 Challenges at the Data Level

The breakthrough of deep learning stage is revolutionary: automatic feature learning gets rid of the limitation of manual design, end-to-end optimization improves the overall performance, and refined representation opens the door to in-depth exploration. But this stage also exposed new problems. The first is data dependence: deep learning models usually need a lot of labeled data to perform well, but high-quality tongue image data sets of traditional Chinese medicine are still scarce. Secondly, the interpretable challenge: although deep learning models perform well, their decision-making process is like a “black box”, which is difficult to meet the clinical requirements of traditional Chinese medicine for clear diagnosis basis. In addition, the integration of TCM knowledge is insufficient: pure end-to-end learning can fit the data, but it does not necessarily conform to the internal logic of TCM theory. These problems point out the direction for the next stage of development.

Compared to the rapid advancements in data collection and model construction, clinical validation research for tongue diagnosis AI models lags significantly, becoming the core bottleneck hindering their transition from the laboratory to clinical practice. Among these, data standardisation is the primary challenge currently facing research into the intelligent diagnosis of the tongue. Regarding collection devices, differences in spectral response characteristics and colour reproduction capabilities between devices lead to the same tongue appearance exhibiting different features on different devices. In terms of environmental control, factors such as light intensity, colour temperature and angle all influence the colour and texture representation of tongue images. Regarding sample size and quality, there is currently a lack of large-scale, multi-centre, high-quality public tongue image datasets, which limits the models’ generalisation ability and comparability.

Firstly, single-centre, small-sample studies predominate, with a lack of multi-centre, large-sample validation. Most existing AI research on tongue diagnosis is conducted using limited samples from a single medical institution; whilst model performance often excels on specific datasets, it declines significantly when transferred to other clinical settings, different populations, or different imaging devices [10]. A study on gastric cancer diagnosis via tongue examination, conducted by teams from Zhejiang Cancer Hospital and West Lake University, adopted a prospective, multicentre cohort design. It recruited 937 gastric cancer patients and 1,911 non-gastric cancer participants from 10 centres, with independent external validation carried out at a further seven centres. The findings were published in *eClinicalMedicine*, a subsidiary journal of *The Lancet*, setting a benchmark for the clinical validation of tongue-based AI .Such high-quality, multicentre studies remain few and far between [11].

Secondly, there is a lack of rigorous randomised controlled trials. As an auxiliary diagnostic tool, the actual value of tongue diagnosis AI systems for clinical outcomes urgently needs to be validated through randomised controlled trials. However, there are currently few RCT studies comparing tongue diagnosis AI systems with independent diagnoses by clinicians or human-machine collaborative diagnoses. Without such evidence-based medical evidence, the clinical value of tongue diagnosis AI is difficult to be recognized by mainstream medical circles.

Great progress has been made in standardization at home and abroad, but there is still a clear gap between the formulation of standards and actual clinical application. On the one hand, most research teams do not strictly abide by the existing standards when collecting data, which makes it difficult to integrate and compare the data of different studies; On the other hand, most of the existing standards are general guidelines, lacking detailed specifications for key clinical links such as tongue picture feature labeling and syndrome diagnosis mapping. A comparative analysis of six domestic and foreign terminology standards for tongue diagnosis shows that these standards are different in terms of structural framework, number of items, terminology translation and definition. Therefore, objectification and quantification of tongue diagnosis data still have room for further development.

The construction of multimodal datasets faces the challenge of aligning heterogeneous data. Data such as tongue images, facial examination images, pulse waveforms and consultation transcripts exhibit significant differences in collection methods, temporal scales and data structures; achieving effective alignment is a prerequisite for fusion analysis. Regarding the standardisation of annotation, different research teams apply varying standards for annotating tongue-image features, and the lack of unified annotation guidelines makes it difficult to integrate and utilise datasets. In terms of adapting to individual differences, physiological variations in tongue images exist across different age groups, genders and ethnicities; Models must possess the ability to adapt to these individual differences to achieve universal application. Cross-cultural applicability is also an issue that requires consideration.

In summary, constructing high-quality, multi-centre, large-scale tongue-image datasets; advancing clinical validation from single-centre exploratory studies to multi-centre randomised controlled trials; and achieving an organic integration between standardisation and clinical application represent the key areas requiring urgent attention in current AI research on tongue diagnosis. This necessitates the collaborative efforts of research institutions, healthcare organisations, standardisation bodies and regulatory authorities.

3.2 Challenges at the Model Level

3.2.1 Challenges in Multimodal Fusion

The technical core of multimodal data fusion lies in effectively integrating data from diverse sources and varying structures, such as tongue images, facial examination, pulse diagnosis, and even textual data from olfactory and interrogative examinations.

At the level of representation learning, we have to distinguish between two different representation strategies: one is called separate representation, which means that each information is represented in its own independent way, that is, each information is mapped to a different feature space; The other is called unified representation, which maps the features in different information into the same feature space to extract features, which has the advantage of more convenient integration of various information [1]. The current research tends to use the method of unified representation to map data from different sources into the same semantic space, which can help us learn their characteristics at the same time and make different information complement each other. For image modes (tongue and facial appearance), convolutional neural network is usually used to extract visual features; For waveform mode (pulse mode), one-dimensional convolution or cyclic neural network can be used to model time characteristics; For the medical consultation records, the semantic representation is obtained by pre-training the language model. On this basis, a cross-modal attention mechanism or shared representation space is designed, so that the features of different modes can be compared and complemented in the same space. In this way, the limitation of single modal information can be overcome, and the model can comprehensively use diagnostic clues from many aspects.

Cross-modal alignment and complementary mechanism enable the model to discover the internal relations between different modes. For example, some changes in the tongue may correspond to specific pulse characteristics; Multi-modal model can automatically learn these associations and supplement them with another model when the information of one model is insufficient. The TCM multi-modal integrated health management system is a typical application in this field. It integrates many kinds of data such as tongue diagnosis images, face diagnosis images and pulse diagnosis waveforms, and analyzes them with deep learning and computer vision technology, which greatly improves the accuracy and objectivity of TCM diagnosis [6].

3.2.2 Interpretability and the Black-Box Problem

There is a contradiction between the “black box” characteristics of deep learning model and clinical acceptance. Although deep learning models perform well in classification tasks, their decision-making process is not transparent enough, and it is difficult to meet the clinical requirements of traditional Chinese medicine for clear diagnosis basis. Because the visualization and interpretation tools are not perfect, the key to improve the interpretability of the model is to establish the corresponding relationship between the model concerned areas and the theoretical diagnostic criteria of traditional Chinese medicine.

In the specific scene of tongue diagnosis, the black-box feature of deep learning model brings three specific and unavoidable challenges:

Firstly, the uncertainty of feature localisation. When a deep learning model classifies a tongue image as “Qi deficiency syndrome”, we cannot ascertain whether the model based its judgement on the pale-white tongue body, the plump and tender tongue, or the thin white coating; or whether it relied on changes in light and shadow within the image background, artefacts along the tongue’s edges, or even random noise signals present in the training data. This “feature confusion” is particularly dangerous in tongue diagnosis—the pathological features of a tongue image are often subtle; cracks may occupy only a fraction of a thousandth of the image area, and tooth marks may be merely slight undulations along the edges, whilst the model may well be capturing statistical coincidences unrelated to TCM theory.

Secondly, the lack of traceability in decision-making logic. The value of clinical diagnosis in Traditional Chinese Medicine lies in the consistency of its “principles, methods, formulas and medicines”: from observation, auscultation, inquiry and palpation to pattern differentiation and formulation of treatment principles, and finally to the prescription of formulas and medicines, every step is grounded in reason. However, deep learning models merely output an isolated result label, unable to demonstrate the reasoning process behind “why this syndrome rather than that” or “what are the key points for distinguishing between these syndromes”.

Thirdly, the unpredictability of error patterns. As we cannot comprehend the model’s internal workings, it is difficult to predict under what circumstances it will fail. Does it misjudge when the lighting is warm-toned? Does it misjudge the tongue appearance in elderly patients? Or does it misjudge when a certain rare type of tongue coating appears? This ambiguity regarding the “failure boundaries” makes it difficult to assess and control the risks associated with the model’s clinical application.

If these “black-box” characteristics are not addressed, they will trigger a series of chain reactions in the process of intelligentising TCM tongue diagnosis: TCM emphasises “differentiating syndromes to seek the root cause”; if the basis for diagnosis is unknown, the rationality of treatment plans cannot be discussed. This dilemma of “knowing the what but not the why” directly constrains the possibility of intelligent diagnostic systems moving from the laboratory to clinical practice.

At the quality control level, models lacking interpretability make it difficult to conduct effective error analysis and iterative optimisation. When a model’s performance declines on a particular batch of data, we cannot determine whether this is due to a shift in the data distribution or whether the model has learned a spurious correlation; when the model exhibits systematic bias (such as low recognition rates for tongue patterns in people of a certain skin tone), it is difficult to trace the root cause of the bias and make targeted corrections.

Currently, research efforts to address these issues are underway, but there is a fundamental gap.

Firstly, there is a lack of “semantic alignment” between interpretability methods and Traditional Chinese Medicine (TCM) diagnostics. Current mainstream visualisation tools (such as activation mapping) can display the image regions of interest to the model, but these regions correspond merely to “pixel-level hotspots”, rather than TCM diagnostic features such as “tooth marks”, “cracks” or “punctures”. There is currently no mature solution for automatically mapping high-response regions on heatmaps to diagnostic features with TCM semantic meaning. Even if the model does indeed focus on the tongue margin, we still need to manually determine whether this corresponds to “tooth marks”; this interpretation itself introduces new uncertainty.

Secondly, attempts at knowledge fusion suffer from a “mechanism disembedding”. Integrating TCM knowledge graphs or rule bases with deep learning models is currently one of the mainstream approaches to enhancing interpretability. However, most of these methods are still at the level of “post-processing”-first, the black box model is used to make predictions, and then the knowledge base is used to constrain or verify the results. This knowledge fusion method has not really changed the internal decision-making mechanism of the model, nor can it make the model think like TCM practitioner. The deeper problem is that the statistical correlation mechanism of deep learning model is fundamentally different from the dialectical logic mechanism of TCM. How to combine them organically instead of simply stacking them is still an unsolved theoretical problem.

Simply put, the interpretability of intelligent tongue diagnosis is essentially the gap between data-driven statistical learning mode and cognitive mode driven by TCM theory. To fill this gap, we need not only technological innovation-such as developing an interpretable deep learning framework and constructing a visualization method that conforms to the semantics of traditional Chinese medicine-but also theoretical breakthrough. We need to deeply study the cognitive mechanism of deep learning model, see what are the similarities and differences between it and dialectical thinking of traditional Chinese medicine, and find ways to make them talk on the epistemological level. Only in this way can we really create an interpretable intelligent diagnosis system that is in line with the theoretical logic of traditional Chinese medicine and has clinical practical value.

These challenges are not only technical problems, but also involve understanding the nature of TCM diagnosis. Multimodal fusion should not only focus on the technical integration of data, but also explore its relationship with TCM cognitive mechanism. Explainability ability is not only the technical requirement to open the “black box”, but also the basic requirement for the rationality of the diagnosis process. To overcome these challenges, we must rely on the deep intersection and joint efforts of computer science, fundamental TCM theory, cognitive science and other disciplines.

4. SUMMARY AND OUTLOOK

4.1 Technological Trends

This study systematically tracks how the technology of analyzing and diagnosing tongue images with computer vision developed step by step in Traditional Chinese Medicine (TCM), and also establishes a three-dimensional framework to help us understand the progress in this field. The research shows that this field started with the traditional image processing method based on manual design and prior knowledge, and later developed into the deep learning method characterized by data-driven and end-to-end learning. Now, it is moving towards a multi-modal fusion and cognitive computing that can imitate the overall thinking of TCM.

The future technological development will advance along the direction of using data and knowledge together. Adding the empirical rules of traditional Chinese medicine into the model systematically will become an important method to make the model more consistent with the theory of traditional Chinese medicine. Dynamic learning and adaptive mechanism will enable the model to continuously optimize itself according to new data and adapt to everyone's different situations and changing clinical needs. Cross-modal attention mechanism will further improve the effect of multi-modal fusion, so that the model can intelligently focus on the most diagnostic models and regions. Constructing a pre-training model specifically for the field of Chinese medicine will become a research hotspot. By pre-training on a large-scale multimodal data set of Chinese medicine, a special basic model for this field can be obtained.

In clinical application, intelligent diagnosis assistant system will become more and more important. Clinical decision support system can quickly help doctors analyze tongue diagnosis, improve diagnosis efficiency and reduce misdiagnosis and missed diagnosis. Telemedicine application will break geographical restrictions and make high-quality TCM medical resources benefit more patients. During rehabilitation treatment, regular monitoring of tongue picture changes can evaluate the treatment effect and adjust the treatment plan in time.

In the field of health management, the application or platform of public health education based on large model allows the public to upload their own photos of tongue images and obtain preliminary analysis of health status and health advice. This will help to improve the public's understanding of tongue diagnosis of traditional Chinese medicine, urge everyone to pay more attention to personal health, spread the knowledge of TCM health preservation and promote the promotion of TCM culture.

In order to cope with the interpretive problem of deep learning model and be closer to the dialectical method of traditional Chinese medicine, the latest research is trying to combine the symbolic AI technology such as knowledge map and causal reasoning. These efforts are to integrate the knowledge in ancient Chinese medicine classics and doctors" experience rules into the data-driven model. By constructing a standardized TCM knowledge graph, which contains rich theoretical knowledge, clinical experience and prescriptions, combined with TCM symptom extraction and recommendation methods, the reliability and interpretability of the diagnosis and treatment model can be improved. The development of dual-drive diagnosis system makes the diagnosis and treatment process of traditional Chinese medicine more intelligent and scientific. By embedding TCM theory into data-driven models, these models can not only learn the statistical laws in data, but also abide by the basic principles of TCM theory, thus achieving deeper integration.

It has become a research direction that everyone is exploring to use those large-scale language models that have been trained to do multimodal tasks in traditional Chinese medicine, such as adjusting and optimizing them. These large models are amazing. They can integrate information from different sources, such as pictures of tongue diagnosis, patients" discomfort and past medical history, and make a comprehensive analysis and judgment. By learning many cases of tongue diagnosis, they can constantly improve their own diagnostic methods, so that they can automatically identify the appearance of the tongue and intelligently help us judge what disease we have.

4.2 Standardisation and Regulation

Standardization is the basis of promoting the sustainable development of intelligent tongue diagnosis. Data acquisition standards must uniformly stipulate the technical requirements for acquisition equipment, environmental conditions and operating procedures. It is necessary to establish a unified algorithm evaluation index performance evaluation system to facilitate the comparison between different studies. System integration specifications must specify technical standards such as system architecture, interface standards and data formats to promote interoperability between systems.

In terms of ethics and regulations, it is necessary to establish a reliable data privacy protection mechanism, formulate clinical verification guidelines, and improve industry supervision policies to ensure the safe, reliable and compliant application of this technology.

After more than 20 years of development, the standardization of AI-assisted tongue diagnosis has made great progress, and the preliminary technical specification framework from data acquisition to algorithm evaluation has been established. On the terminology level, ISO 23961-1: 2021 "Traditional Chinese Medicine–Diagnostic Vocabularity-Part 1: Tongue

Diagnosis" led by Wang Yiqin's team is the first ISO international standard on the terminology of tongue diagnosis of traditional Chinese medicine, and it was selected as "Top Ten Academic Progress of Traditional Chinese Medicine in 2021" [12].

The standardization work is becoming more and more systematic and perfect from the initial stage. In the future, the focus will shift from hardware standards to data standards, especially to formulate labeling specifications about tongue features, which can help computers better understand the meaning of tongue. At the same time, we have to make special standards for some special medical scenes and different groups of people. This more and more comprehensive standard system will lay a solid foundation for the large-scale application of tongue diagnosis AI, the cooperation between different hospitals and the promotion to the whole world, so that the traditional Chinese medicine tongue diagnosis will gradually move from relying on experience to standardized clinical practice.

4.3 Research Limitations and Outlook

At present, the research still faces some challenges, such as insufficient data standardization, unclear model and lack of multi-center verification. Future research should focus on establishing high-quality multimodal data sets, developing more understandable AI and knowledge-guided models, and promoting strict clinical trial verification, so as to form a complete closed-loop system.

Looking back on the development of more than 20 years, we can clearly see a route of technological progress: from manual design to automatic learning, from single mode to multimodal integration, from disease prediction to syndrome differentiation and treatment. This evolution, on the one hand, is strongly promoted by the progress of general computer vision and AI technology, on the other hand, it is always guided by the inherent characteristics of TCM theory and the actual clinical needs. This chapter systematically combs the technical development path of research methods in this field from the traditional image processing stage to the deep learning stage, and then to the current multi-modal fusion and cognitive computing stage. With the in-depth development of intelligent interconnection and the era of big data, the large-scale tongue diagnosis model of traditional Chinese medicine will show broad application prospects in scientific research, medical practice and health education, and bring profound changes and positive effects to the modernization of traditional Chinese medicine, the improvement of medical service level and the progress of public health.

REFERENCES

[1] Liu, Y., & Lu, X. Z. (2005). Development of a portable tongue imaging device [in Chinese]. Journal of Tianjin University of Traditional Chinese Medicine, (3), 164-165.

[2] Li, Q. L., Xue, Y. Q., Wang, J. Y., et al. (2006). Research on the application of hyperspectral imaging systems in traditional Chinese medicine tongue diagnosis [in Chinese]. Journal of Infrared and Millimeter Waves, *25*(6), 465-468.

[3] Ting Li, Ling Zuo, Pengyu Wang, Liangfu Yang, Zijia Liu, Xu Wang. & Sijia Wang. (2025). Tongu Expert: A Deep Learning-Based Algorithm Platform for Fine-Grained Extraction and Classification of Tongue Phenotypes.Phenomics,5(2),1-14. https://doi.org/10.1007/S43657-024-00210-9.

[4] Alireza Hekmat Ardakani, Hamid R. Rabiee, Armin Behnamnia, Amir Hooman Kazemi, Amir Sobh Rakhshankhah, Ramtin Mahmoudi Kashani. & Mehrdad Karimi. (2025).AI-driven tongue image analysis for diagnosing and predicting coronary artery disease. Scientific Reports,15(1),44722-44722.https://doi.org/10.1038/S41598-025-28417-4.

[5] Qi, Z., Xu, J. T., Zhang, Z. F., et al. (2015). Advances in the clinical application of tongue diagnosis objectification based on digital image processing technology [in Chinese]. Chinese Journal of Traditional Chinese Medicine, *30*(8), 2849-2851.

[6] Zhao, W., Zhang, J., Xu, J. J., Xin, J. L., Zhou, C. E., Li, S. Z., & Li, C. D. (2020). Current status and implementation pathways for the intelligent development of integrated four diagnostic methods [in Chinese]. Journal of Traditional Chinese Medicine, *61*(1), 58–62, 67. https://doi.org/10.13288/j.11-2166/r.2020.01.013

[7] Xu, J. T., Jiang, T., & Liu, S. (2024). Current status and prospects of machine learning-based diagnostic analysis of tongue morphology and characteristics [in Chinese]. Digital Chinese Medicine, *7*(1), 3-12.

[8] Jiang, T., Tu, L. P., & Xu, J. T. (2024). Review and outlook on intelligent diagnostic technologies for tongue diagnosis in traditional Chinese medicine [in Chinese]. Chinese Journal of Traditional Chinese Medicine Information, *31*(7), 182-187.

[9] Su, Y. L., Hu, X. Y., Ma, S. J., Zhang, Y. N., Abulizi, A., & Abulikemu, H. (2024). A review of artificial intelligence research in the field of traditional Chinese medicine diagnosis and treatment [in Chinese]. Computer Engineering and Applications, *60*(16), 1-18.

[10] Zhang, M., Jin, X., Li, Y. M., et al. (2025). Clinical research on intelligent healthcare in traditional Chinese tongue diagnosis based on multimodal large models [in Chinese]. Pharmacology and Clinical Applications of Traditional Chinese Medicine, *41*(3), 72-76.

[11] Huang, X. (2023, March 15). "When the stomach is unwell, the tongue knows first!" This TCM study featured in a subsidiary journal of The Lancet [in Chinese]. Economic Information Daily, p. 005.

[12] Hou, C. L., Cui, Y. J., Xu, Y., Sang, Z., Wang, Y. Q., & Hao, Y. M. (2022). A study on international standards for the English translation of terminology in traditional Chinese medicine tongue diagnosis [in Chinese]. World Science and Technology—Modernization of Traditional Chinese Medicine, *24*(2), 815-821.