Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 5 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i05.53671

PAPER

Multi-Model Approach for Tongue Image Classification in Traditional Thai Medicine

Kasikrit Damkliang[1](✉), Julara Chumnaul[1], Teerawat Sudkhaw[2], Thitinan Yingtawee[1], Nasma Saearm[1]

[1]Division of Computational Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand

[2]Faculty of Traditional Thai Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand

kasikrit.d@psu.ac.th

ABSTRACT

Nowadays, complementary medicine is gaining widespread acceptance and is widely accepted, particularly within traditional Thai medicine (TTM). Tongue inspection is a primary method for diagnosing health conditions, as it reflects organ functionality. However, diagnostic results can vary depending on the expertise of TTM practitioners. In this work, we propose methods that incorporate transfer learning (TL) from deep learning (DL), machine learning (ML), and statistical models, using various tongue features. We introduced a collected dataset for evaluation. Experimental results demonstrated that the DenseNet121 model, trained on tongue images pre-processed with histogram equalisation (HE), achieved the best performance, with accuracy, sensitivity, and specificity of 0.89, 0.83, and 0.92, respectively. Model ensembling and paired t-tests were used to analyse the results. Finally, we identified the best approach and models for potential clinical use to assist in the pre-diagnostic analysis of tongue images for TTM practitioners and general users via our web application at http://bioservices.sci.psu.ac.th/.

KEYWORDS

feature analysis, multinomial logistic regression (LR), traditional Thai medicine (TTM), tongue image classification, transfer learning (TL)

1 INTRODUCTION

Traditional medicine has been used for illness treatment over a long historical period in many parts of the world. In particular, traditional Thai medicine (TTM) inherits a rich cultural heritage and indigenous wisdom, becoming increasingly popular and widely used in Thailand. TTM is based on a holistic medical approach and unique theoretical principles, incorporating knowledge from Ayurveda in Indian culture, traditional Chinese medicine (TCM), Buddhist teachings, Thai folk culture, spirituality, and astrology [1], [2], [3].

Deep learning (DL) methods have been widely used for tongue image classification, specifically in TCM, to reduce the errors associated with subjective judgement

Damkliang, K., Chumnaul, J., Sudkhaw, T., Yingtawee, T., Saearm, N. (2025). Multi-Model Approach for Tongue Image Classification in Traditional Thai Medicine. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(5), pp. 47–62. https://doi.org/10.3991/ijoe.v21i05.53671

Article submitted 2024-12-06. Revision uploaded 2025-02-26. Final acceptance 2025-02-26.

and to automate tongue diagnosis [4], [5]. Various DL techniques, including transfer learning (TL), have been applied to classify TCM syndromes, diseases, and physical constitutions based on tongue images. In particular, deep TL has been proposed and applied to assist in TCM tongue diagnosis, addressing challenges such as the scarcity of clinical diagnosis data and improving model interpretability [4], [6], [7].

Despite its growing usage, there is currently no AI-assisted information system available to support TTM practitioners in the pre-diagnosis of tongue images based on TTM knowledge. Additionally, general users interested in understanding their health can use such a system to analyse tongue images and receive health advice generated by the system.

In this work, we present an expanded analysis of tongue images aimed at improving classification outcomes by utilising a variety of features derived from the images. Our methods integrate TL from DL, ML, and statistical models. Through this approach, we have identified the most effective models for potential pilot clinical use, offering pre-diagnostic tongue image analysis tools for TTM practitioners and general users.

2 RELATED WORKS

In this section, we review recent studies on tongue image analysis, particularly those emphasising the role of AI, as summarised in Table 1.

Table 1. Studies on tongue image analysis, emphasising the role of AI in classification

Author	Data	Method	Target	Accuracy
Li et al. [8]	Tongue images obtained from a diagnosis instrument	ViT combined with Grad-CAM and K-means	Diabetic tongue	0.84
Zhou et al. [9]	Tongue images of 330	ResNet-34	Tooth marks recognition	0.92
Shi et al. [10]	Tongue characteristic data	RF, LR, SVM, and NN	Syndrome classification	0.88
Xu et al. [11]	1,858 tongue images	UNet and NN	Five classes of clinical significance	0.92
Ma et al. [12]	Tongue images	Pretrained models and LBP technique	Nine physique types	0.59

In TCM, Li et al. [8] proposed a method for tongue image classification in diabetes patients using a multi-step feature extraction approach that utilises original images and a tongue diagnosis acquisition system (TDAS). Features such as colour spaces, textures, and coating ratios were extracted by the TDAS. A vector quantised-variational autoencoder (VQ-VAE), combined with K-means, was used to analyse the differential interpretation of these features. Three DL models—ViT [13], DenseNet121 [14], and ResNet-50 [15]—were trained and validated. As a result, the ViT model demonstrated the best classification accuracy of 84.4% in a five-fold cross-validation evaluation.

Specifically in TCM, tongue image analysis has been applied to tasks such as colour classification, fur colour classification, shape classification, and crack classification. The authors utilised multi-feature fusion models and TL techniques to improve accuracy and reduce data requirements. However, subjective factors can still affect the judgement standards [16].

In TCM, tooth marks are often associated with spleen and blood disorders. Zhou et al. [9] proposed a method for recognising tooth marks in three classes consisting of Qi-deficiency, Yang-deficiency, and Yin-deficiency. A dataset was created using the

tongue region of interest (ROI) from 330 images, which were used to train ResNet-34 models [17] in an end-to-end manner. The authors claimed that the proposed method provided convenience for clinical diagnosis and interpretation, achieving an accuracy of 0.92, precision of 0.87, recall of 0.94, and F1 score of 0.90.

In cancer-related research, Shi et al. [10] proposed a method for establishing a syndrome classification model using tongue characteristic data and pulse information from non-small-cell lung cancer patients. The authors employed four machine learning (ML) classifiers, including random forest (RF), logistic regression (LR), support vector machine (SVM), and neural networks (NN), to classify symptoms. The results showed that the NN model achieved the best accuracy of 0.88, indicating the feasibility of using tongue and pulse data for non-small-cell lung cancer diagnosis.

Xu et al. [11] proposed an automatic multi-task joint learning approach that combined segmentation and classification tasks for a dataset of 1,858 tongue images, classified into five clinically significant categories in TCM. A customised UNet model and a discriminative filter learning model based on deep NN were utilised. The experimental results at the pixel level demonstrated that their method was highly consistent with human perception, achieving an accuracy of 0.92.

Ma et al. [12] proposed a system framework for identifying tongue images using a complexity perception approach to distinguish between easy and difficult training data. Tongue images of three different sizes from clinics were used in combination with models such as VGG-19 [18], Inception-V3 [19], and VGG-16 [18], as well as the local binary pattern (LBP) and colour moment techniques. Their results demonstrated that the system was applicable to real-world clinical settings in traditional Chinese medicine.

While previous studies have applied various AI techniques for tongue image classification, most have been conducted within the TCM framework. However, AI-assisted classification in TTM remains unexplored. Moreover, existing models often lack generalisability due to limited datasets or subjective feature selection. To address these challenges, we propose a multi-model approach that integrates DL, ML, and statistical methods for Tridhat classification.

To bridge this gap, our study introduces a novel dataset and systematically evaluates multiple analytical approaches. We analyse various features derived from our collected tongue images and apply feature extraction techniques, ML classifiers, and statistical models based on methodologies found in related works. The following section details our approach in depth.

3 MATERIALS

This section presents the data acquisition processes, data pre-processing steps for the training process, and analysis approaches in detail. All protocols adhered to the Declaration of Helsinki and received approval from the Ethics Committee of the Faculty of Traditional Thai Medicine at Prince of Songkla University, Thailand (Ethical Application Ref: EC.66/TTM.01-011). Written informed consent was obtained from all participants or their guardians. The approval is valid from October 17, 2023, to October 16, 2024.

3.1 Datasets

Data acquisition was conducted with volunteers aged between 18 and 60 years who visited the Traditional Thai Medicine Hospital at Prince of Songkla University, as shown in Table 2. A DSLR camera (Nikon D3400 with lens specifications of 18–55mm f/3.5–5.6G)

and a mobile phone camera (iPhone 13 with lens specifications of 1.5–5.1mm f/1.6–2.4) were used to capture tongue images under a calibrated environmental setting.

Table 2. Demographic and tongue characteristics of 284 patients from the TTM Hospital, Prince of Songkla University. The data are presented as n (%) patient prevalence, minimum, maximum, and mean age

Variable	Category	Study Population
Sex	Male	71 (24.8)
	Female	215 (75.2)
Age	Maximum	59
	Minimum	18
	Mean	21.85
Tridhat of tongue	Pitta	175 (61.6)
	Vata	22 (7.7)
	Kapha	87 (30.6)

After inspecting all 276 volunteers, we collected 284 cases, accounting for image file damage in two cases. The images were cropped to ensure a minimum width of 500 pixels, annotated for physical structures, and classified into Tridhat categories by three independent TTM practitioners, each with at least five years of experience, as shown in Figure 1A. Interobserver variability was assessed, yielding a kappa value of 0.30 based on Fleiss' Kappa calculation [20]. We used cropped tongue images taken by different devices to create their respective datasets.

Inspired by Vega-Huerta et al. (2024) [21], we used oversampling for minority classes and under sampling for the majority class to ensure balanced classes of the images. Based on an average class size of 94.67 of the number of 284 subjects, we selected a roundup of the average size into 100 subjects. The original imbalanced dataset was augmented to create a balanced dataset, with each class consisting of 100 images using *Undersampling* and *RandomOverSampler* from the *imblearn* package of Python [22], as shown in Figure 1B. In addition, augmentation of horizontal flip was also conducted during the training process.

Guided by [23] and our previous work [24], the dataset was initially split into a training set (70%) and a separate unseen testing set (30%). The training set was then further divided into training (70%) and validation (30%) subsets. As a result, the final dataset distribution consisted of 147 images for training, 63 for validation, and 90 for testing.

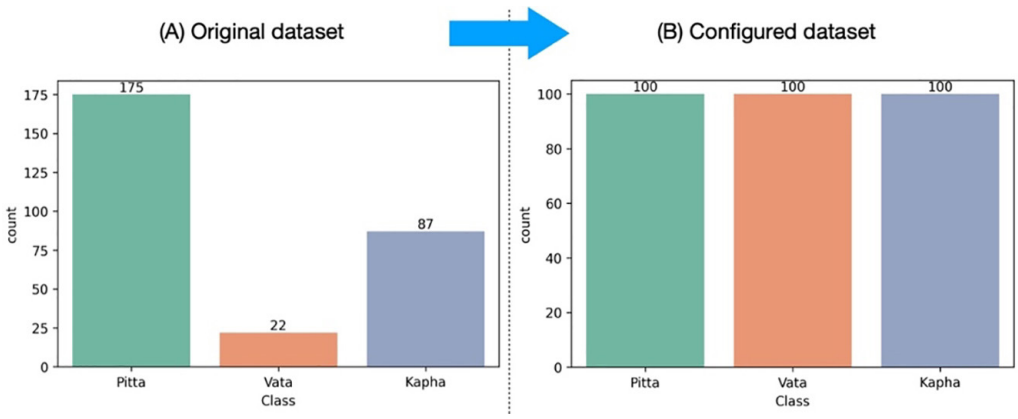

Fig. 1. (A) The original imbalanced dataset was adjusted to create a balanced dataset, (B) The original imbalanced dataset was adjusted to create a balanced dataset

3.2 Physical feature selection

We prepared physical features for both the ML and statistical approaches. Initially, eight physical features were considered: shape, colour, red spots, moisture, coating, middle line, crack line, and teeth marks, as shown in Figure 2. To identify relevant features in the statistical approach, we conducted chi-square (χ^2) tests [25] to assess their association with the target variable. This test evaluates whether the observed distribution of feature values significantly differs from expectations under independence. A low p-value ($p < 0.05$) indicates a significant relationship, justifying feature selection for further analysis.

Based on the chi-square test results, five features with strong statistical relationships were selected: shape, colour, red spots, middle line, and crack line, as presented in Table 3. Features such as moisture, coating, and teeth marks were excluded due to their weak statistical association with the target variable ($p > 0.05$). Additionally, in the ML approach, we further analysed pixel-based features to complement the selected physical features, ensuring a more comprehensive analysis.

Table 3. Selected physical features and their values analysed by chi-square

Selected Physical Feature	Value
Shape	1 = Thin
	2 = Swollen
	3 = Normal (Reference group)
Colour	1 = Pale Shiny
	2 = Pale
	3 = Red Shiny
	4 = Dark Red
	5 = Reddish Purple
	6 = Red (Reference group)
Red spots	1 = Yes
	2 = No (Reference group)
Middle line	1 = Yes
	2 = No (Reference group)
Crack line	1 = Yes
	2 = No (Reference group)

4 ANALYSIS METHODS

Building upon these insights, we developed a comprehensive framework that enhances classification accuracy and interpretability by incorporating both pixel-based and physical feature-based approaches, as presented in Figure 2. The following section describes our methodology in detail.

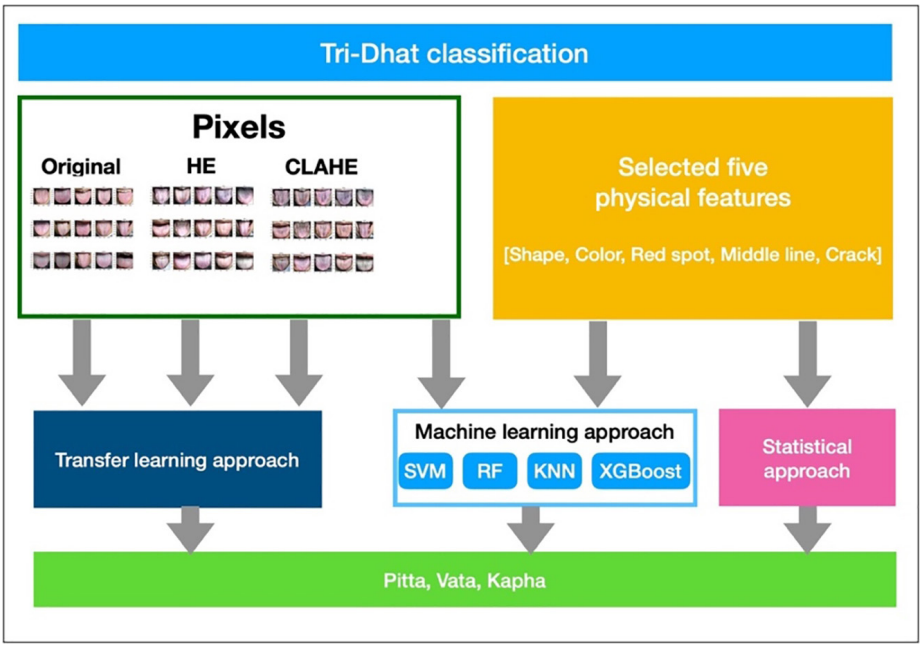

Fig. 2. Analysis framework for Tridhat classification

4.1 Pixel-based methods: Transfer learning and machine learning

Based on the literature review [26], the original RGB images were pre-processed using histogram equalisation (HE) and contrast-limited adaptive HE (CLAHE) to enhance contrast. These processed images were used in both TL and ML approaches.

In the TL approach (see Figure 3), we selected DenseNet121 as the backbone model, as it demonstrated the best performance in our experiments. The dataset included images captured from both DSLR and mobile cameras. To enhance dataset diversity, we applied three different random seed values (1337, 42, and 2024) for splitting.

In the ML approach, we implemented five-fold cross-validation on four classifiers: SVM, RF, k-Nearest Neighbours (KNN), and Extreme Gradient Boosting (XGBoost). These models were trained using two feature sets: pixel data and the five selected physical features [27]. The same dataset splits were used as in the TL approach. Default parameters were applied to all classifiers, except for XGBoost, which was configured with a max depth of 10, a learning rate of 0.3, 200 rounds, and the evaluation metric set to *mlogloss*.

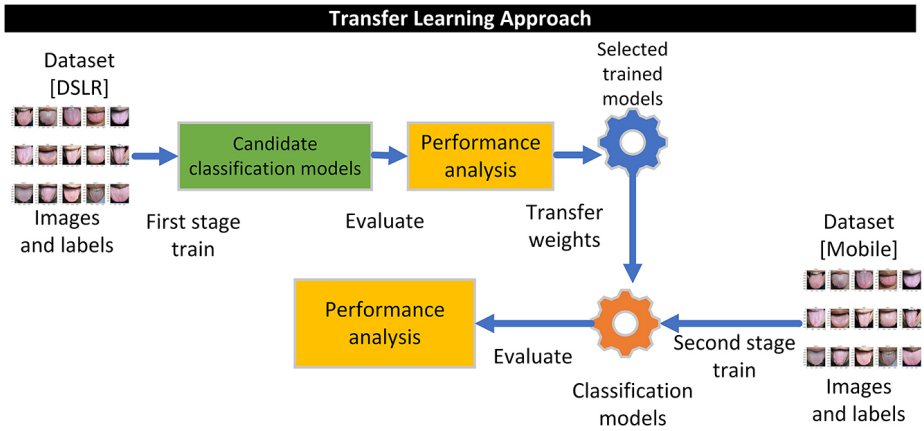

Fig. 3. The transfer learning approach used in the deep learning method

4.2 Physical feature-based methods: statistical models

We applied a multinomial LR model [28] to classify tongue images into three unordered categories: Pitta, Vata, and Kapha. This model is particularly suitable when the response variable consists of more than two categories, as defined in Equation (1).

In this model, $P(Y = j)$ represents the probability of class j, while β_j denotes the model coefficients for that class are denoted X is the feature vector, and K is the number of classes (three in this study). The model enables a comprehensive analysis by examining the relationships between categorical outcomes and multiple predictor variables derived from tongue image characteristics.

Both DSLR and mobile camera datasets were used to validate the statistical model, ensuring consistency across different imaging devices.

$$P(Y = j) = \frac{e^{\beta_j X}}{\sum_{k=1}^{K} e^{\beta_k X}} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$F1_{score} = \frac{2TP}{2TP + FP + FN} \tag{3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

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

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

5 RESULTS AND DISCUSSION

Performance evaluations were conducted for each combination of the dataset and its respective approach.

5.1 Transfer learning and machine learning models

The TL approach was trained and evaluated three times using distinct random seed numbers (1337, 42, and 2024). These seed values were used to partition the dataset into training, validation, and testing sets.

The performance metrics included precision, F1 score, accuracy, sensitivity, and specificity, each defined in Equations 2 through 6. The performance evaluation results were further analysed using two-tailed paired t-tests and single-factor analyses [29].

5.2 Performance analysis and model ensemble

For the pixel-based dataset, performance comparisons of the TL approach are presented in Figure 4. The DenseNet121 model, trained with HE-pre-processed images, outperformed both RGB and CLAHE across all evaluation metrics.

We also employed a model ensemble (ME) evaluation technique for each respective pixel pre-processing method. The results showed that HE outperformed both RGB and CLAHE across all metrics.

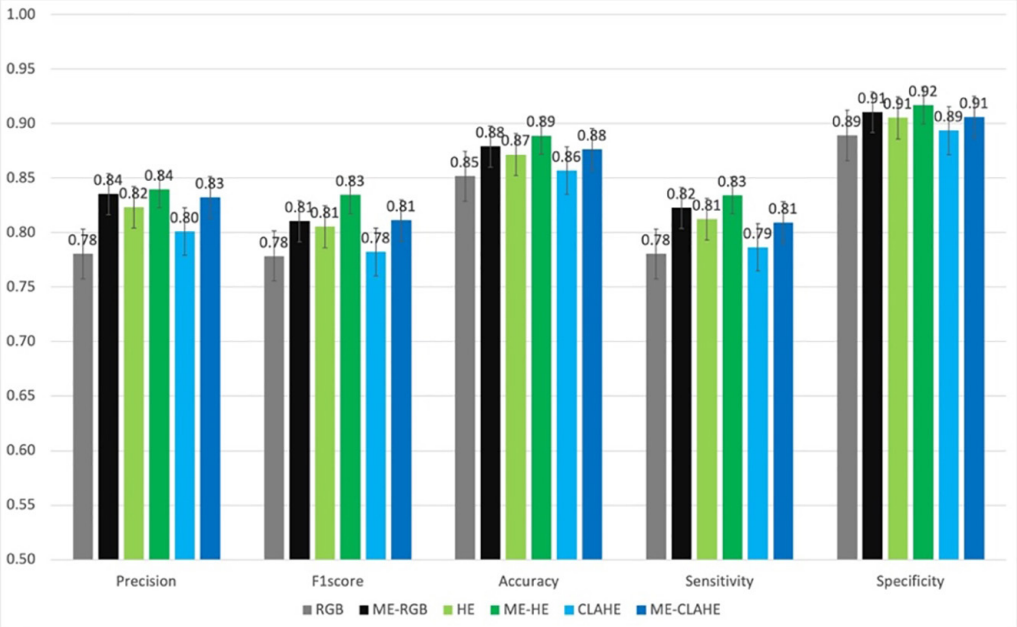

Fig. 4. Performance comparisons for three testing sets of the TL approach using pixel-based features and DenseNet121 as the backbone model

Paired t-tests were conducted to assess the statistical significance of the performance differences, with the alpha level set to 0.05. As shown in Table 4, the ME technique demonstrated better performance than standard evaluations across all pairs, as the p-values were less than 0.05. When comparing the ME techniques, HE also significantly outperformed RGB and CLAHE, with p-values of 0.0314 and 0.0108, respectively.

Table 4. Paired t-tests of TL performance using pixel-based features

Pair	t-Stat	p-Value
RGB vs ME-RGB	–5.9910	0.0039*
HE vs ME-HE	–6.4086	0.0030*
CLAHE vs ME-CLAHE	–6.8556	0.0024*
ME-RGB vs ME-HE	–3.2489	0.0314*
ME-HE vs ME-CLAHE	4.4984	0.0108*

For the physical feature dataset, performance comparisons of the ML approach for all classifiers show that each classifier achieved similar performances, except for the accuracy and specificity of RF and SVM, which were better than the others. However, we selected the RF model as the baseline for comparison with other approaches.

Figure 5 presents the comparisons of the best-performing models from each approach. In the ML approach, the RF classifier trained with the HE-pixel dataset (Pixel-HE-RF) did not significantly outperform the RF model trained with the five

physical features (Five-PF-RF), as indicated by a p-value of 0.0613 (with alpha set to 0.05). In contrast, the DenseNet121 model trained with the HE-pixel dataset (Pixel-HE-DenseNet121) in the TL approach significantly outperformed the RF classifier across all metrics, with a p-value of 2.60E-06 (alpha set to 0.05).

For performance comparisons using the five selected physical features as inputs, a paired t-test revealed a p-value of 0.3342 between the RF classifier and the statistical model, indicating no significant difference in classification effectiveness between these approaches.

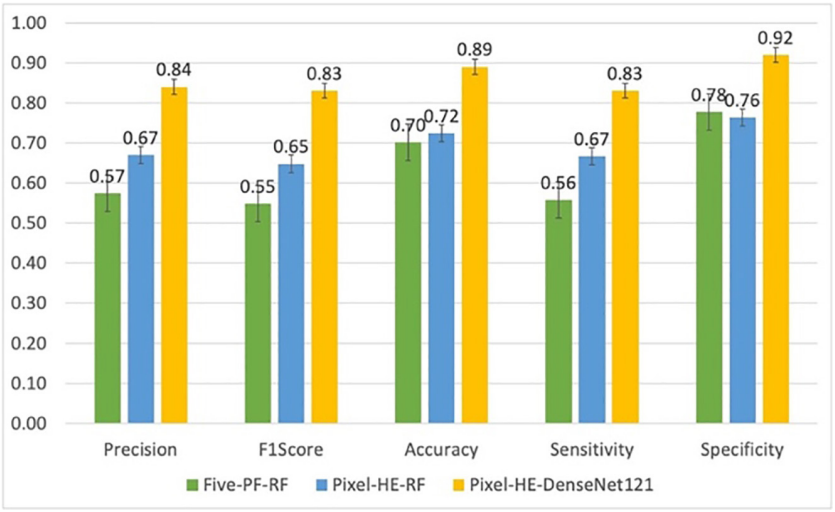

Fig. 5. Performance comparisons of the best models using pixel-based inputs in the TL and ML approaches

5.3 Statistical models

As shown in Table 3, the coding scheme for each selected physical feature was applied in the multinomial LR model. Each feature was assigned a reference group that served as the baseline for comparison. The results of the multinomial LR model using data from DSLR and mobile phone cameras are presented as follows.

Multinomial logistic model with data based on DSLR camera. According to Equations 7 and 8 (multinomial logistic model based on data from DSLR camera images), individuals in the Vata group, relative to Pitta, are more likely to have a thin tongue shape (IRR = 3.083), tongue colours of pale shiny, pale, or red shiny (IRR = 14.547, 3.271, and 14.243, respectively), or cracks on the tongue (IRR = 1.846).

$$\ln\left(\frac{\pi_{\text{Vata}}}{\pi_{\text{Pitta}}}\right) = -2.484 + 1.126 \times \text{Shape}_1 - 0.539 \times \text{Shape}_2 + 2.677 \times \text{Color}_1 + 1.185 \times \text{Color}_2 + 2.656 \times \text{Color}_3 - 19.321 \times \text{Color}_4 - 19.033 \times \text{Color}_5 - 0.791 \times \text{Redspot}_1 - 0.442 \times \text{Midline}_1 + 0.613 \times \text{Crack}_1 \quad (7)$$

$$\ln\left(\frac{\pi_{\text{Kapha}}}{\pi_{\text{Pitta}}}\right) = -1.381 - 0.082 \times \text{Shape}_1 + 1.466 \times \text{Shape}_2 + 2.427 \times \text{Color}_1 + 2.185 \times \text{Color}_2 + 0.995 \times \text{Color}_3 - 0.754 \times \text{Color}_4 + 0.823 \times \text{Color}_5 - 0.638 \times \text{Redspot}_1 - 0.866 \times \text{Midline}_1 - 0.587 \times \text{Crack}_1 \quad (8)$$

For Kapha relative to Pitta, individuals with a swollen tongue shape (IRR = 4.331) or tongue colours of pale shiny, pale, red shiny, or reddish-purple (IRR = 11.325, 8.886, 2.705, and 2.278, respectively) are more likely to belong to the Kapha group than the Pitta group.

Moreover, the results of the goodness-of-fit indicate that the multinomial LR model fits the data from DSLR camera images reasonably well (p-value > 0.05), especially for the Pitta and Kapha categories. However, the model has difficulty accurately classifying the Vata group, with an overall classification accuracy of 72.0%, as shown in Table 5.

Table 5. Classification accuracy for Tridhat categories using data from DSLR camera images

Observed	Pitta	Vata	Kapha	Percent Correct
Pitta	135	3	26	82.3%
Vata	11	2	7	10.0%
Kapha	28	0	56	66.7%
Overall Percentage	64.9%	1.9%	33.2%	72.0%

Multinomial logistic model with data based on mobile camera. Considering equations 9 and 10 (multinomial logistic model based on mobile camera data), individuals in the Vata group, relative to Pitta, are more likely to have a thin tongue shape (IRR = 3.453), a tongue colour that is pale or red shiny (IRR = 6.666 and 13.611, respectively), or cracks on the tongue (IRR = 1.042).

$$\ln\left(\frac{\pi_{Vata}}{\pi_{Pitta}}\right) = -2.748 + 1.239 \times Shape_1 - 0.177 \times Shape_2 - 15.403 \times Color_1 + 1.897 \times Color_2 + 2.611 \times Color_3 - 15.293 \times Color_4 - 15.160 \times Color_5 - 0.826 \times Redspot_1 - 0.687 \times Midline_1 + 0.041 \times Crack_1 \quad (9)$$

$$\ln\left(\frac{\pi_{Kapha}}{\pi_{Pitta}}\right) = -0.511 - 0.430 \times Shape_1 + 1.042 \times Shape_2 - 15.201 \times Color_1 + 1.504 \times Color_2 - 16.821 \times Color_3 - 0.575 \times Color_4 - 15.896 \times Color_5 - 0.872 \times Redspot_1 - 0.942 \times Midline_1 - 0.953 \times Crack_1 \quad (10)$$

For Kapha relative to Pitta, individuals with a swollen tongue shape (IRR = 2.835) or a pale tongue colour (IRR = 4.499) are more likely to belong to the Kapha group than the Pitta group. Furthermore, the multinomial LR model adequately represents the data when using mobile camera images (p-value > 0.05). Similar to the DSLR-based model, it performs well in identifying Pitta, moderately for Kapha, and poorly for Vata, with an overall classification accuracy of 70.1%, as shown in Table 6.

Table 6. Classification accuracy for Tridhat categories using data from mobile camera images

Observed	Pitta	Vata	Kapha	Percent Correct
Pitta	151	1	28	83.9%
Vata	12	0	8	0.0%
Kapha	35	0	46	56.8%
Overall Percentage	70.5%	0.4%	29.2%	70.1%

Comparing the performance of the DSLR-based and mobile camera-based models, Table 7 summarises the metrics for the multinomial logistic model used for Tridhat classification based on DSLR and mobile camera data. The multinomial logistic model performs better with DSLR camera data across all metrics, including precision, F1 score, accuracy, sensitivity, and specificity. This suggests that DSLR images may provide more reliable features for the Tridhat classification task.

Table 7. Performance of the multinomial LR model for Tridhat classification using DSLR and mobile camera data

Camera	Precision	F1 Score	Accuracy	Sensitivity	Specificity
DSLR	0.602	0.564	0.720	0.530	0.796
Mobile	0.441	0.455	0.701	0.469	0.766

Table 8. Performance comparisons with related works

Author	Domain	Method	Accuracy
Xu et al. [11]	TCM	UNet and NN	0.92
Zhou et al. [9]	TCM	ResNet-34	0.92
Shi et al. [10]	TCM	RF, LR, SVM, and NN	0.88
Li et al. [8]	TCM	ViT combined with Grad-CAM and K-means	0.84
This work	TTM	DenseNet121 with HE pre-processed RGB images	0.89

5.4 Compare with related works

Based on the accuracy performance reported in the literature review, we indirectly compared our results with previous studies, as summarised in Table 8. This comparison considers dataset size, diversity, and methodological differences in TCM and TTM domains. Furthermore, our work is among the first to integrate DL, ML, and statistical models for tongue image classification within the TTM domain.

Our model achieved an accuracy of 0.89, which is close to the results reported by Xu et al. (0.92) [11], Zhou et al. (0.92) [9], and Shi et al. (0.88) [10]. Despite methodological differences, these studies collectively demonstrate the potential of AI-assisted tools to improve diagnostic consistency across medical domains.

6 WEB APPLICATION DEPLOYMENT

For the real-world deployment, we selected three optimal models to serve as a pre-diagnostic tool to aid TTM practitioners in tongue analysis. We chose the TL approach using the DenseNet121 model with HE pixel images as inputs. When physical features were used as inputs, we opted for the RF classifier. Additionally, we employed a statistical model using the physical features.

We initially deployed our web-based application system at http://bioservices.sci.psu.ac.th/, as shown in Figure 6. The analysis is conducted at the subject level (patient level), allowing users to upload multiple tongue images via the upload page (see Figure 6A) and select the 'analyze' button. The result page (see Figure 6B) reports the classification result for each tongue image, including the respective probabilities among the three classes. Based on the subject-level classification, a final classification

result for the target class is provided to the user. In addition, the user can save the transaction, which will be logged as part of their personal analysis history.

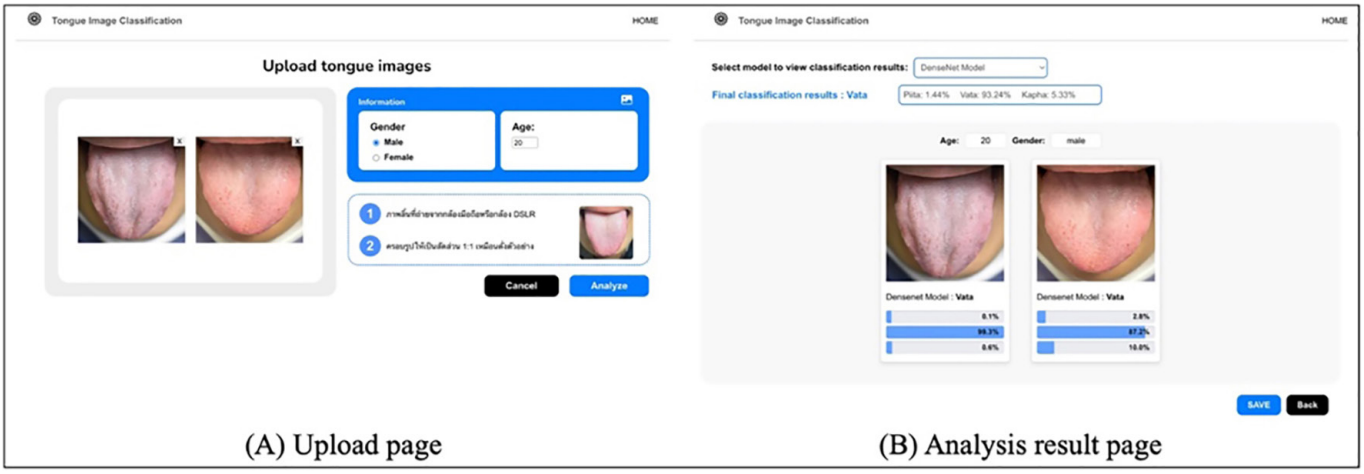

Fig. 6. (A) Upload page, where users can upload tongue images, (B) Analysis result page, displaying classification results and probabilities for each image

7 LIMITATIONS AND FUTURE WORK

Despite applying augmentation techniques, we encountered limitations with the dataset, including a limited number of patients and class imbalance. To mitigate these challenges in future research, we recommend expanding the dataset by collecting more diverse tongue images across different demographics and medical conditions. Additionally, capturing images in RAW format with high-end cameras will help preserve pixel quality and enable more advanced pre-processing techniques during dataset preparation.

Furthermore, the classification results from the models are based on a limited dataset and methodologies specific to TTM. Tridhat classification provides only an overall health condition of a subject within a certain period (e.g., a week ago). Therefore, it cannot directly indicate specific symptoms that would lead to targeted treatments or remedies, even in complementary medicine. Close inspection and consultation with experts remain essential.

To enhance the clinical applicability of tongue image analysis in TTM, future studies could: (1) incorporate a larger and more diverse dataset covering multiple age groups, health conditions, and geographic regions; (2) combine tongue images with additional health indicators (e.g., pulse, medical history) for more precise diagnosis; (3) investigate explainable AI techniques to provide transparent and interpretable insights for TTM practitioners; and (4) conduct clinical validation studies to test model performance in real-world TTM practice.

By addressing these challenges and expanding research directions, future studies can further improve the reliability, accuracy, and applicability of AI-driven tongue image classification in traditional Thai medicine.

8 CONCLUSION

This study presented a novel approach to tongue image classification in TTM by leveraging DL, ML, and statistical models. Leveraging a newly curated dataset

from our TTM university hospital, we analysed both pixel-based features (in various colour modes) and physical characteristics to classify tongue images into Tridhat categories.

Our findings demonstrate that AI-driven classification methods can assist in standardising TTM diagnostics, addressing the subjectivity often associated with practitioner-dependent assessments. Through extensive evaluation, we identified the DenseNet121 model, trained with RGB pixel features pre-processed using HE, as the best-performing model, achieving an accuracy of 0.83, a sensitivity of 0.89, and a specificity of 0.92. Additionally, paired t-tests and model ensembling were applied to validate model performance, further strengthening the reliability of our results.

This study makes significant contributions to the scientific community by:

1. pioneering the integration of DL, ML, and statistical models for automated Tridhat classification in traditional Thai medicine;
2. introducing a high-quality dataset tailored for tongue image classification, serving as a benchmark for future AI-based TTM studies;
3. demonstrating the feasibility of AI models in delivering consistent, objective, and scalable assessments, thereby reducing diagnostic variability among TTM practitioners; and
4. bridging the gap between AI and real-world complementary medicine by integrating the best-performing model into a publicly accessible web application, making automated TTM diagnostics more accessible to both practitioners and the general public.

9 ACKNOWLEDGEMENTS

This study was supported by the Faculty of Science and the Faculty of Traditional Thai Medicine at Prince of Songkla University. OpenAI was used for editing and grammar enhancement.

10 CONFLICTS OF INTEREST STATEMENT

The authors declare that they have no competing interests.

11 DATA SHARING

The dataset used in this study is publicly available at https://zenodo.org/doi/10.5281/zenodo.12525501.

12 REFERENCES

[1] P. Maki *et al.*, "Ethnopharmacological nexus between the traditional Thai medicine theory and biologically based cancer treatment," *J Ethnopharmacol*, vol. 287, p. 114932, 2022. https://doi.org/10.1016/j.jep.2021.114932

[2] K. He, "Traditional Chinese and Thai medicine in a comparative perspective," *Complement Ther. Med.*, vol. 23, no. 6, pp. 821–826, 2015. https://doi.org/10.1016/j.ctim.2015.10.003

[3] V. Tantiveerkul, *Textbook of Traditional Thai Medicine, Volumes 1–3*. Bangkok: Traditional Medicine School, Wat Phra Chetuphon Vimolmangklararam Rajwaramahaviharn, 1957.

[4] X. Wu, H. Xu, Z. Lin, S. Li, H. Liu, and Y. Feng, "Review of deep learning in classification of tongue image," *Journal of Frontiers of Computer Science and Technology*, vol. 17, no. 2, pp. 303–323, 2023. https://doi.org/10.3778/j.issn.1673-9418.2208052

[5] Z. Tian *et al.*, "Current status and trends of artificial intelligence research on the four traditional Chinese medicine diagnostic methods: A scientometric study," *Ann. Transl. Med.*, vol. 11, no. 3, p. 145, 2023. https://doi.org/10.21037/atm-22-6431

[6] X. Zhang, Z. Chen, J. Gao, W. Huang, P. Li, and J. Zhang, "A two-stage deep transfer learning model and its application for medical image processing in traditional Chinese medicine," *Knowl. Based Syst.*, vol. 239, p. 108060, 2022. https://doi.org/10.1016/j.knosys.2021.108060

[7] X. Wang *et al.*, "Syndrome types classification method of skin diseases based on tongue hierarchical feature fusion," *Res. Sq.*, 2023. https://doi.org/10.21203/rs.3.rs-3490132/v1

[8] J. Li *et al.*, "A multi-step approach for tongue image classification in patients with diabetes.," *Comput. Biol. Med.*, vol. 149, p. 105935, 2022. https://doi.org/10.1016/j.compbiomed.2022.105935

[9] J. Zhou *et al.*, "Weakly supervised deep learning for tooth-marked tongue recognition," *Front. Physiol.*, vol. 13, 2022. https://doi.org/10.3389/fphys.2022.847267

[10] Y. Shi *et al.*, "A new method for syndrome classification of non-small-cell lung cancer based on data of tongue and pulse with machine learning," *Biomed. Res. Int.*, vol. 2021, pp. 1–14, 2021. https://doi.org/10.1155/2021/1337558

[11] Q. Xu *et al.*, "Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network," *IEEE J. Biomed Health Inform.*, vol. 24, no. 9, pp. 2481–2489, 2020. https://doi.org/10.1109/JBHI.2020.2986376

[12] J. Ma, G. Wen, C. Wang, and L. Jiang, "Complexity perception classification method for tongue constitution recognition," *Artif. Intell. Med.*, vol. 96, pp. 123–133, 2019. https://doi.org/10.1016/j.artmed.2019.03.008

[13] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *The Ninth International Conference on Learning Representations*, 2021. https://openreview.net/pdf?id=YicbFdNTTy

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015. https://doi.org/10.48550/arXiv.1512.03385

[16] Z. Guo, S. Feng, L. Wang, and M. Zhang, "A tongue image classification method in TCM based on multi feature fusion," in *Cognitive Computation and System, ICCCS 2023, Communications in Computer and Information Science*, F. Sun and J. Li, Eds., vol. 2029, 2024. https://doi.org/10.1007/978-981-97-0885-7_2

[17] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016. https://doi.org/10.48550/arXiv.1602.07261

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, pp. 1–14, 2014. https://doi.org/10.48550/arXiv.1409.1556

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[20] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971. https://doi.org/10.1037/h0031619

[21] H. Vega-Huerta, K. R. Pantoja-Pimentel, S. Y. Quintanilla-Jaimes, G. L. E. Maquen-Niño, P. De-La-Cruz-VdV, and L. Guerra-Grados, "Classification of Alzheimer's disease based on deep learning using medical images," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 10, pp. 101–114, 2024. https://doi.org/10.3991/ijoe.v20i10.49089

[22] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. http://jmlr.org/papers/v18/16-365.html

[23] G. L. E. Maquen-Niño, J. G. Nuñez-Fernandez, F. Y. Taquila-Calderon, I. Adrianzén-Olano, P. De-La-Cruz-VdV, and G. Carrión-Barco, "Classification model using transfer learning for the detection of pneumonia in chest X-ray images," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 5, pp. 150–161, 2024. https://doi.org/10.3991/ijoe.v20i05.45277

[24] K. Damkliang, T. Wongsirichot, and P. Thongsuksai, "Tissue classification for colorectal cancer utilizing techniques of deep learning and machine learning," *Biomedical Engineering: Application, Basis, and Communications*, vol. 33, no. 3, p. 2150022, 2021. https://doi.org/10.4015/S1016237221500228

[25] R. A. Fisher, "On the interpretation of χ^2 from contingency tables, and the calculation of P," *J. R. Stat. Soc.*, vol. 85, no. 1, pp. 87–94, 1922. https://doi.org/10.1111/j.2397-2335.1922.tb00768.x

[26] M. H. Tania, K. Lwin, and M. A. Hossain, "Advances in automated tongue diagnosis techniques," *Integr. Med. Res.*, vol. 8, no. 1, pp. 42–56, 2019. https://doi.org/10.1016/j.imr.2018.03.001

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[28] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression* (2nd ed.). New York, NY: Wiley, 2000. https://doi.org/10.1002/0471722146

[29] J. Han, M. Kamber, and J. Pei, "8 – classification: Basic concepts," in *Data Mining (Third Edition),* in The Morgan Kaufmann Series in Data Management Systems, J. Han, M. Kamber, and J. Pei, Eds., 2012, pp. 327–391. https://doi.org/10.1016/B978-0-12-381479-1.00008-3

13 AUTHORS

Kasikrit Damkliang received B.Sc. in computer science, M.Eng. in computer engineering, and the Ph.D. in computer engineering from the Prince of Songkla University (PSU), Hat Yai, Thailand, in 2005, 2009, and 2019, respectively. He is currently serving as an Assistant Professor with the Division of Computational Science, Faculty of Science, PSU. His research interests include medical image analysis, biosignal analysis, deep learning and machine learning, bioinformatics, web service, cloud computing, and workflow technology (E-mail: kasikrit.d@psu.ac.th).

Julara Chumnaul received B.Sc. in Statistics from Maejo University and M.Sc. in Applied Statistics from Chiang Mai University, Thailand, respectively, in 2007 and 2010, and received Ph.D. in Mathematical Sciences (Statistics track) from Mississippi State University, USA, in 2019. Currently, she is working as an Assistant Professor in the Division of Computational Science, Faculty of Science, Prince of Songkla University, Thailand. Her research interests include power-law process (PLP), the theory of repairable system reliability, and statistical inference.

Teerawat Sudkhaw received B.Sc. in Traditional Thai Medicine and Master of Traditional Thai Medicine from the Faculty of Traditional Thai Medicine, PSU, Hat Yai, Thailand, in 2010 and 2021, respectively. He is currently a Traditional Thai Medicine Doctor at the Traditional Thai Medicine Hospital, PSU. His research interests include the treatment of diseases through traditional Thai medicine, the development and research of herbal medicines, and Thai massage.

Thitinan Yingtawee and **Nasma Saearm** are senior students in the Information and Communication Technology (ICT) program, Division of Computational Science, Faculty of Science, PSU, Thailand.