ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Leveraging Transfer Learning for Tri-Dhat Classification of Tongue Images in Traditional Thai Medicine

Kasikrit Damkliang[1], Teerawat Sudkhaw[2], Thitinan Yingtawee[3], Nasma Saearma[4], Kotchakorn Moosigapong[5], Patcharawalai Jaisamut[6], Julalak Chokpaisarn[7], Sirinat Laman[8], Anintita Intan[9] and Anyamanee Ladam[10]

ABSTRACT

Traditional Thai medicine (TTM) is a popular and increasingly accepted treatment option. In TTM, tongue diagnosis is a highly efficient method for assessing overall health, yet its accuracy can vary significantly depending on the practitioner's expertise. In this work, we hypothesize that deep learning-based transfer learning (TL) methods can achieve high accuracy in the Tri-Dhat classification of tongue images, a system that aligns with TTM principles and categorizes the tongue into three types: Vata, Pitta, and Kapha. We propose an approach that uses raw pixel data and artificial intelligence (AI) to support TTM diagnoses. For our analysis, we used a unique dataset of genuine tongue images collected from our university's TTM hospital. To prepare the data, we performed class balancing and data augmentation. We then developed TL techniques with a variety of pretrained deep learning models. For performance comparisons, we utilized two-tailed paired t-tests and single-factor ANOVA. Our experiments showed that the DenseNet121 and Xception models produced the most significant results with cropped image datasets, including both DSLR- and mobile-taken images. Notably, an ensemble of these models yielded the highest average predictions. This ensemble achieved an accuracy of 0.96, a precision of 0.94, an F1 score of 0.96, a sensitivity of 0.96, and a specificity of 0.97. These results were further supported by a p-value of 0.0003 from the ANOVA analysis. We suggest that our methods could be effectively deployed in real-world scenarios to aid TTM practitioners in their diagnoses.

Article information:

Keywords: Traditional Thai Medicine, Tongue Image Classification, Deep Learning, Transfer Learning, Tri-Dhat, Variant Analysis

Article history:
Received: January 19, 2025
Revised: July 17, 2025
Accepted: August 7, 2025
Published: August 9, 2025
(Online)

DOI: 10.37936/ecti-cit.2025193.260498

1. INTRODUCTION

Ancient medical traditions have relied on knowledge accumulated over more than 3,000 years, using non-intrusive diagnoses of organs, including the tongue, pulse, and face. These diagnoses help evaluate the overall health of the heart, liver, spleen, lungs, and kidneys in the human body [1].

Historically, insights into tongue diagnostics have been primarily informed by two ancient medical traditions: Ayurveda from traditional Indian medicine and traditional Chinese medicine (TCM) [2]. TCM has emerged as the leading methodology in tongue diagnosis for both diagnosis and classification [1].

Building on this foundation, the realm of Ayurveda offers insights into the functions of internal organs through the characteristics of the tongue, classified into three doshas: Vata, Pitta, and Kapha [3] [4] [5]. This classification aligns with the Tri-Dhat concept in traditional Thai medicine (TTM). Additionally, Kapha can be called Semha or Saled in the TTM [6].

[1,3,4] The authors are with Division of Computational Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand., E-mail: kasikrit.d@psu.ac.th, 6410210470@psu.ac.th, 6410210163@psu.ac.th

[2,5−10] The authors are with Faculty of Traditional Thai Medicine, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand. [2,5,7] The authors are with Traditional Thai Medical Research and Innovation Center, Faculty of Traditional Thai Medicine, Prince of Songkla University. Hat Yai, Songkhla 90110, Thailand., E-mail: teerawat.sud@psu.ac.th, kotchakorn.m@psu.ac.th, patcharawalai.j@psu.ac.th, julalak.c@psu.ac.th, 6411410068@psu.ac.th, 6411410077@psu.ac.th, 6411410188@psu.ac.th

[1] Corresponding author: kasikrit.d@psu.ac.th

Fundamental inspection methods in TTM consist of pulse examination, evaluation of the body's elements (earth, water, wind, and fire), and Tri-Dhat functions. Tongue diagnosis, in particular, is a highly efficient method for determining overall health [7] [8].

Table 1 summarizes the characteristics of the Tri-Dhat that are relevant to health [7] [9]. Symptoms of excess Pitta include feeling hot and the presence of red or black spots. Symptoms of excess Vata include pain, diarrhea, and movement disorders, which can lead to the distortion of the body structure or the displacement of organs. Symptoms of excess Kapha include feeling full, sleepy, and a sensation of something covering the heart.

Despite the rich history of tongue diagnosis in TTM, there is currently a lack of studies and evidence on its variations. Practitioners typically rely on visual cues such as color, shape, moisture, movement, and coating to identify Tri-Dhat imbalances, drawing on their expertise [6] [10] [11]. Despite its importance, there is a notable absence of research and in silico methodologies for analyzing tongue images, as well as tools or software to support TTM practitioners in their diagnoses.

In this work, we propose in-silico methods for Tri-Dhat analysis and classification of tongue images in TTM using artificial intelligence (AI). We leveraged genuine tongue images collected from subjects at our university's TTM hospital. We introduce designed datasets and analysis approaches based on TL techniques of deep learning (DL) methods. The experimental results and performance evaluations conducted using statistical tools, demonstrate the effectiveness of our proposed approach in supporting TTM practitioners in their diagnoses.

2. RELATED WORKS

In this section, we review recent related works on tongue image analysis, particularly those emphasizing the role of AI, as summarized in Table 2.

Joshi *et al.* [2] developed a computerized, pragmatic assessment method for the tridoshas using 120 tongue images collected from adults not undergoing any treatment. They employed K-nearest neighbors (KNN), neural networks, and decision trees (DTs) [19] for the classification task, with the DT model demonstrating superior performance, achieving sensitivities ranging from 0.72 to 0.83.

In parallel to the domain of TCM, Tian *et al.* [1] conducted a systematic review of the current status and trends in AI research on TCM diagnostic methods developed since 2007, encompassing tongue, pulse, and facial diagnoses. Their study revealed that the majority of research efforts have been concentrated on tongue diagnosis and classification, predominantly employing deep neural networks.

To further illustrate this point, Wang *et al.* [12] utilized 1,548 tongue images, captured using various types of equipment, to develop a classification method aimed at recognizing unhealthy tongues with tooth marks. This method, based on the ResNet34 [20] convolutional neural network (CNN) architecture, achieved an overall accuracy exceeding 0.90. Importantly, it demonstrated successful model generalization across images captured by different devices.

By expanding the application scope, Xiang *et al.* [14] and Li *et al.* [15] explored the potential of AI for diagnosing diabetes through tongue analysis. Xiang *et al.* [14] collected data from 165 subjects across 11 medical institutions in Tianjin, China, focusing specifically on diabetes mellitus. This study developed a method that applies the random forest (RF) algorithm [19] to analyze fundus photography in conjunction with physical and physiological features of the tongue and pulse. This approach achieved precision, sensitivity, and F1 scores of 0.89, 0.67, and 0.76, respectively.

Conversely, Li *et al.* [15] adopted a machine learning approach to specifically analyze tongue features for diabetes prediction. The study analyzed data from 570 individuals, including diabetic, prediabetic, and normal individuals, collected from 2011 to 2019. It utilized fusion data of color and texture and employed a ResNet-50-based model optimized with a genetic algorithm (GA) and XGBoost [21] for feature extraction. The model achieved precision, sensitivity, and F1 scores of 0.84, 0.81, and 0.80, respectively, demonstrating the relevance of tongue characteristics in detecting diabetes.

In 2022, Li *et al.* [16] used tongue images obtained from a diagnostic instrument to classify diabetes. A tongue diagnosis system was utilized to extract features from these images, which were then analyzed using K-means and validated using Vision Transformer (ViT) [22] incorporating Grad-weighted Class Activation Mapping (Grad-CAM). ViT achieved the highest Top-1 classification accuracy of 0.88.

In the domain of cancer, Ding *et al.* [13] proposed a method for syndrome classification and prediction for primary liver cancer (PLC). The method used 10,602 medical records from PLC patients, including tongue diagnosis information and other features, to train SVM and Bayesian networks, which were incorporated using particle swarm optimization. Their method achieved an accuracy of 0.86. Another interesting study was proposed by Shi *et al.* [17]. The study collected tongue images from 219 patients with non-small cell lung cancer (NSCLC), the most common histological type of lung cancer. They proposed a method for NSCLC stage classification using tongue features, tumor markers, and neural networks (NN), support vector machines (SVM), decision trees, and logistic regression. The neural network achieved the highest accuracy of 0.77.

In 2023, the integration of TCM with DL has advanced towards analyzing more specific pathological

Table 1: *Characteristics of Tri-Dhat relevant to health*

Tri-Dhat	Pitta	Vata	Kapha
Other names	Dee	Lom	Semha/Saled
Symbolized	Fire	Air	Water
Physical structures	The tongue is medium-sized with a red or dark red color. It may have red spots, a red tip, or appear eroded. A yellow or yellowish-white coating may be present.	The tongue is typically thin, dry, and has a pale or light pink color with no coating. It may show small cracks all over or a deep fissure down the center. A trembling tongue may also be observed.	The tongue is characteristically thick and large, with teeth marks on the sides or a swollen appearance. It has a pinkish-white color and is often covered with a thick white or sticky coating. The tongue looks moist, sticky, or has mucus.
Function	It is the source of heat, digestion, hunger, thirst, thought, intelligence, memory, and emotions.	It is the source of power, movement, speech, and control of the mind, as well as the feeling of understanding and the nervous system.	It makes the body soft, mellow, strong, and patient. Semha has a nourishing function and supports the brain.
Effect of unbalanced or excess	Feeling hot like fire, burning sensations, sourness, profuse sweating with a bad odor, and the presence of red or black spots	Pain, numbness, cracking, deterioration, splitting, breaking, a feeling of constriction, diarrhea, and movement disorders. Loss of senses such as smell, sound, heat, sweating, anxiety, sadness	Feeling full, sleepy, bloated, excessive salivation, weight gain, a sensation of something covering the heart, pale skin, and a heavy feeling

Table 2: *Studies on tongue image analysis, particularly emphasizing the role of AI in terms of classification*

Author	Data	Method	Target	Performance
Joshi *et al.* (2020) [2]	120 images	KNN	Tridoshas of Ayurveda	0.83 sensitivity
Wang *et al.* (2020) [12]	1,548 images	ResNet34	Unhealthy tongue with tooth marks	0.90 accuracy
Ding *et al.* (2021) [13]	Tongue diagnosis information	SVM, Bayesian networks, particle swarm optimization	Liver cancer	0.86 accuracy
Xiang *et al.* (2021) [14]	165 subjects of fundus with tongue and pulse	RF	Diabetes mellitus	0.89 precision, 0.67 sensitivity, 0.76 F1 score
Li *et al.*(2021) [15]	570 cases with tongue features	ResNet50, GA, and XGBoost	Diabetes	0.84 precision, 0.81 sensitivity, 0.80 F1 score
Li *et al.* (2022) [16]	Tongue images obtained from a diagnosis instrument	ViT combined with Grad-CAM and K-means	Diabetic tongue	0.88 accuracy
Shi *et al.* (2023) [17]	Tongue images from 219 lung cancer patients	NN	Lung cancer stages	0.77 accuracy
Lu *et al.* (2024) [18]	1,083 tongue images from 741 patients	DenseNet-201	Fibrosis vs. Non-fibrosis	0.81 accuracy, 0.82 sensitivity, 0.81 specificity

structures of the tongue. Yan *et al.* [23] introduced a novel method for tongue crack segmentation, aimed at characterizing the pathologies of the spleen and stomach. The dataset, comprising 176 tongue images with cracks and 140 images without cracks, was sourced from Shanghai University of TCM and augmented to train a segmentation model. This model outperformed several established methods, including the mask region-based convolutional neural network (Mask R-CNN) [24], DeeplabV3+ [25], U-Net [16], UNet++ [27], and semantic segmentation with adversarial learning (SegAN) [28], in comparative analyses.

Similarly, Feng *et al.* [28] presented a progress report on the objectification of tongue diagnosis over the past decade, including a comparative analysis of various segmentation models. The study utilized real tongue image datasets and found that models based on the U-Net architecture consistently outperformed others in terms of precision, sensitivity, and the mean intersection over union (MIoU) metric. Additionally, the research highlighted the challenges associated with using mobile device-captured tongue images in complex environmental conditions.

Another study by Lu *et al.* [29] presented an early-stage disease screening method for hepatic fibrosis using 1,083 tongue images taken from 741 patients with a DSLR camera, and a neural network model, the DenseNet201. The authors employed rigorous data augmentation to generate 13,381 images for training the model. Ultrasound elastography examinations obtained from an instrument were used as a reference standard. The results showed an accuracy of 0.845 and 0.814 in the validation and test sets, respectively.

In this work, we utilized images captured from both a digital single-lens reflex (DSLR) camera and a mobile device camera within controlled environment settings. We proposed a classification method based on the Tri-Dhat concept from the TTM, which consists of three classes: Vata, Pitta, and Kapha.

3. MATERIALS AND METHODS

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

3.1 Data Acquisition

Data acquisition was conducted with volunteers aged between 18 and 60 years who visited the Thai Traditional Medicine Hospital, Prince of Songkla University. Table 3 summarizes the criteria for data acquisition in our study. The physical and mental health status of these volunteers was assessed.

The sample size was determined using Yamane's formula (1973), as defined in Equation 1 [30], where N represents the average monthly patient visits to the hospital (535 patients), with a confidence level of 95% (at $\alpha = 0.05$) in 2022.

$$n = \frac{N}{1 + N(e)^2} \tag{1}$$

Two sets of tongue images were collected from each patient: the first set was taken using a DSLR camera (Nikon D3400 with lens specifications of 18-55mm f/3.5-5.6G) and the second set was captured using a mobile phone camera (iPhone 13 with lens specifications of 1.5-5.1mm F/1.6-2.4). Both sets of photographs were taken alongside a calibration color chart to ensure consistency with true-to-life visuals. A total of 286 patients, including 71 males and 215 females, were included, yielding 572 images, as presented in Table 4.

Table 3: *Data acquisition criteria*

Attribute	Detail
Target population	Tongue images
Study population	Cases with age ranges from 18 to 60 years old.
Inclusion criteria	Normal physical conditions Volunteer gives consent
Exclusion criteria	Tongue affected by any disease TTM practitioners cannot identify Tri-Dhat Volunteer denies consent
Sample size	286 cases

Table 4: *Demographic and tongue characteristics of 286 patients from the Traditional Thai Medicine Hospital, Prince of Songkla University. The data are presented as n (%) patient prevalence, min, max, and mean age*

Variable	Category	Study population
Sex	Male	71 (24.8)
	Female	215 (75.2)
Age	Maximum	59
	Minimum	18
	Mean	21.85
Tri-Dhat of tongue	Pitta	175 (61.6)
	Vata	22 (7.8)
	Kapha	87 (30.6)

3.2 Data Preprocessing and Labelling

The tongue images were verified against the calibration chart for color accuracy. Images captured with the DSLR camera were saved in Joint Photographic Experts Group (JPEG) format, while those taken with the mobile phone camera were stored in the High Efficiency Image File Format (HEIC). The HEIC format is advantageous because it maintains higher image quality at half the file size of JPEG.

Subsequently, the images were cropped to ensure a minimum width of 500 pixels and then, annotated to identify physical structures, including color, shape, moisture, coating, cracks, and teeth marks. Subsequently, these images were classified into their respective Tri-Dhat categories by three independent TTM practitioners (T.S., K.M., and P.J.), each with at least five years of experience. The practitioners followed established TTM diagnostic criteria, using the annotated features to determine the Tri-Dhat classification (see the physical structures in Table 1). The initial class label for each image was determined by a majority vote among the practitioners. From an original pool of 286 cases, two were excluded due to lost image files, resulting in a final dataset of 284 cases for analysis.

In TTM knowledge, there is a clear correlation between age and Tri-Dhat types: Kapha dominates for the age range of 1 to 16 years, Pitta for 17 to 32 years, and Vata for those over 32 years old [31]. This theoretical distribution aligns with the demographics of our study. Since most of our volunteers were university students, staff, and other visitors, the average age was 21.85 years. As a result, our tongue Tri-Dhat class distribution was sharply imbalanced, with a disproportionate number of cases in the Pitta category.

To assess the consistency of the classification rules among the practitioners, we measured interobserver variability using Fleiss' Kappa [32]. The formula is defined in Equations 2 to 4, where $\bar{P}$ is the observed agreement (the average proportion of times that the TTM practitioners agreed on a particular category for an image) and $\bar{P}_e$ is the expected agreement by chance. The variables are defined as follows: n_{ij} is the number of TTM practitioners who assigned the i-th subject to the j-th category, P_j is the proportion of all assignments that were to the j-th category, N is the total number of subjects, n is the number of ratings per subject, and k is the number of categories.

To assess the reliability of using a majority vote, we calculated the inter-rater agreement among the three TTM practitioners using Fleiss' Kappa. A value of 1.0 represents perfect agreement, while values from 0.81–0.99 are considered almost perfect, 0.61–0.80 substantial, and 0.41–0.60 moderate. Our calculated kappa value was 0.30. This value falls into the 'fair' agreement category (0.21–0.40), indicating that a simple majority vote would not be sufficiently reliable and could lead to misclassification due to significant minority opinions.

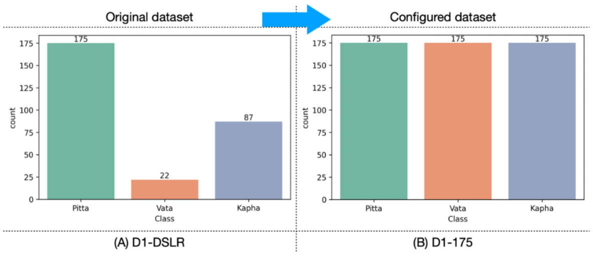

Fig.1: *(A) Distribution of imbalanced classes for Tri-Dhat tongue images taken with a DSLR camera in Dataset D1, featuring 284 images. The classes were determined by majority voting from three TTM practitioners, resulting in a Fleiss' kappa of 0.30. (B) Balanced distribution of Tri-Dhat classes for tongue images taken with a DSLR camera in Dataset D1, comprising 525 images.*

Therefore, we determined that a consensus-based approach was necessary for the final labels. The kappa score was not used to decide the labels directly; rather, the low score justified our decision to move beyond majority voting. For every image where there was disagreement, all three practitioners engaged in a discussion to finalize the classification, ensuring a single, validated label for each image, as detailed in Table 4.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2}$$

$$\bar{P} = \frac{1}{N \cdot n(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - n \tag{3}$$

$$\bar{P}_e = \frac{1}{N \cdot (n-1)} \sum_{j=1}^{k} P_j^2 \tag{4}$$

3.3 Dataset Creation

We compiled a dataset named "Dataset D1." The image sizes varied according to the physical anatomy of the subject. A second dataset, "Dataset D2," was created where the region of interest (ROI) representing only the tongue pixels was extracted from each image, and the background was set to zero.

Considering real-world applications, where a TTM practitioner typically captures a tongue image and crops it similarly to that in Dataset D1, extracting the ROI can be labor-intensive and time-consuming. To address this, we assembled "Dataset D3," which includes images such as those in Dataset D1 accompanied by their ground truth ROI masks.

All datasets were rigorously curated to support the analysis approaches, which will be discussed in the subsequent section.

Figure 1 (A) displays the distribution of labeled tongue images across the imbalanced classes of the dataset. To address this imbalance, we applied the RandomOverSampler algorithm [33], achieving a balanced distribution, as shown in Figure 1 (B) with each class now containing 175 images. Consequently, the total dataset size increased to 525 images.

The dataset was split into training and testing sets at a 70:30 ratio. Additionally, 30% of the training set was further divided into a validation set, resulting in 256, 111, and 158 images for the training, validation, and testing sets, respectively

Exhaustive empirical experiments were conducted. Initially, we determined the optimal image size for model training. A size of 299 × 299 pixels was selected and consistently used for all models except for ViT_b16, which used a default image size of 384 × 384 pixels.

Data augmentation was configured and performed during training. A batch of augmented images was fed into the built-in image preprocessing function of a pretrained model, then was horizontally flipped.

3.4 Experimental Design

To systematically investigate the effectiveness of the TL technique for Tri-Dhat classification, we designed a study with three distinct analysis approaches (A1, A2, and A3). This technique is recognized for its simplicity and efficiency [34] [35]. Approaches A1 and A2 focus on classification tasks, as shown in Figures 2 and 3, respectively. Conversely, Approach A3 combines segmentation and classification tasks, utilizing Dataset D3, as depicted in Figure 4.

3.4.1 Approach A1

In Approach A1, we conducted an exploratory analysis using a combined dataset to identify optimal classification models. For this purpose, we trained various candidate pretrained models, incorporating transferred weights from ImageNet [36]. This approach utilizes both Dataset D1 (a diverse range of cropped images that vary in size) and Dataset D2 (images where only the tongue's highly specific ROI is extracted). Both DSLR-taken and mobile-taken images from these datasets were employed to evaluate whether applying the TL technique to this comprehensive dataset could improve the model's ability to generalize across different image preprocessing methods. The classification task for this approach is shown in Figure 2.

The candidate classification models included DenseNet121 [37], EfficientNetB7 [38], InceptionResNetV2, InceptionV3 [39], MobileNet, MobileNetV2 [40], VGG16, VGG19 [41] Xception [42] [43], and the Vision Transformer (ViT) [22]. Each classification model served as a feature extractor, with a fixed architecture and configurations across all approaches (A1, A2, and A3), as illustrated in Figure 5.

3.4.2 Approach A2

In Approach A2, we designed a two-stage training process to evaluate the effectiveness of TL between images from different acquisition devices. This method utilized Dataset D1, which consists of both DSLR-taken and mobile-taken images. In the first training stage, a model was trained on the DSLR-taken images (D1-DSLR), which were chosen due to their high pixel quality and strong empirical performance. Subsequently, the trained weights from this stage were transferred to a new model, which was then fine-tuned on the mobile-taken images (D1-Mobile). The purpose of this two-stage approach was to determine if knowledge gained from high-quality DSLR images could be leveraged to enhance model robustness and improve classification performance on mobile-taken images. The classification task for this approach is depicted in Figure 3.

3.4.3 Approach A3

Approach A3 was created to explore a more practical, two-stage process, designed with real-world applications in mind. This approach combines an initial tongue segmentation task with a subsequent classification task, using Dataset D3 for both. For instance, a TTM practitioner might input a cropped tongue image, and the model would focus exclusively on the tongue area to aid in diagnosis. The combined process is shown in Figure 4.

Initially, we trained and validated candidate U-Net-based segmentation models using images and their respective ground truth ROI masks. This was to determine the optimal initial weights for models that demonstrated superior focus on the ROI of the tongue. The candidates for these U-Net-based models included our previously proposed segmentation model [44], HuggingFace Transformers [45], and various architectures of pretrained segmentation models enhanced with ImageNet weights [46].

In the subsequent classification task, we employed the encoder path of the segmentation model as a backbone to construct the architecture for the classification task, as illustrated in Figure 5. Only the trained weights from the encoder path were transferred to the classification model. This model was then trained and validated using images from Dataset D3 (without ground truth masks) and Tri-Dhat class labels. Ultimately, this resulted in optimally trained models for Tri-Dhat classification through the use of TL techniques.

3.4.4 Configurations, Classification Architecture, and Model Selection

We used a range of pretrained DL models from both the ViT-based and CNN-based families to conduct the TL technique. For all three approaches, the classification architecture with its configurations is presented in Figure 5. This architecture consists

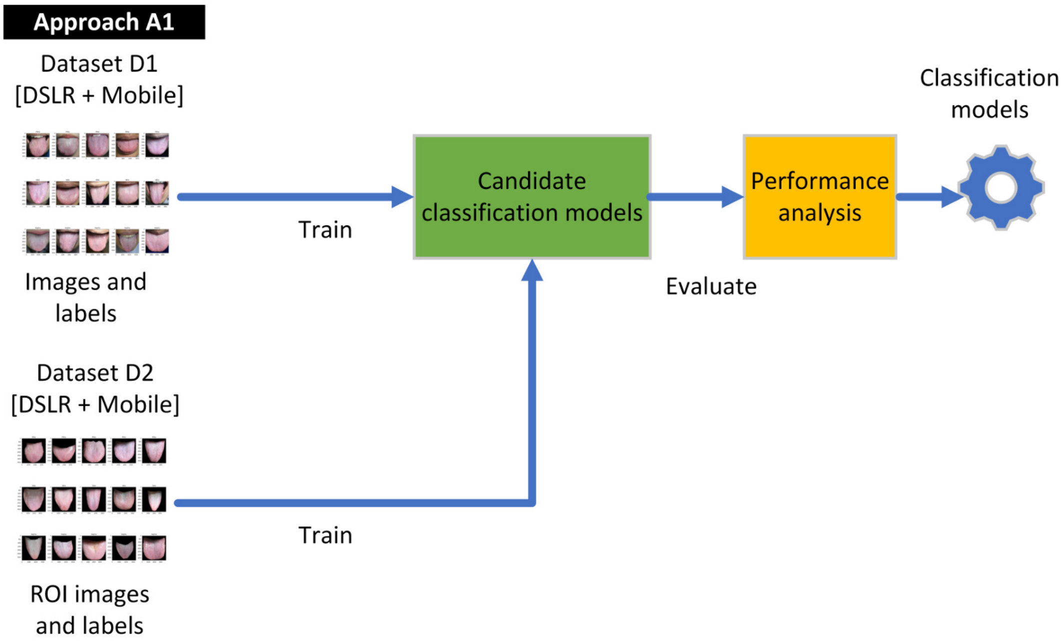

Fig.2: *Tongue image analysis for Approach A1 utilizes the D1 and D2 datasets employing a simplicity TL technique of candidate pretrained models.*

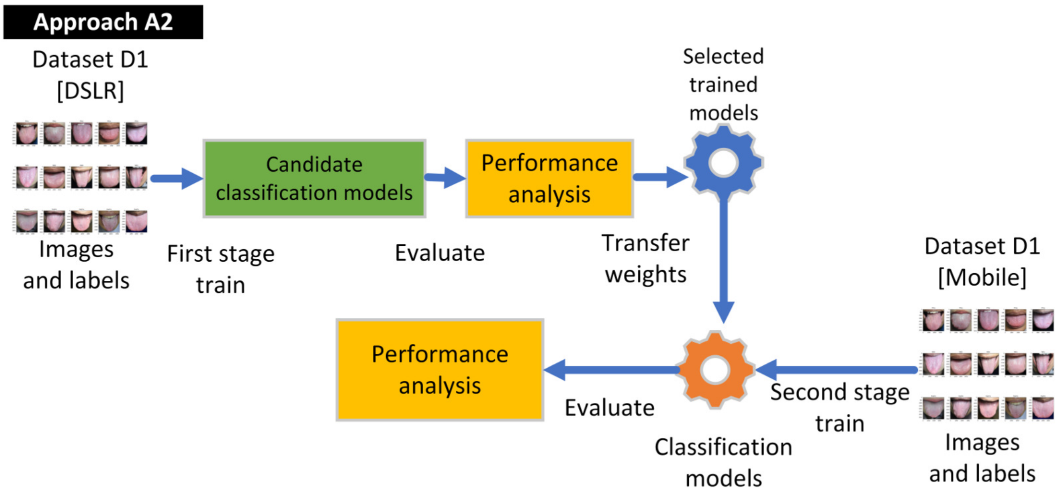

Fig.3: *Tongue image analysis for Approach A2 consists of two training stages: weights from the trained model with DSLR images are transferred to a new model, which is then trained using mobile images.*

of five dense layers with 1024, 512, 64, 32, and 16 units, each employing a ReLU activation function, except for the ViT model, which uses a GeLU activation function. The final output layer, aimed at three predictive classes, utilizes a softmax activation function. A dropout rate of 0.2 was applied to all dense layers. All hyperparameters and their configurations were frozen for the proposed approaches (details in the Dataset and code availability section). The four models that showed the best performance were then selected for a statistical analysis, and these models were subjected to further performance analyses on the unseen testing set.

4. RESULTS

Performance evaluations were conducted for all approaches using the specifically designed datasets. Each model, under every approach, was trained and evaluated three times using distinct random seed numbers (1337, 42, and 2024). These seeds helped partition the dataset into training, validation, and testing sets. The performance metrics consisted of precision, F1 score, accuracy, sensitivity, and specificity, each defined in Equations 5 through 9. The performance evaluation results were analyzed using two-tailed paired t-tests and single-factor analyses [19].

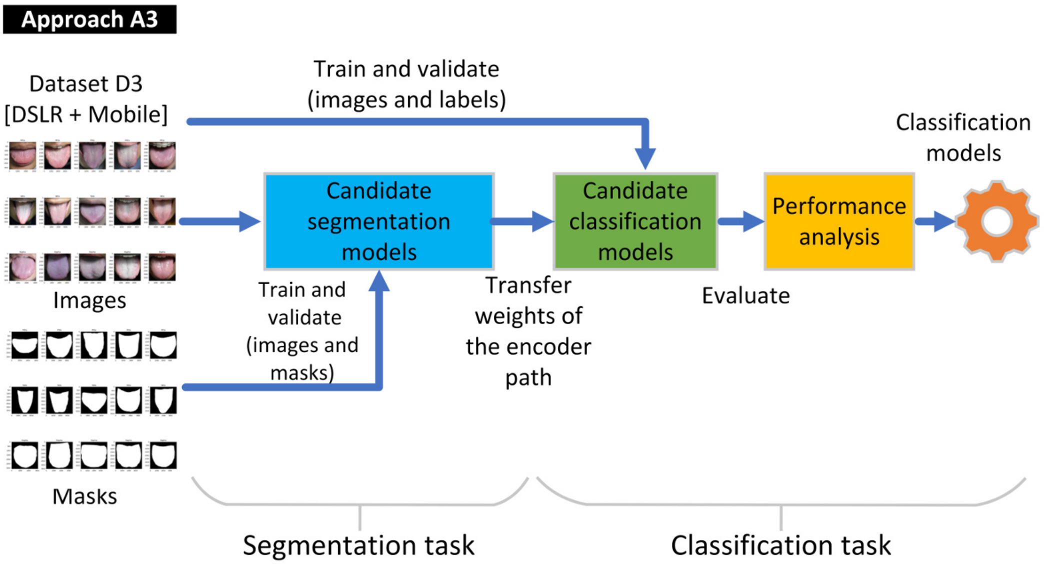

Fig.4: *Tongue image analysis for Approach A3 consists of segmentation and classification tasks. Initially, candidate U-Net-based segmentation models are employed using images and their respective ground truth ROI masks. In the subsequent classification task, the encoder path of the segmentation model is used as a backbone to construct the architecture for the classification task.*

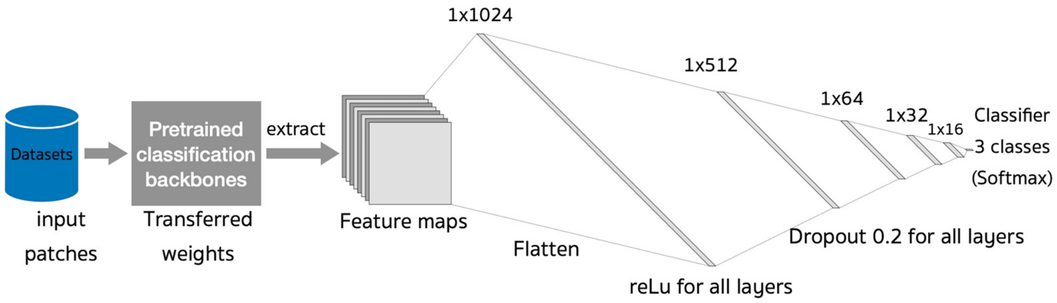

Fig.5: *Classification architecture with configurations utilized for all proposed approaches.*

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

$$F1_{score} = \frac{2TP}{2TP + FP + FN} \quad (6)$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (8)$$

$$specificity = \frac{TN}{TN + FP} \quad (9)$$

4.1 Model Evaluations of Approach A1

For Approach A1, Performance evaluation comparisons were conducted across the D1 and D2 datasets for all pretrained models, focusing on sensitivity. The DenseNet121 model using the D2-Mobile dataset achieved the highest sensitivity at 0.85, followed by the MobileNetV2 model on the same dataset (0.84) and the InceptionResNetV2 model on the D2-DSLR dataset (0.83).

Focusing on specificity performance across the D1 and D2 datasets for all the pretrained models, The DenseNet121 and MobileNetV2 models outperformed others on the D2-Mobile dataset with a specificity of 0.92, followed by InceptionV3 and MobileNet (0.91) and Xception (0.90). For the D1-DSLR dataset, the DenseNet121 and Xception models performed best, each achieving a specificity of 0.91.

Regarding the F1 score, the performance trends mirrored those observed for sensitivity and specificity. The D2-Mobile dataset showed the highest F1 score of 0.85, while the D1-DSLR, D1-Mobile, and D2-DSLR datasets each achieved a maximum F1 score of 0.82, produced by DenseNet121

However, all scores of the trained models and datasets were statistically analyzed using a two-tailed paired t-test (with an alpha threshold of 0.05). No significant differences were found among these met-

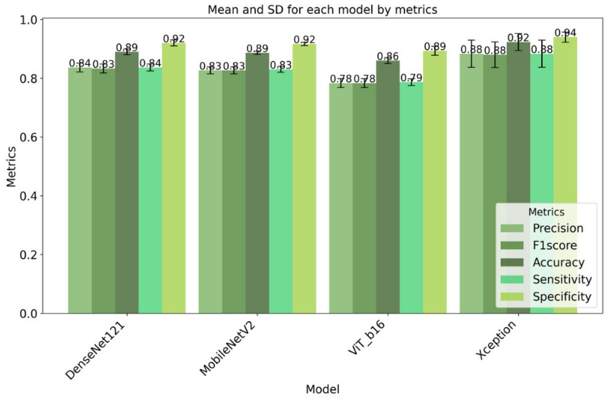

Fig.6: *Mean sensitivity and specificity with standard deviations (SD) for selected models from test sets under varying random seed numbers in Approach A2 using Dataset D1 (A2-D1). The Xception model achieved the highest mean sensitivity and specificity, 0.88 and 0.94, respectively, surpassing those of the DenseNet121, MobileNetV2, and ViT_b16 models.*

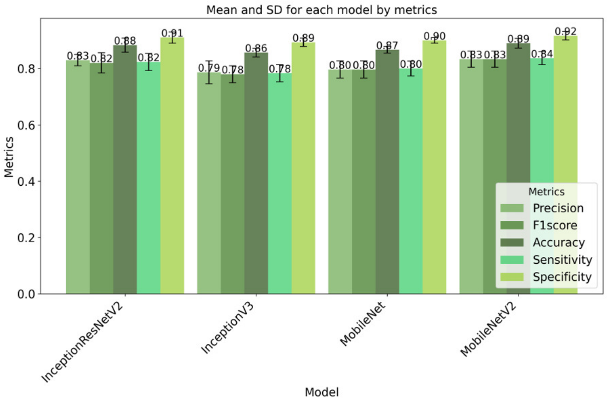

Fig.7: *Mean sensitivity and specificity with standard deviations (SD) for selected models from test sets under varying random seed numbers in Approach A2 using Dataset D2 (A2-D2). The MobileNetV2 model led to the best mean sensitivity and specificity of 0.84 and 0.92, respectively, followed by the InceptionResNetV2, MobileNet, and InceptionV3 models.*

rics across the designed datasets, as all p-values exceeded the alpha threshold.

Considering real-world applications where TTM practitioners may not need to isolate the tongue ROI, we applied the TL technique from Approach A2. This approach commenced with first-stage training using models trained on the D1-DSLR dataset due to its high pixel quality and continued with second-stage training using the D1-Mobile dataset.

4.2 Model Evaluations of Approach A2

In Approach A2, the top performances of the four models trained with Dataset D1 (referred to as experiments A2-D1) are depicted in Figure 6. The Xception model achieved the highest average sensitivity and specificity, at 0.88 and 0.94, respectively, outperforming the DenseNet121, MobileNetV2, and ViT_b16 models.

For experiments A2-D2 with the same approach, the MobileNetV2 model exhibited the best average sensitivity and specificity (0.84 and 0.92, respectively). This performance was followed by the InceptionResNetV2, MobileNet, and InceptionV3 models, as illustrated in Figure 7.

After comparing the outcomes of both experiments A2-D1 and A2-D2, we selected the best-performing model from each and performed a two-tailed paired t-test to compare performance across all metrics. The t-test results revealed significant differences, with a p-value of 0.0012 (alpha set at 0.05), highlighting the superior performance of the Xception model from experiments A2-D1, for which the average sensitivity and specificity were 0.88 and 0.94, respectively.

4.3 Model Evaluations of Approach A3

In Approach A3, we initially utilized our previously developed segmentation model [44] to build a feature extractor for segmentation tasks. After successful training on the D3-DSLR dataset (images and masks), the encoder path with its trained weights was employed as the feature extractor for the classification task. We analyzed the model architecture to identify the last convolutional layer, marking the end of the encoder path. The intermediate outputs of the last convolutional layer (conv2d_155) demonstrated the effective focus of the encoder path, supported by segmentation task performance evaluations showing a Dice coefficient of 0.8524 and precision, sensitivity, and accuracy of 0.9806.

However, when we utilized the encoder path for the classification task, its performance significantly decreased. Consequently, the encoder path of the segmentation model was deemed suboptimal for this task.

Considering alternative models, such as HuggingFace Transformers [45], performance evaluation results from the ViT-based model (ViT_b16) in Approach A2 indicated that further training of HuggingFace Transformers is unnecessary due to their construction on a foundation of transformer blocks similar to those in the ViT-based model.

Regarding pretrained segmentation models [46], the performance results from traditional pretrained models in Approach A2 provided sufficient evidence to conclude that further training is unnecessary, as segmentation models are built using the backbones of these traditional pretrained models.

4.4 Model Selection Method

Based on their performance across all metrics, we selected the three best models for Dataset D1: A2-D1-Xception, A2-D1-DenseNet121, and A1-D1-DSLR-Xception. Single-factor ANOVA revealed a p-value of 0.0108 (F-critical = 3.48, at $p < 0.05$), indicating significant differences in the performances of these models. The A1-D1-DSLR-Xception model, trained solely on DSLR-taken images, was found to lack robustness and was therefore not selected. Conversely, the two best models from Approach A2 (TL), A2-D1-Xception and A2-D1-DenseNet121 were chosen for deployment.

For Dataset D2, MobileNetV2 with the A1-D2-Mobile experimental setting achieved the best performance. However, this model was not selected because its training data, taken exclusively from a mobile phone camera, typically lack robustness across various data sources. Additionally, its reliance on tongue ROI images would necessitate labor-intensive preprocessing steps for deployment. Nevertheless, the encoder of the segmentation model from Approach A3 was also not chosen for application due to its poor performance across all the metrics.

We summarize the performance scores of the proposed approaches in Figure 8. Among the four models, the DenseNet121 and Xception models from the A2-D1 experiment were selected for application based on a two-tailed paired t-test that yielded a p-value of 0.0012 (t-Stat = -8.28, at $p < 0.05$), indicating significant differences among them.

Additionally, we conducted model ensemble (ME) performance evaluations for these chosen models. To thoroughly assess the ME, we compared the A2-D1-DenseNet121 and A2-D1-Xception models, along with their ensembles that included both DSLR-taken (DSLR-ME) and mobile-taken images (Mobile-ME). We used unique test sets, generated with different seed numbers, across three evaluation rounds, as depicted in Figure 9.

Notably, the DSLR-ME evaluations yielded the highest average predictions, achieving impressive scores: precision of 0.94, F1 score of 0.96, accuracy of 0.96, sensitivity of 0.96, and specificity of 0.97. A subsequent single-factor ANOVA reported a p-value of 0.0003 (F-critical = 3.24, at $p < 0.05$), which confirmed statistically significant differences among the various evaluation methods. Based on these results, we selected the ME evaluation method for inference on the unseen testing set.

4.5 Model Ensemble Evaluations of a Testing Set

Figure 10 presents two confusion matrices from the ME evaluations using a test set (seed 1337) from the D1-DSLR dataset. These matrices are shown in both absolute numbers of predicted cases and normalized forms. The dataset, comprising Tri-Dhat tongue images, initially had imbalanced classes (as shown in Figure 1), with Pitta being the most prevalent at 175 cases, followed by Kapha at 87, and Vata at 22. This imbalance was addressed by data augmentation, balancing each class to 175 cases. The test set included 48 Pitta, 64 Kapha, and 46 Vata images.

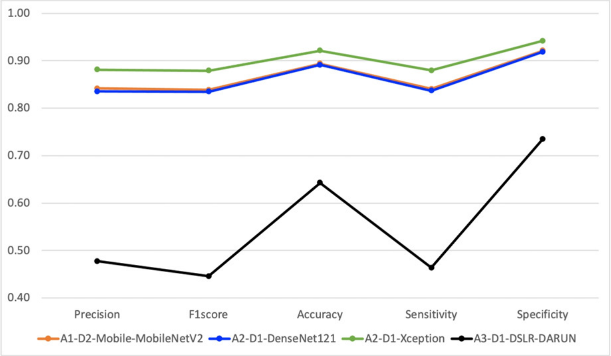

Fig.8: *A comparison of approaches A1, A2, and A3 highlights that the A2-D1-DenseNet121 and A2-D1-Xception models were selected for application. This decision was based on a two-tailed t test, which yielded a significant p-value of 0.0012 (t-stat = -8.28, p <* 0.05*).*

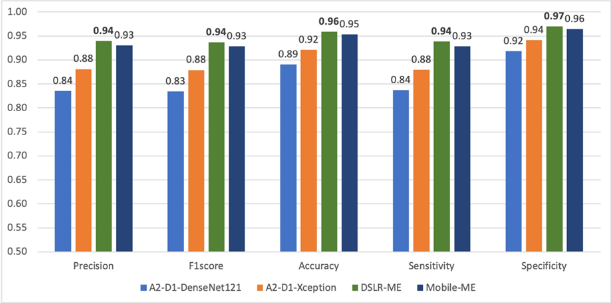

Fig.9: *Performance evaluation comparisons for the A2-D1-DenseNet121 and A2-D1-Xception models, along with their model ensemble (ME), using both DSLR-taken and mobile-taken images (denoted as DSLR-ME and Mobile-ME, respectively). The text in bold highlights the highest average values achieved by the MEs with DSLR image predictions.*

Our ME correctly predicted 41 cases for Pitta, with seven cases misclassified, resulting in a sensitivity of 0.85 for Pitta. For the other classes, the ME correctly predicted all cases for Vata and Kapha, each achieving a sensitivity of 1.00. Overall, the ME achieved an average sensitivity of 0.95 across all classes. Remarkably, even without data augmentation particularly through the omission of the RandomOverSampler algorithm. Pitta's sensitivity of 0.85 was notable as an unbiased and satisfactory score, reflecting the robustness of our designed datasets and experimental methodologies.

Furthermore, we are the first team to analyze

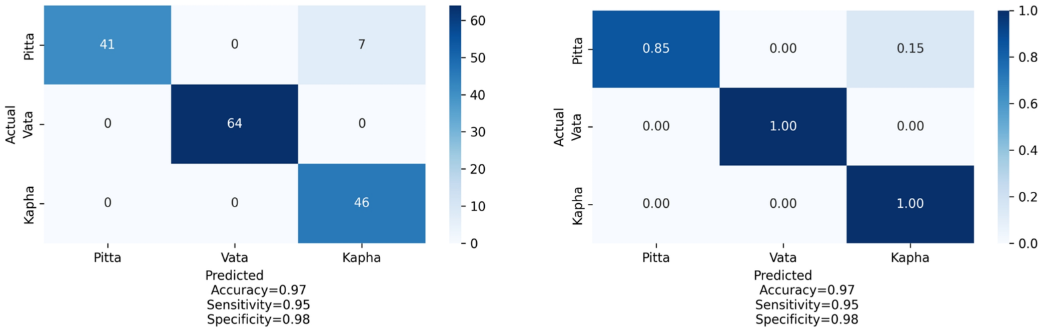

Fig.10: *Two confusion matrices of model ensemble evaluations using the test set (seed 1337) for both visualization styles, including numbers of predicted cases and normalization*

Table 5: *Performance comparisons with related works*

Author	Method	Performance
Joshi *et al.* (2020) [2]	KNN	0.83 sensitivity
Wang *et al.* (2020) [12]	ResNet34	0.90 accuracy
Lu *et al.* (2024) [18]	DenseNet-201	0.82 sensitivity
Our proposed	Transfer learning of DenseNet121, Xception, and their model ensemble	0.94 precision, 0.96 sensitivity, 0.96 F1 score

tongue images using AI based on TTM knowledge. As a result, we compared our results with indirectly related works, especially those related to the classification of tongue images, as presented in Table 5. Our ME method surpassed the sensitivity of Joshi *et al.* [2] and the accuracy of Wang *et al.* [12].

5. DISCUSSION

5.1 Ablation Analysis

Since we utilized only the *RandomOverSampler* algorithm for the D1, D2, and D3 datasets, increasing their minority classes (Vata and Kapha) to match the majority class, Pitta, in this section, we created another balanced dataset for D1 using both the *RandomUnderSampler* and *RandomOverSampler* algorithms. With an average class size of 94.67, we selected 100 images for each class. First, we reduced the Pitta class from 175 to 100. Then, we increased the Vata class from 22 and the Kapha class from 87 to 100 each.

The DenseNet121 and Xception models were trained and evaluated three times using different random seed numbers. The performance comparisons revealed the average metric scores obtained from the D1-175 dataset surpassed those from the D1-100 dataset for both models. Additionally, two-tailed paired t-tests revealed that the model performances for D1-100 dropped significantly, with p-values of 0.0010 and 0.0015 (alpha set at 0.05). This indicates that dataset organization is a crucial factor affecting model performance.

5.2 Dataset Limitations

The subjects or cases involved imbalanced classes and exhibited a lower kappa agreement score of 0.30, which falls at the lower end of the fair agreement category. This suggests a need to review the classification criteria or provide additional training for TTM practitioners to ensure a higher level of consistency in their evaluations. In addition, we suggest increasing the number of subjects from multiple clinical sites to cover all age ranges for handling the degree of data imbalance.

The source devices used in this study were a DSLR camera and a mobile phone camera, with files saved in JPEG and HEIC formats, respectively. These formats are highly compressed and result in the loss of pixel information. We recommend capturing images in a raw pixel data format and then processing them into less compressed file types such as TIFF or PNG files to better analyze their performance.

The Tri-Dhat classification of individuals can vary periodically depending on their physical and mental conditions. We recommend periodic and follow-up data collection in future studies to accommodate these variations.

5.3 Methodological Limitations

The proposed Approach A2 represents an optimal TL technique given our current constraints, paving the way for new and complex analytical techniques in the future. However, for real-world deployment and application, we recommend using a DSLR camera as the primary source and a mobile phone camera as a secondary option.

Based on the literature review, another model from the DenseNet family, DenseNet201, was also trained and evaluated. Its performance was close to that of the DenseNet121 model, catching our attention for potential optimization of this model family in future work.

Raw pixels of tongue images, along with data augmentation, were used to train the models in this work. We also have ground truth for other physical conditions of these images, such as color, shape, moisture, coating, cracks, and teeth marks, which will be incorporated to support the Tri-Dhat classification in future work.

Only the RGB color space of the raw pixels was used in this work. Analyzing tongue images using various color spaces and chromatic features, as well as edge and contour detection techniques, is superior for potential feature engineering and selection [11]. Additionally, edge and contour detection techniques aligning with state-of-the-art DL and traditional machine learning approaches such as DeepLabV3+ [25], UNet++ [27], SegAN [28], and XGBoost [21] have attracted our attention for further analysis which can be forwarded to disease classification in future work [47].

6. CONCLUSION

In this work, we focused on tongue image analysis using raw pixels and AI to support TTM diagnoses, specifically targeting the Tri-Dhat classification: Vata, Pitta, and Kapha. We collected 572 tongue images from subjects at our university's TTM hospital using both DSLR and mobile phone cameras in a calibrated environment. These images were cropped, assessed, and labeled for Tri-Dhat physical conditions by three experienced TTM practitioners, with interobserver variability measured by Fleiss' kappa, resulting in a kappa value of 0.30, indicating fair agreement. However, a consensus-based approach was used for the final labels.

We utilized these labeled images to design and create three distinct datasets to support our three proposed analysis approaches. We addressed issues of class imbalance and data augmentation in these datasets.

Approach A1 focused on classification tasks, Approach A2 integrated classification with transfer learning techniques, and Approach A3 combined segmentation with classification tasks. Various pretrained models were trained for these approaches, and their performances were evaluated using two-tailed paired t-tests and single-factor ANOVA. The null hypothesis for the ANOVA was that there was no significant difference in performance across the various approaches, while the alternative hypothesis was that a significant difference did exist.

Approach A2 provided the most significant results, with the DenseNet121 and Xception models demonstrating exceptional performance with the D1 dataset, which featured cropped images. ME evaluations for these models incorporated both DSLR-taken and mobile-taken images. Notably, the ME evaluations that utilized DSLR-taken images yielded the highest average predictions, achieving a precision of 0.94, an F1 score of 0.96, an accuracy of 0.96, a sensitivity of 0.96, and a specificity of 0.97. The ANOVA result, with a p-value of 0.0003, indicated that we could reject the null hypothesis, confirming a statistically significant difference in performance among the approaches.

Notably, even without data augmentation, the unbiased sensitivity of 0.85 for Pitta was considered satisfactory, reflecting the robustness of our designed datasets and experimental methodologies. We suggest that our methods could be effectively deployed in real-world scenarios to aid TTM practitioners in their diagnoses.

DATA AND CODE AVAILABILITY

The dataset used during the current study for publicly accessible are available at `https://dx.doi.org/10.21227/56cx-0f96`. Code implemented using Tensorflow [48], Keras [49], and ViT_Keras [50] in Python is available at `https://github.com/kasikrit/Tonge-Image-Classification/`.

ACKNOWLEDGEMENT

This research was partially supported by the Traditional Thai Medicine Faculty Research Fund, Prince of Songkla University (Grant No. TTM6411/2566).

AUTHOR CONTRIBUTIONS

Conceptualization, K.D.; Data Curation, T.S., K.M., P.J, N.S., T.Y., S.A., A.I., and A.L.; Formal Analysis, K.D.; Funding Acquisition, T.S.; Investigation, K.D.; Methodology, K.D.; Project Administration; K.D. and T.S.; Resources, J.C.; Software, K.D.; Supervision, K.D. and T.S.; Validation, K.D. and T.S.; Visualization, K.D.; Writing – Original Draft Preparation, K.D. and T.S.; Writing – Review & Editing, K.D. and T.S.; All authors reviewed the manuscript.

References

[1] Z. Tian, D. Wang, X. Sun, Y. Fan, Y. Guan, N. Zhang, M. Zhou, X. Zeng, Y. Yuan, H. Bu, and H. Wang, "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, Feb 2023.

[2] M. S. Joshi, V. Umadevi, and A. B. N. Raj, "Computerized pragmatic assessment of Prakriti Dosha using tongue images- Pilot study," *Indian Journal of Science and Technology*, vol. 13, no. 48, pp. 4679-4698, 2020. [Online].

Available: https://doi.org/10.17485/IJST/v13i48.1626.

[3] A. K. Sharma, *Diagnostic Methods in Ayurveda*. India: MLBD, 2008.

[4] A. R. Jung, H.-Y. Lee, and M.-S. Hwang, "A Comparative Study on the Tongue Diagnosis between Korean medicine and Ayurveda," *The Journal of Korean Medicine*, vol. 40, no. 2, pp. 63-71, 2019, published online: June 30, 2019. [Online]. Available: https://doi.org/10.13048/jkm.19017.

[5] W. S. Kacera, *Ayurvedic Tongue Diagnosis*. Delhi, India: Motilal Banarsidass, 2007.

[6] J. K. Anastasi, L. M. Currie, and G. H. Kim, "Understanding diagnostic reasoning in TCM practice: tongue diagnosis," *Altern Ther Health Med*, vol. 15, no. 3, pp. 18-28, May-Jun 2009.

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

[8] P. P. Prasatwet, *Medical Studies: General Medicine*, vol. 1. Bangkok: Samakkhi, 1986.

[9] P. Sapcharoen, *Compilation of Theory of Thai Traditional Medicine, Volume 1: The Canon of Diagnosis*, vol. 1. Bangkok: Institute of Thai Traditional Medicine, 2537.

[10] D. Miryala, P. Parvataneni, and G. Aliperi, "Computer aided image enhancement of tongue for diagnosis in ayurvedic medical treatment," *Applied Medical Informatics*, vol. 34, no. 1, pp. 46-56, 2014.

[11] 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, Mar 2019.

[12] X. Wang, J. Liu, C. Wu, J. Liu, Q. Li, Y. Chen, X. Wang, X. Chen, X. Pang, B. Chang, J. Lin, S. Zhao, Z. Li, Q. Deng, Y. Lu, D. Zhao, and J. Chen, "Artificial intelligence in tongue diagnosis: Using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 973-980, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037020300325.

[13] L. Ding, X. you Zhang, D. yao Wu, and M. ling Liu, "Application of an extreme learning machine network with particle swarm optimization in syndrome classification of primary liver cancer," *Journal of Integrative Medicine*, vol. 19, no. 5, pp. 395-407, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095496421000650.

[14] Y. Xiang, L. Shujin, C. Hongfang, W. Yinping, Y. Dawei, D. Zhou, and L. Zhiqing, "Artificial intelligence-based diagnosis of diabetes mellitus: Combining fundus photography with traditional chinese medicine diagnostic methodology," *BioMed Research International*, vol. 2021, p. 5556057, 2021. [Online]. Available: https://doi.org/10.1155/2021/5556057.

[15] J. Li, P. Yuan, X. Hu, J. Huang, L. Cui, J. Cui, X. Ma, T. Jiang, X. Yao, J. Li, Y. Shi, Z. Bi, Y. Wang, H. Fu, J. Wang, Y. Lin, C. Pai, X. Guo, C. Zhou, L. Tu, and J. Xu, "A tongue features fusion approach to predicting prediabetes and diabetes with machine learning," *Journal of Biomedical Informatics*, vol. 115, p. 103693, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046421000228.

[16] J. Li, J. Huang, T. Jiang, L. Tu, L. Cui, J. Cui, X. Ma, X. Yao, Y. Shi, S. Wang, Y. Wang, J. Liu, Y. Li, C. Zhou, X. Hu, and J. Xu, "A multi-step approach for tongue image classification in patients with diabetes," *Computers in Biology and Medicine*, vol. 149, p. 105935, 2022.

[17] Y. Shi, H. Wang, X. Yao, J. Li, J. Liu, Y. Chen, L. Liu, and J. Xu, "Machine learning prediction models for different stages of non-small cell lung cancer based on tongue and tumor marker: a pilot study," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 197, 2023. [Online]. Available: https://doi.org/10.1186/s12911-023-02266-5.

[18] X. Lu, H. Hu, W. Li, J. Deng, L. Chen, M. Cheng, H. Huang, W. Ke, W. Wang, and B. Sun, "Exploring hepatic fibrosis screening via deep learning analysis of tongue images," *Journal of Traditional and Complementary Medicine*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2225411024000294.

[19] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and Techniques*, 3rd ed. MA, USA: Morgan Kaufmann, 2012.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.03385, 2015.

[21] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *CoRR*, vol. abs/1603.02754, 2016. [Online]. Available: http://arxiv.org/abs/1603.02754.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *The Ninth International Conference on Learning Representations*, Vienna, Austria, 4 May 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[23] J. Yan, J. Cai, Z. Xu, R. Guo, W. Zhou, H. Yan, Z. Xu, and Y. Wang, "Tongue crack recognition using segmentation based deep learn-

ing," *Scientific Reports*, vol. 13, no. 1, p. 511, 2023. [Online]. Available: `https://doi.org/10.1038/s41598-022-27210-x`.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980-2988, Oct 2017.

[25] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: `http://arxiv.org/abs/1802.02611`.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, pp. 234-241, 2015.

[27] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856-1867, June 2020.

[28] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "SegAN: Adversarial Network with Multi-scale L1 Loss for Medical Image Segmentation," *Neuroinformatics*, vol. 16, no. 3, pp. 383-392, 2018. [Online]. Available:`https://doi.org/10.1007/s12021-018-9377-x`.

[29] L. Feng, W. Xiao, C. Wen, Q. Deng, J. Guo, and H. Song, "Objectification of Tongue Diagnosis in Traditional Medicine, Data Analysis, and Study Application," *J Vis Exp*, no. 194, Apr 2023.

[30] T. Yamane, *Statistics : an introductory analysis*, 2nd ed. New York: Harper and Row, 1967.

[31] B. Prapaspong, S. Suwanphokhin, and U. Chaiklang, *Royal Medical Literature: Medical Knowledge and Literary Heritage of the Nation.* Bangkok: The Institute of the Thai Language, Department of Education, Ministry of Education, 1999.

[32] J. L. Fleiss, "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378-382, 1971.

[33] 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. [Online]. Available: `http://jmlr.org/papers/v18/16-365.html`.

[34] K. Damkliang, T. Wongsirichot, and P. Thongsuksai, "TISSUE CLASSIFICATION FOR COLORECTAL CANCER UTILIZING TECHNIQUES OF DEEP LEARNING AND MACHINE LEARNING," *Biomedical Engineering: Applications, Basis and Communications*, vol. 33, no. 03, p. 2150022, 2021. [Online]. Available: `https://doi.org/10.4015/S1016237221500228`.

[35] K. Damkliang, T. Wongsirichot, C. Khongrak, and P. Suwannarat, "An Optimal Deep Learning Approach to BCa Tissue Detection using Case Studies," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 17, no. 1, pp. 60-72, Feb. 2023. [Online]. Available: `https://ph01.tci-thaijo.org/index.php/ecticit/article/view/250441`.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, Miami, FL, USA, pp. 248-255, June 2009.

[37] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *CoRR*, vol. abs/1608.06993, 2016.

[38] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," CoRR, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[39] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," CoRR, vol. abs/1602.07261, 2016.

[40] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: `http://arxiv.org/abs/1704.04861`.

[41] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, vol. abs/1409.1556, pp. 1-14, 2014.

[42] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800-1807, 2017.

[43] K. Zhang, X. Zhang, and F. Ahmad, "Tongue image texture classification based on xception," *9th International Conference on Computing and Pattern Recognition*, pp. 261-266, 2020.

[44] K. Damkliang, P. Thongsuksai, K. Kayasut, T. Wongsirichot, C. Jitsuwan, and T. Boonpipat, "Binary semantic segmentation for detection of prostate adenocarcinoma using an ensemble with attention and residual U-Net architectures," *PeerJ Computer Science*, vol. 9, no. e1767, 2023. [Online]. Available: `https://doi.org/10.7717/peerj-cs.1767`.

[45] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer,

P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “HuggingFace's Transformers: State-of-the-art Natural Language Processing,” https://huggingface.co/docs/transformers/index, 2020.

[46] P. Lakubovskii, “Segmentation Models,” https://github.com/qubvel/segmentation_models, 2019.

[47] K. Damkliang, J. Chumnaul, T. Sudkhaw, T. Yingtawee, and N. Saearma, “Multi-Model Approach for Tongue Image Classification in Traditional Thai Medicine,” *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 21, no. 5, pp. 1-16, 2025. [Online]. Available: https://doi.org/10.3991/ijoe.v21i05.53671.

[48] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems,” https://www.tensorflow.org/, 2015, software available from tensorflow.org.

[49] F. Chollet, J. J. Allaire, S. Pal, A. Gulli, R. Atienza, and S. Keydana, “Keras,” https://keras.io, 2015, [Online; accessed 16-November-2023].

[50] F. Morales, “vit-keras: Implementation of Vision Transformer (ViT) in Keras,” https://github.com/faustomorales/vit-keras, 2022.

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, ORCID: https://orcid.org/0000-0002-5342-7302.

Teerawat Sudkhaw received the B.Sc. degree in Thai Traditional Medicine and the Master of Thai Traditional Medicine degree from the Faculty of Traditional Thai Medicine, Prince of Songkla University (PSU), Hat Yai, Thailand, in 2010 and 2021, respectively. He is currently a Thai Traditional Medicine practitioner at the Thai Traditional 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 (E-mail: teerawat.sud@psu.ac.th).

Thitinan Yingtawee was a senior student in the Information and Communication Technology (ICT) program, Division of Computational Science, Faculty of Science, Prince of Songkla University, Thailand. She received a B.Sc. degree in Information and Communication Technology from Prince of Songkla University in 2024 (E-mail: 6410210470@psu.ac.th).

Nasma Saearma was a senior student in the Information and Communication Technology (ICT) program, Division of Computational Science, Faculty of Science, Prince of Songkla University, Thailand. She received a B.Sc. degree in Information and Communication Technology from Prince of Songkla University in 2024 (E-mail: 6410210163@psu.ac.th).

Kotchakorn Moosigapong received B.Sc. in Thai Traditional Medicine and Master of Thai Traditional Medicine from the Faculty of Traditional Thai Medicine, Prince of Songkla University (PSU), Hat Yai, Thailand, in 2012 and 2016, respectively. She is currently a Traditional Thai Medicine Doctor at the Traditional Thai Medicine Hospital, PSU and a Co-lecturer in Traditional Thai Medicine curriculum about physical examination, Thai scripture and internship. Her research interests include traditional Thai and folk wisdom, biological activity of Thai herbal formulation described in Thai pharmaceutical textbook and clinical study (E-mail: kotchakorn.m@psu.ac.th).

Patcharawalai Jaisamut received a B.Sc. in Thai Traditional Medicine from Prince of Songkla University (PSU) in 2009 and a Ph.D. in Pharmaceutical Science from PSU in 2016. She is currently an Assistant Professor at the Faculty of Traditional Thai Medicine, PSU, Thailand, and also a licensed Thai traditional medicine doctor specializing in disease diagnosis using Thai traditional medical principles. Her research focuses on the development of herbal formulations and oral drug delivery systems. These systems aim to improve the solubility, stability, and efficacy of Thai herbal medicines (E-mail: patcharawalai.j@psu.ac.th).

Julalak Chokpaisarn earned her B.Sc. in Thai Traditional Medicine from Prince of Songkla University (PSU) in 2011 and her Ph.D. in Microbiology from the Faculty of Science, PSU in 2016. She is currently an Assistant Professor at the Faculty of Traditional Thai Medicine, PSU, where she works as both a lecturer and a researcher. She also serves as a traditional Thai doctor at the Traditional Thai Medicine Hospital, PSU, specializing in disease diagnosis and treatment based on Thai traditional medical principles and herbal medicine. Her research focuses on studying and validating the effectiveness of traditional Thai medicine in treating various diseases, such as myofascial pain syndrome and post-COVID syndrome. In addition, she aims to develop innovative herbal medicines that are modern, evidence-based, and effective in disease treatment (E-mail: julalak.c@psu.ac.th).

Sirinat Laman was a senior student in the Bachelor of Thai Traditional Medicine Program, Faculty of Traditional Thai Medicine, Prince of Songkla University (PSU), Thailand. She earned the Bachelor of Thai Traditional Medicine (B.TM.) degree from PSU in 2024 (E-mail: 6411410068@psu.ac.th).

Anintita Intan was a senior student in the Bachelor of Thai Traditional Medicine Program, Faculty of Traditional Thai Medicine, Prince of Songkla University (PSU), Thailand. She earned the Bachelor of Thai Traditional Medicine (B.TM.) degree from PSU in 2024 (E-mail: 6411410077@psu.ac.th).

Anyamanee Ladam was a senior student in the Bachelor of Thai Traditional Medicine Program, Faculty of Traditional Thai Medicine, Prince of Songkla University (PSU), Thailand. She earned the Bachelor of Thai Traditional Medicine (B.TM.) degree from PSU in 2024 (E-mail: 6411410188@psu.ac.th)