Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal CIELAB phenotyping and machine learning classification of five traditional Chinese medicine constitutions in university students

Yuexin Wei
Changchun University of Chinese Medicine
Lu Han
Changchun University of Chinese Medicine
Zhuoran Li
lizr@ccucm.cn.edu

Changchun University of Chinese Medicine
Yinghao Zhu
Changchun University of Chinese Medicine

Article

Keywords: traditional Chinese medicine constitution, facial color, tongue color, CIELAB, machine learning, incremental value

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11278313/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study aimed to characterize differences in facial and tongue CIELAB color distributions across five traditional Chinese medicine constitution categories and to assess changes in internal classification performance after adding facial features to tongue-based models. This single-center cross-sectional study included 779 university students aged 18–25 years. Standardized imaging provided 39 CIELAB features. Five constitution labels were derived from the 60-item Constitution in Chinese Medicine Questionnaire, which was not used as a model input. L2-regularized multinomial logistic regression and CatBoost were developed and evaluated using stratified nested cross-validation with five outer folds repeated three times and five inner folds. Macro-averaged F1 was the primary performance measure. Participant-level bootstrap resampling estimated 95% confidence intervals for performance, between-model differences, and modality increments. After multiplicity correction, 38 of 39 features differed significantly across categories. Macro-averaged F1 was 0.776 (0.739–0.808) for CatBoost and 0.752 (0.715–0.786) for logistic regression; the paired difference was 0.024 (− 0.005–0.053). Adding facial features to 21 combined tongue features yielded paired increments of 0.025 (0.002–0.048) and 0.030 (0.001–0.059), respectively. Tongue features provided most of the classification information, while facial features yielded modest internal performance gains. Combined features may inform constitution phenotyping and multimodal research; independent validation remains necessary.

Introduction

Traditional Chinese medicine (TCM) constitution refers to a set of relatively stable morphological, physiological, and psychological characteristics shaped by the combined influences of inherited predisposition and acquired environmental factors [1]. Differences in constitution are associated with susceptibility to external pathogenic factors and disease predisposition. TCM constitution is commonly assessed through an integrated evaluation using the four traditional diagnostic methods or classified using self-reported constitution questionnaires. Wang Qi and colleagues developed the Constitution in Chinese Medicine Questionnaire (CCMQ), which was incorporated into the Classification and Determination of Traditional Chinese Medicine Constitution issued by the China Association of Chinese Medicine in 2009, becoming an important tool for constitution classification and assessment [2]. Although widely used, the questionnaire relies on participants' recall of their symptoms and health status and may be influenced by differences in understanding, subjective perceptions, and recent health conditions [3]. With constitution labels defined using questionnaires, characterizing the phenotypic features of different constitution categories through standardized, reproducible, and objective quantitative measures has become an important focus of research on the objective assessment of TCM constitution.

Facial and tongue images are important sources of information for visual inspection in traditional Chinese medicine (TCM) and have been widely used in constitution classification research because of their noninvasive nature and ease of acquisition. Meng et al. summarized facial and tongue features used for constitution identification through expert consultation and developed a diagnostic nomogram

for blood stasis constitution, exploring the feasibility of using these features for constitution assessment [4]. Traditional visual inspection obtains diagnostic information by observing facial complexion, tongue color, tongue shape, and tongue coating, and diagnostic judgments depend heavily on practitioners' clinical experience. In addition, facial and tongue color measurements can be readily affected by factors such as lighting conditions, camera color response, and color correction procedures. Standardized image acquisition and color correction are important prerequisites for obtaining comparable features [5]. The CIELAB color space defined by the Commission Internationale de l'Éclairage represents color using lightness (L*), the red–green axis (a*), and the yellow–blue axis (b*) [6] and can be used to compare color distributions and their degree of overlap across populations. However, between-group differences in individual color features do not directly indicate their contribution to multiclass classification. To assess the value of different modalities, the classification performance of different information sources must also be compared in the same group of participants to evaluate the additional contribution of a newly introduced modality.

Previous studies have used tongue image features for traditional Chinese medicine (TCM) constitution classification and identification, improving model performance through ensemble learning or modality fusion. Zhao et al. [7] and Liu et al. [8] investigated constitution identification using tongue image features. Gu et al. further used facial images, tongue images, pulse waveforms, and health information to classify balanced and biased constitutions [9]. Cui et al. combined tongue images, facial images, and infrared thermal images to explore the objective identification of subhealth status using multiple modalities [10]. In disease screening, Wang et al. used a dual-branch deep neural network incorporating tongue and facial images for binary classification of coronary heart disease. They performed single-modality ablation analyses to compare classification performance using facial images, tongue images, and their combined input, finding that the facial–tongue fusion model outperformed the single-modality models [11]. These studies support the use of facial, tongue, and other information for objective classification, although their target labels, input formats, and validation approaches differ. For the classification of five questionnaire-defined primary constitution categories, direct comparisons within the same participant sample under consistent validation conditions are still needed to assess how much additional classification information facial color provides beyond combined tongue color features.

The study population consisted of university students aged 18–25 years. Five constitution categories—qi deficiency, yang deficiency, yin deficiency, phlegm-dampness, and damp-heat—defined using the 2009 version of the 60-item Constitution in Chinese Medicine Questionnaire served as reference labels. Analyses were based on 39 facial and tongue CIELAB features obtained using the same standardized imaging device. Color distributions across constitution categories were first described, followed by an evaluation of internal classification performance. Multinomial logistic regression and CatBoost models were compared before and after the addition of facial features using identical outer cross-validation splits. This comparison aimed to assess the complementary value of facial features beyond existing tongue information and inform subsequent research on TCM constitution assessment integrating facial and tongue information.

Methods

Study design and participants

This single-center cross-sectional study was approved by the Medical Ethics Committee of the Third Affiliated Clinical Hospital of Changchun University of Chinese Medicine (approval no. CZDSFYLL2025XS □□-062). All methods were performed in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki. Written informed consent was obtained from all registered participants, and each participant was enrolled only once.

Recruitment was organized by the university, and students registered voluntarily. Questionnaire completion and facial and tongue image acquisition were conducted in batches in the same practical training room on campus. The inclusion criteria were: (a) university students aged 18–25 years, irrespective of sex; (b) the ability to understand the study and willingness to provide written informed consent; (c) no indication in the recruitment records of acute or chronic disease, or of oral or tongue disorders or facial skin conditions that could affect facial or tongue image acquisition or color interpretation; and (d) the ability to complete the constitution questionnaire and participate in facial and tongue image acquisition. The exclusion criteria were: (a) incomplete questionnaire responses that prevented the calculation of valid transformed constitution scores; (b) failure to obtain facial or tongue images meeting the quality requirements after repeat photography; (c) image segmentation failure in the accompanying software after image acquisition, resulting in failure to output one or more of the 39 CIELAB features; and (d) failure to meet the prespecified criteria for assigning a primary constitution reference label within the five target categories. The latter included a highest transformed score below 40 among the eight biased constitution categories, a tie for the highest score, or a unique highest score corresponding to a constitution outside the five target categories. Each excluded participant was counted once, with the reason for exclusion assigned according to the first applicable criterion.

Following recruitment registration and health information screening, 1,020 participants proceeded to questionnaire completion and image acquisition. After sequential screening based on questionnaire completeness, image usability, feature export quality, and reference label criteria, the final sample comprised 779 participants (Fig. 1). Image acquisition took place from 26 December 2025 to 25 May 2026. A total of 169 participants were excluded because they did not meet the prespecified criteria for the five-class reference labels. The available records did not permit a breakdown of these exclusions by specific reason: a highest transformed score below 40, a tie for the highest score, or a unique highest score corresponding to a constitution outside the five prespecified categories. Therefore, only the total number of exclusions under this criterion is reported.

TCM constitution labels

The constitution questionnaire was administered through the consultation module of the TCM tongue and pulse information acquisition and management system (YM-SA-I; Tianjin Tianzhong Yimai

Technology Development Co., Ltd., Tianjin, China). The assessment followed the Classification and Determination of Traditional Chinese Medicine Constitution (ZYYXH/T 157—2009), issued by the China Association of Chinese Medicine [2]. The accompanying software automatically calculated and exported the raw and transformed scores for each constitution dimension. The exported transformed scores were used directly to assign constitution reference labels. Questionnaire responses and transformed scores were used only to define these labels and were not included as model predictors.

A previous survey of 3,217 university students in Hubei Province found that yang deficiency, yin deficiency, damp-heat, qi deficiency, and phlegm-dampness were among the most frequently identified categories in students with a single biased constitution [12]. These five constitutional categories were therefore selected for this study. The criteria for defining the primary constitution were prespecified using the transformed scores for the biased constitution dimensions of the 2009 version of the Constitution in Chinese Medicine Questionnaire. A biased constitution was designated as the primary constitution if its transformed score was ≥ 40 and was the unique highest score among the eight biased constitution categories. Only participants whose primary constitution was qi deficiency, yang deficiency, yin deficiency, phlegm-dampness, or damp-heat were included. Participants with a tie for the highest transformed score or a primary constitution outside these five categories were excluded. The questionnaire-defined labels were used only as a reference for this study, not as diagnoses made by TCM practitioners or as a clinical gold standard.

Facial and tongue image acquisition

Facial and tongue images were collected using the respective acquisition modules of the YM-SA-I system. The device was equipped with an enclosure that blocked ambient light and a simulated daylight source. The specified illumination parameters were a nominal illuminance of 3600 lx with a tolerance within ± 8%, a correlated color temperature of 5000–5800 K, and a color rendering index > 95. The imaging device had 24 million effective pixels. Image acquisition was carried out jointly by two researchers who had received standardized training. Participants were instructed to avoid strenuous exercise, alcohol, and colored beverages during the 24 h before image acquisition, to refrain from eating for at least 30 min beforehand, and not to apply makeup or lipstick before facial image acquisition. Facial images were acquired first, followed by tongue images, with three images of each type collected for each participant. Images of a given type were retaken only if all three failed to meet the quality requirements. Image quality was evaluated primarily in terms of complete coverage of the target region, image clarity, the degree of motion blur, tongue posture, and exogenous staining. Software image segmentation and the completeness of feature output were checked separately during subsequent feature extraction. After acquisition, the two researchers jointly selected the highest-quality facial and tongue images for each participant while blinded to the constitution categories. Matching image numbers between the selected facial and tongue images were not required.

Following image selection, the accompanying software was used to analyze the facial and tongue images by region and extract CIELAB color features. L*, a*, and b* values were obtained for each of 13

regions, yielding 39 continuous features. The six facial regions—the overall facial region, forehead, nose, left cheek, right cheek, and lip region—provided 18 features. The four tongue body regions—the whole tongue body, left tongue edge, right tongue edge, and tongue tip—provided 12 features, while the three tongue coating regions—the whole tongue coating, central tongue coating, and tongue root coating—provided nine features. The exported feature data were subsequently checked for completeness. If image segmentation failure prevented the software from outputting one or more of the 39 CIELAB features, the participant was excluded, and missing feature values were not imputed. This criterion resulted in the exclusion of 26 participants from the final analysis dataset.

Statistical analysis of CIELAB phenotypes

The distributions of the 39 CIELAB features were described using medians and quartiles. For selected representative entries, the 5th–95th percentile range (P5–P95) was also reported to characterize the empirical distribution range in the study sample. Differences in each feature across the five primary constitution categories were first assessed using the Kruskal–Wallis test. The P values from all 39 tests were adjusted together using the Benjamini–Hochberg procedure [13]. For each feature, each constitution category was subsequently compared with the pooled group comprising the other four categories using a two-sided Mann–Whitney U test, resulting in 195 comparisons. The P values from all 195 comparisons were adjusted together using the same procedure, and signed rank-biserial correlation coefficients were calculated to describe the magnitude and direction of the differences.

To explore differences among feature groups, representative features were selected for each combination of primary constitution category and information source. Within each combination, the feature with the largest absolute rank-biserial correlation coefficient was selected from the set of features with an adjusted q-value < 0.05. This procedure yielded 15 entries for phenotype presentation, involving 10 distinct CIELAB features. Feature selection at this stage was used solely for phenotype presentation and did not affect the selection of model inputs. All 39 prespecified features were retained in the classification models. The stability of the feature rankings was assessed using 2,000 bootstrap resamples, stratified by constitution category, with participants as the sampling unit. Resampling was performed within the fixed candidate sets defined by q < 0.05 in the original analysis. For each candidate feature, the 95% confidence interval for the effect size and the frequency of selection as the top-ranked feature were estimated.

Model development

The 39 CIELAB features from each participant were used as predictors for L2-regularized multinomial logistic regression and CatBoost models [14], with model development and internal testing conducted using repeated stratified nested cross-validation [15, 16]. The outer cross-validation consisted of stratified five-fold cross-validation repeated three times. Within each outer training set, stratified five-fold cross-validation was performed to select hyperparameters based on macro-averaged F1. The

preprocessing pipeline included median imputation and checks for constant features and was fitted only within the corresponding training folds. Standardization parameters for multinomial logistic regression were also estimated within the training folds, while CatBoost was applied without standardization. None of the 39 CIELAB features had missing values in the analysis dataset. The outer held-out data were not used to fit the preprocessing pipeline, select hyperparameters, or train the models. The base random seed was 20260818, and seeds for the outer and inner cross-validation splits and model training were derived from this seed according to the repetition and fold indices.

For hyperparameter selection, multinomial logistic regression used L2 regularization and the lbfgs solver, with a maximum of 10,000 iterations and a convergence tolerance of 1×10^{-4}. Four C values (0.05, 0.2, 1.0, and 5.0) were evaluated with either no class weighting or balanced class weights, yielding eight candidate combinations. For CatBoost, four configurations were evaluated, with tree depth, learning rate, number of iterations, and L2 leaf regularization coefficient set to (4, 0.03, 600, 3), (4, 0.05, 400, 5), (6, 0.03, 600, 5), and (6, 0.05, 400, 7), respectively. Each configuration was evaluated with automatic class weighting either disabled or set to Balanced, yielding eight candidate combinations. For each model, the hyperparameter combination with the highest mean macro-averaged F1 across the inner validation folds was selected.

Each participant received one held-out prediction in each of the three repetitions of the outer cross-validation. The resulting three sets of five-class prediction probabilities were averaged separately for each class, and the class with the highest mean probability was assigned as the final predicted class. Given the unequal sample sizes across the five classes, macro-averaged F1, which gives equal weight to each class, was used as the primary performance metric. Accuracy, balanced accuracy, and macro-averaged precision were calculated from the final predicted classes [17]. Macro-averaged one-versus-rest AUC and the multiclass Brier score were calculated from the averaged prediction probabilities, and one-versus-rest probability calibration curves were plotted for each class [18].

For each participant, the squared differences between the five predicted class probabilities and the corresponding one-hot-encoded true label were summed. The multiclass Brier score was then calculated as the mean of these sums across all participants. Ordinary participant-level bootstrap resampling was performed with 5,000 resamples drawn with replacement, and 95% confidence intervals were estimated using the percentile method. Performance differences between the two algorithms were estimated using paired bootstrap resampling. In each resample, the same participant indices were used for both models, and the difference in each metric was calculated as CatBoost minus multinomial logistic regression. Bootstrap resampling was restricted to recalculating metrics from the saved outer held-out predictions; models were neither retrained nor retuned during resampling.

Modality ablation and incremental value

To quantify the relative contributions of facial, tongue body, and tongue coating features, logistic regression and CatBoost models were developed separately using five CIELAB feature sets: 18 facial

features, 12 tongue body features, nine tongue coating features, 21 combined tongue features, and all 39 features. All feature sets used the same outer cross-validation splits as the main analysis, with hyperparameters selected through inner cross-validation within the corresponding outer training sets. Incremental performance was assessed separately within each algorithm, using the model with all 39 features as the reference. Paired bootstrap resampling of held-out predictions from the same participants was used to estimate differences in macro-averaged F1 between the full model and each reduced feature model, together with their 95% confidence intervals. Particular emphasis was placed on the change in performance after adding facial features to the combined tongue features.

Sensitivity analyses

Age and sex were added to the 39 CIELAB features to form a 41-feature set and assess the influence of demographic variables on classification performance. The 41-feature and 39-feature models used identical participant splits for outer cross-validation, with hyperparameters selected separately through inner cross-validation within each outer training set. Paired bootstrap resampling of held-out predictions for the same participants was used to estimate performance differences between the two feature sets and their 95% confidence intervals.

The strict-label analysis included only participants whose transformed scores for all seven biased constitution categories other than their primary constitution were < 40, without restricting the transformed score for the balanced constitution. This subset comprised 396 participants. Both models used the 39 CIELAB features and were evaluated using nested cross-validation with newly generated folds within the subset. Hyperparameter selection was restricted to the outer training sets.

Learning curve analysis was conducted only for the CatBoost model using the 39 CIELAB features. Within each outer training set, subsets comprising 25%, 50%, 75%, and 100% of participants were obtained through class-stratified sampling. For each of the first three proportions, sampling was repeated three times per fold, yielding 45 runs per proportion; at 100%, one run was performed per fold, yielding 15 runs. Each model was refitted using the hyperparameters selected for the corresponding outer fold in the main analysis, without further tuning. The outer test sets remained unchanged. Test performance at each proportion was reported as the mean and standard deviation. Standard deviations described variability across runs and did not represent 95% confidence intervals.

Robustness to alternative data partitions was assessed through three post hoc internal sensitivity analyses: leave-one-source-group-out analysis across 17 source file groups, leave-one-date-out analysis across 26 calendar dates, and chronological holdout analysis. Calendar dates were parsed from the "Detection time" field in the frozen analysis dataset. Chronological holdout analysis was exploratory. The 60th, 70th, and 80th percentiles of the distribution of participants' recorded dates were used to define three cutoffs, each corresponding to the latest observed date on or before the respective percentile date. Participants recorded on or before each cutoff formed the training set, and those recorded afterward

formed the test set. Participants recorded on the same date were kept together. The cutoff dates and the composition of the training and test sets are provided in Supplementary Table S1.

All three holdout analyses used fixed configurations from the main analysis candidate grid, documented in the analysis metadata before the formal runs. Multinomial logistic regression used C = 0.2 without class weighting. CatBoost used 400 iterations, a tree depth of 4, a learning rate of 0.05, and an L2 leaf regularization coefficient of 5, without class weighting. No hyperparameter search was performed on the held-out test groups. For the source-group and date-based analyses, overall performance was calculated by pooling one held-out prediction per participant across all 779 participants. The 95% confidence intervals were estimated using 5,000 cluster bootstrap resamples, with source file groups and calendar dates as the respective clustering units. Each chronological split was evaluated separately on its test set, with only point estimates reported.

Model interpretation and software specifications

CatBoost's built-in ShapValues functionality was used to assess feature contributions in the model based on 39 CIELAB features [19]. Models were fitted to each outer training set using the outer splits and fold-specific hyperparameters from the main analysis. SHAP values for each of the five output classes were calculated for participants held out from the corresponding fold. For each participant, the signed SHAP values for the same feature and output class were averaged across the three repetitions. Global importance was calculated for each feature by taking the absolute values of these averaged SHAP values and then averaging them across all participants and all five output classes. The averaged signed SHAP values were also summarized separately for each output class to illustrate the direction and distribution of the main feature contributions across participants. SHAP values explain the model's raw scores for each class and do not directly represent changes in predicted probabilities. The SHAP distribution for each output class included all 779 participants, rather than only participants whose reference label belonged to that class.

Analyses were performed using Python 3.14.6 with NumPy 2.4.6, pandas 3.0.3, and SciPy 1.18.0. Model development and internal evaluation were implemented using scikit-learn 1.9.0 and CatBoost 1.2.10. Reporting was guided by the STROBE recommendations [20] for study design, participant selection, and phenotypic analyses, and by TRIPOD + AI [21] for model development, performance evaluation, and internal validation.

Code availability

The custom Python scripts used for statistical analysis, model development, and performance evaluation are available from the corresponding authors upon reasonable request.

Results

Participants and baseline characteristics

Following health information screening, 1,020 participants entered the data collection process (Fig. 1). Sequential screening excluded participants with incomplete questionnaires (n = 27), inadequate image quality after repeat acquisition (n = 19), incomplete CIELAB feature output due to software segmentation failure (n = 26), or failure to meet the reference label criteria (n = 169), leaving 779 participants for analysis. The median age of the included participants was 19 years; 319 (40.9%) were male and 460 (59.1%) were female. The distribution of constitution categories was as follows: qi deficiency, 270 participants (34.7%); yang deficiency, 76 (9.8%); yin deficiency, 117 (15.0%); phlegm-dampness, 80 (10.3%); and damp-heat, 236 (30.3%). Age and sex distributions within each category are presented in Table 1. Of the 779 participants, 383 (49.2%) had a transformed score ≥ 40 for at least one other biased constitution in addition to their primary constitution, while the remaining 396 met the criteria for the strict-label subset. The main analysis, therefore, used mutually exclusive research labels based on primary constitution, while the sample still included participants whose scores met the threshold for more than one biased constitution.

Table 1
Age and sex distributions in the overall study sample and by primary constitution category.

Constitution class	n	Age, years	Male, n (%)	Female, n (%)
Qi deficiency	270	19 (18–19)	112 (41.5)	158 (58.5)
Yang deficiency	76	19 (18–20)	36 (47.4)	40 (52.6)
Yin deficiency	117	19 (18–19)	32 (27.4)	85 (72.6)
Phlegm-dampness	80	19 (18–19)	29 (36.3)	51 (63.8)
Damp-heat	236	19 (18–19)	110 (46.6)	126 (53.4)
Total	779	19 (18–19)	319 (40.9)	460 (59.1)

Note: Age is reported as median (25th–75th percentiles) and sex as n (%). Percentages are calculated within rows and may not sum to 100% because of rounding.

Facial and tongue CIELAB phenotypes across the five primary constitution categories

Across facial and tongue regions, 38 of the 39 CIELAB color features showed statistically significant overall differences among the five constitution categories after adjustment for multiple comparisons. Representative features showed lower b* values in the overall facial region and whole tongue body in the qi deficiency category, higher overall facial L* values in the yang deficiency category, and higher values for certain b* features in the yin deficiency and phlegm-dampness categories. Among the 15 representative entries, whole tongue body L* in the damp-heat category showed the largest absolute effect size relative to the pooled participants from the other four categories, with a rank-biserial correlation of − 0.804 (Fig. 2a). Despite these differences, the distributions of whole tongue body b* and

overall facial L* overlapped across categories (Fig. 2b, c). Medians and quartiles of the representative features within each category are presented in Supplementary Table S4. In the stratified bootstrap analysis, with candidate feature sets fixed at those meeting the significance criterion in the original analysis, the frequencies of selection as the top-ranked feature ranged from 41.9% to 99.6%. This suggests that feature ranking stability varied by constitution category and information source. Corresponding confidence intervals for effect sizes, empirical distribution ranges, and selection frequencies are provided in Supplementary Table S5.

Internal performance of the five-class classification models

The CatBoost model using the 39 CIELAB features had a macro-averaged F1 of 0.776 (95% CI: 0.739–0.808) based on outer held-out predictions, compared with 0.752 (0.715–0.786) for multinomial logistic regression. The paired difference in macro-averaged F1, calculated at the participant level as CatBoost minus multinomial logistic regression, was 0.024 (95% CI: −0.005–0.053). For CatBoost, accuracy was 0.806, macro-averaged one-versus-rest AUC was 0.945, and the multiclass Brier score was 0.286 (Table 2).

In the CatBoost model, the qi deficiency and damp-heat categories showed relatively high F1 scores of 0.841 and 0.835, respectively, whereas the phlegm-dampness category had the lowest F1 score, at 0.699. The yang deficiency and phlegm-dampness categories had sensitivities of 0.645 and 0.625, respectively, both lower than those of the other categories (Fig. 3b). The confusion matrix showed that 13 participants in the phlegm-dampness category and 18 in the yang deficiency category were misclassified as qi deficiency (Fig. 3a). Class-specific AUCs ranged from 0.906 to 0.972. ROC curves are presented in Supplementary Fig. S1, and prediction probability calibration curves are shown in Fig. 3c.

Table 2
Internal classification performance.

Performance metric	Multinomial logistic regression	CatBoost
Macro-F1	0.752 (0.715–0.786)	0.776 (0.739–0.808)
Balanced accuracy	0.724 (0.687–0.761)	0.753 (0.715–0.787)
Accuracy	0.787 (0.757–0.815)	0.806 (0.777–0.833)
Macro-averaged precision	0.802 (0.769–0.835)	0.811 (0.777–0.844)
Macro OvR-AUC	0.924 (0.907–0.940)	0.945 (0.931–0.957)
Multiclass Brier score	0.322 (0.291–0.355)	0.286 (0.257–0.316)

Note: Values are point estimates (95% confidence intervals) based on averaged outer held-out prediction probabilities. Confidence intervals were estimated by participant-level bootstrap. OvR-AUC, one-versus-rest area under the receiver operating characteristic curve.

Modality ablation and the incremental value of facial features

Modality ablation analysis revealed the relative contributions of the different feature sets. Classification performance was lower when facial features were used alone, but these features provided supplementary information beyond the combined tongue features. In CatBoost, macro-averaged F1 was 0.553 with facial features, 0.745 with tongue body features, and 0.751 with combined tongue features. Adding facial features to the combined tongue features increased macro-averaged F1 to 0.776, with a paired increment of 0.025 (95% CI: 0.002–0.048). Macro-averaged F1 also increased for multinomial logistic regression, from 0.722 to 0.752, with a paired increment of 0.030 (95% CI: 0.001–0.059) (Fig. 4a, b; Table 3).

In CatBoost, adding facial features increased F1 point estimates for all five categories, with larger gains for yang deficiency (0.691 to 0.742) and phlegm-dampness (0.662 to 0.699) than for the other three categories (Fig. 4c).

Table 3
Modality ablation and incremental value of facial features.

Input group	Number of features	Multinomial logistic regression	CatBoost
Face (18)	18	0.519 (0.478–0.558)	0.553 (0.513–0.592)
Tongue body (12)	12	0.700 (0.660–0.737)	0.745 (0.708–0.779)
Tongue coating (9)	9	0.585 (0.543–0.623)	0.630 (0.589–0.667)
Combined tongue (21)	21	0.722 (0.683–0.758)	0.751 (0.714–0.786)
Face + tongue (39)	39	0.752 (0.715–0.786)	0.776 (0.739–0.808)
Increment from adding facial features	-	0.030 (0.001–0.059)	0.025 (0.002–0.048)

Note: Values are macro-averaged F1 scores (95% confidence intervals). The final row reports the paired F1 increment from adding facial features to combined tongue features. Models used identical outer splits with separate inner tuning.

Model interpretation and feature importance

SHAP analysis identified whole tongue body b*, a*, and L* as the most important global features, with importance values of 0.214, 0.184, and 0.166, respectively. The importance of right tongue edge b* was 0.163. Among the ten highest-ranked features, overall facial L* was the only facial feature (Fig. 5a).

Class-specific SHAP distributions showed that lower whole tongue body b* values generally contributed positively to the qi deficiency class score, whereas higher values generally contributed positively to the yin deficiency and phlegm-dampness class scores. Higher overall facial L* values mainly contributed positively to the yang deficiency class score, while lower whole tongue body L* values mainly contributed positively to the damp-heat class score (Fig. 5b–f). These patterns describe the model's reliance on color features and the direction of their contributions.

Internal sensitivity analyses

Sensitivity analyses were conducted to assess model performance under different conditions (Supplementary Table S2). With age and sex included, macro-averaged F1 was 0.779 (95% CI: 0.743–0.811) for CatBoost, with a paired difference of 0.003 (− 0.009–0.016) relative to the 39-feature model. For multinomial logistic regression, macro-averaged F1 was 0.751 (0.713–0.785), with a paired difference of − 0.001 (− 0.010–0.008). No clear improvement in internal performance was observed for either model.

A total of 396 participants were included in the strict-label subset: qi deficiency (n = 174), yang deficiency (n = 22), yin deficiency (n = 49), phlegm-dampness (n = 38), and damp-heat (n = 113). After nested cross-validation was repeated within this subset, macro-averaged F1 was 0.734 (0.671–0.787) for CatBoost and 0.731 (0.664–0.785) for multinomial logistic regression. Both sample size and class composition were changed in this analysis; differences from the main analysis therefore cannot be attributed solely to the exclusion of participants meeting the score threshold for additional biased constitutions.

In the CatBoost learning curve analysis, mean macro-averaged F1 on the outer test sets was 0.719, 0.753, 0.761, and 0.774 when 25%, 50%, 75%, and 100% of the participants in each outer training set were used, respectively. Full results are presented in Supplementary Table S3. These values were obtained by averaging the metrics across outer test runs, whereas the main analysis metrics were calculated from each participant's averaged prediction probabilities.

For the source-group and date-based holdout analyses, performance was calculated after pooling held-out predictions for all participants. The three chronological holdout splits were evaluated separately on their respective test sets; results are presented in Supplementary Table S2. These analyses describe performance under different internal partitions of data collected at the same center using the same device. They do not substitute for independent external validation or establish robustness to temporal drift.

Discussion

Between-group differences in color features, internal classification performance, and the contributions of different modalities were evaluated using facial and tongue CIELAB features and questionnaire-defined constitution labels from 779 university students aged 18–25 years. Although most CIELAB features showed overall differences among categories, the distributions of some representative features still

overlapped substantially. In the classification analysis, tongue features provided most of the predictive information. Classification performance was lower with facial features alone, but modest increases in macro-averaged F1 were observed for both algorithms when facial features were added to the combined tongue features.

Incremental value of facial features for tongue-based classification

Previous studies have explored the value of tongue features for constitution classification from different perspectives. Zhao et al. improved classification models through ensemble learning and feature selection [7], while Liu et al. enriched model inputs by combining conventional tongue features with deep features [8]. In the present study, the additional contribution of facial color features was evaluated beyond the combined tongue body and tongue coating color features. With the same participants and outer cross-validation splits, modest performance gains were observed for both algorithms after facial features were added. Performance with facial features alone reflects their standalone classification ability, whereas the change in performance after adding them to the combined tongue features reflects their incremental contribution. These two aspects should therefore be evaluated separately.

In the current dataset, performance with tongue body features was already close to that with combined tongue features, and further gains were obtained after facial features were added. This suggests that color features from different sources provide partially complementary information. However, an increase in macro-averaged F1 alone does not establish practical benefit. The value of adding facial image acquisition to constitution classification therefore requires further assessment, taking into account acquisition time, measurement repeatability, and independent validation results.

For information integration, Zhou et al. used a Transformer to jointly process imaging data, text, and structured clinical information, providing an example of joint modeling across different information sources [22]. In the present study, classification performance was compared across prespecified combinations of facial and tongue color features. With larger samples, future analyses could further compare the performance of different information integration strategies.

Color phenotypes and their relevance to classification

Color phenotype analysis indicated that differences among primary constitution categories should be described by region and color coordinate. For example, representative differences for qi deficiency involved b* in the overall facial region and whole tongue body, whereas higher overall facial L* was observed in the yang deficiency category. Both yin deficiency and phlegm-dampness showed higher whole tongue body b* values, with overlap remaining across categories. This suggests that changes in a single color feature may be shared by several categories, making it difficult to establish one-to-one constitution classification rules from individual color features.

SHAP analysis further showed that whole tongue body color features contributed strongly to the combined model, while overall facial L* was also among the ten highest-ranked features. These findings help explain how the model uses facial and tongue color information. The phenotype and SHAP analyses provide descriptions from different perspectives, but both were based on the same dataset and therefore do not constitute independent validation.

Differences between models and class-specific performance

The paired difference in macro-averaged F1 between CatBoost and multinomial logistic regression was 0.024, but the 95% confidence interval crossed zero. Thus, no clear evidence was obtained that CatBoost was superior to multinomial logistic regression in macro-averaged F1, the primary performance metric. Class-specific performance was also uneven. Sensitivities for yang deficiency and phlegm-dampness were 0.645 and 0.625, respectively, and some participants in these categories were predicted to belong to the qi deficiency category. These results may be related to smaller sample sizes, overlapping color distributions, and the simplification of co-occurring constitution profiles when mutually exclusive primary constitution labels are assigned. Although class-weight settings were compared during inner hyperparameter tuning, the effects of class imbalance cannot be ruled out. The respective contributions of these factors cannot be determined from the current analyses. Zhang et al. used HWmixer-MLP for multi-label constitution recognition using facial and tongue images, providing a methodological reference for modeling co-occurring constitutions [23]. Future studies could include larger samples from the yang deficiency and phlegm-dampness categories and compare the performance of mutually exclusive primary constitution classification with that of multi-label classification.

Study limitations

Different facial and tongue color feature sets were compared in the same participants using identical outer cross-validation splits, allowing the standalone classification performance of facial features to be distinguished from their incremental contribution beyond combined tongue features. Questionnaire data were used only to define reference labels and were not included as model predictors. The main analysis used 39 prespecified CIELAB features, with repeated stratified nested cross-validation and participant-level paired bootstrap resampling providing a common framework for evaluating internal performance and modality increments.

These findings should be interpreted in light of the label definitions, study population, and scope of validation. Constitution reference labels were derived from the questionnaire. Transformed scores were automatically calculated and exported by the accompanying device software and could not be independently recalculated. The sample was restricted to university students aged 18–25 years, and the findings cannot be directly generalized to other age groups or constitution categories not included in the study. Although participants meeting the score threshold for multiple biased constitutions were included,

mutually exclusive primary constitution labels were used in the main analysis. Multi-label recognition of co-occurring constitutions was not assessed.

Color measurement reliability is another limitation. All images were acquired at a single center using the same device, and the consistency of CIELAB features across repeated acquisitions or measurements on different days was not evaluated. Available records did not fully document the device software version, independent calibration using a standard color chart, monitoring of device drift over time, or each participant's adherence to behavioral instructions during the 24 h before acquisition. Consequently, the effects of acquisition procedures and device-related factors on color measurements cannot be fully excluded.

The source-group, date-based, and chronological holdout analyses were post hoc exploratory internal sensitivity analyses using consistent fixed configurations. Although these configurations had been documented in the analysis metadata before the formal holdout runs, the rationale for individual hyperparameter choices and historical trial-run logs were incomplete, preventing full reconstruction of the configuration selection process. Performance confidence intervals were estimated by resampling saved outer held-out predictions without retraining or retuning the models in each resample. They therefore do not fully capture the variability associated with these processes.

The current evidence is limited to internal validation using data from the same center and device; performance in other populations and acquisition settings has not been established. Xue et al. evaluated a multimodal model for the differential diagnosis of dementia across multiple independent cohorts [24], providing an example for future validation designs. Future studies should keep the current label rules and feature definitions fixed and conduct independent external validation in different age groups and across multiple centers and devices to further assess model transportability.

Conclusion

This study demonstrated the feasibility of multimodal fusion for objective classification of TCM constitutions using facial and tongue CIELAB features. Tongue features provided the main classification information, while adding facial features yielded a modest, statistically significant improvement in internal model performance. Given the restriction to university students from a single center, model generalizability requires further validation in independent populations and across different devices. These findings may inform the future development of multimodal systems for intelligent TCM assessment.

Declarations

Data availability

The de-identified participant-level CIELAB feature data and corresponding constitution reference labels analyzed in this study are available from the corresponding authors upon reasonable request.

Funding

This work was supported by the Science and Technology Research Project of the Jilin Provincial Department of Education (grant no. JJKH20261386KJ).

Acknowledgements

Not applicable.

Authors' contributions

Yuexin Wei: Writing – original draft, Visualization, Software.

Lu Han: Validation, Investigation, Formal analysis.

Zhuoran Li: Writing – review and editing, Supervision, Resources.

Yinghao Zhu: Conceptualization, Methodology, Project administration.

Consent for publication

Not applicable.

Additional Information

Competing interests

The authors declare that they have no competing interests.

References

1. Wang, Q. Individualized medicine, health medicine, and constitutional theory in Chinese medicine. *Front. Med.* **6**, 1–7 (2012).
2. China Association of Chinese Medicine. Classification and determination of traditional Chinese medicine constitution (ZYYXH/T 157–2009). *World J. Integr. Tradit West. Med.* **4**, 303–304 (2009). (in Chinese).
3. Lu, T. et al. Valid and convenient questionnaire assessment of Chinese body constitution: item characteristics, reliability, and construct validation. *Patient Prefer Adherence*. **16**, 1875–1884. https://doi.org/10.2147/PPA.S373512 (2022).
4. Meng, Y. et al. Traditional Chinese medicine constitution identification based on objective facial and tongue features: a Delphi study and a diagnostic nomogram for blood stasis constitution. *Evid. Based Complement. Alternat. Med.* 6950529 (2022). (2022). https://doi.org/10.1155/2022/6950529
5. Qi, Z. et al. The classification of tongue colors with standardized acquisition and ICC profile correction in traditional Chinese medicine. *Biomed. Res. Int.* 3510807 (2016). (2016).

https://doi.org/10.1155/2016/3510807

6. Commission Internationale de. l'Éclairage. *Colorimetry*, 4th edn. CIE 015:2018 (CIE, 2018). https://doi.org/10.25039/TR.015.2018
7. Zhao, M., Zhou, H., Wang, J., Liu, Y. & Zhang, X. A new method for identification of traditional Chinese medicine constitution based on tongue features with machine learning. *Technol. Health Care*. **32**, 3393–3408. https://doi.org/10.3233/THC-240128 (2024).
8. Liu, Y. et al. Study on a traditional Chinese medicine constitution recognition model using tongue image characteristics and deep learning: a prospective dual-center investigation. *Chin. Med.* **20**, 84. https://doi.org/10.1186/s13020-025-01126-w (2025).
9. Gu, T. Y., Yan, Z. Z. & Jiang, J. H. Classifying Chinese medicine constitution using multimodal deep-learning model. *Chin. J. Integr. Med.* **30**, 163–170. https://doi.org/10.1007/s11655-022-3541-8 (2024).
10. Cui, D. et al. A multimodal deep learning model for objective identification of sub-health status in traditional Chinese medicine. *Comb. Chem. High. Throughput Screen.* **29**, 2490–2511. https://doi.org/10.2174/0113862073484389260528213445 (2026).
11. Wang, Y. et al. Explainable multimodal deep learning fusion of tongue and facial images for noninvasive coronary heart disease screening. *Sci. Rep.* https://doi.org/10.1038/s41598-026-71871-x (2026).
12. Yao, F. X. et al. Analysis on the relationship between the pattern of five movements and six qi, Chinese medicine constitution and physical health level of 3217 university students in Hubei province. *China J. Traditional Chin. Med. Pharm.* **39**, 1955–1965 (2024). (in Chinese).
13. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. B (Methodol)*. **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).
14. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances Neural Inform. Process. Systems*. **31**, 6638–6648 (2018).
15. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **7**, 91. https://doi.org/10.1186/1471-2105-7-91 (2006).
16. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
17. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag*. **45**, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002 (2009).
18. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. on behalf of the Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230. https://doi.org/10.1186/s12916-019-1466-7 (2019).
19. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67. https://doi.org/10.1038/s42256-019-0138-9 (2020).

20. von Elm, E. et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann. Intern. Med.* **147**, 573–577. https://doi.org/10.7326/0003-4819-147-8-200710160-00010 (2007).
21. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378 (2024). https://doi.org/10.1136/bmj-2023-078378
22. Zhou, H. Y. et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* **7**, 743–755. https://doi.org/10.1038/s41551-023-01045-x (2023).
23. Zhang, M., Wen, G., Yang, P., Wang, C. & Chen, C. Multi-label body constitution recognition via HWmixer-MLP for facial and tongue images. *Expert Syst. Appl.* **269**, 126383. https://doi.org/10.1016/j.eswa.2025.126383 (2025).
24. Xue, C. et al. AI-based differential diagnosis of dementia etiologies on multimodal data. *Nat. Med.* **30**, 2977–2989. https://doi.org/10.1038/s41591-024-03118-z (2024).

Figures

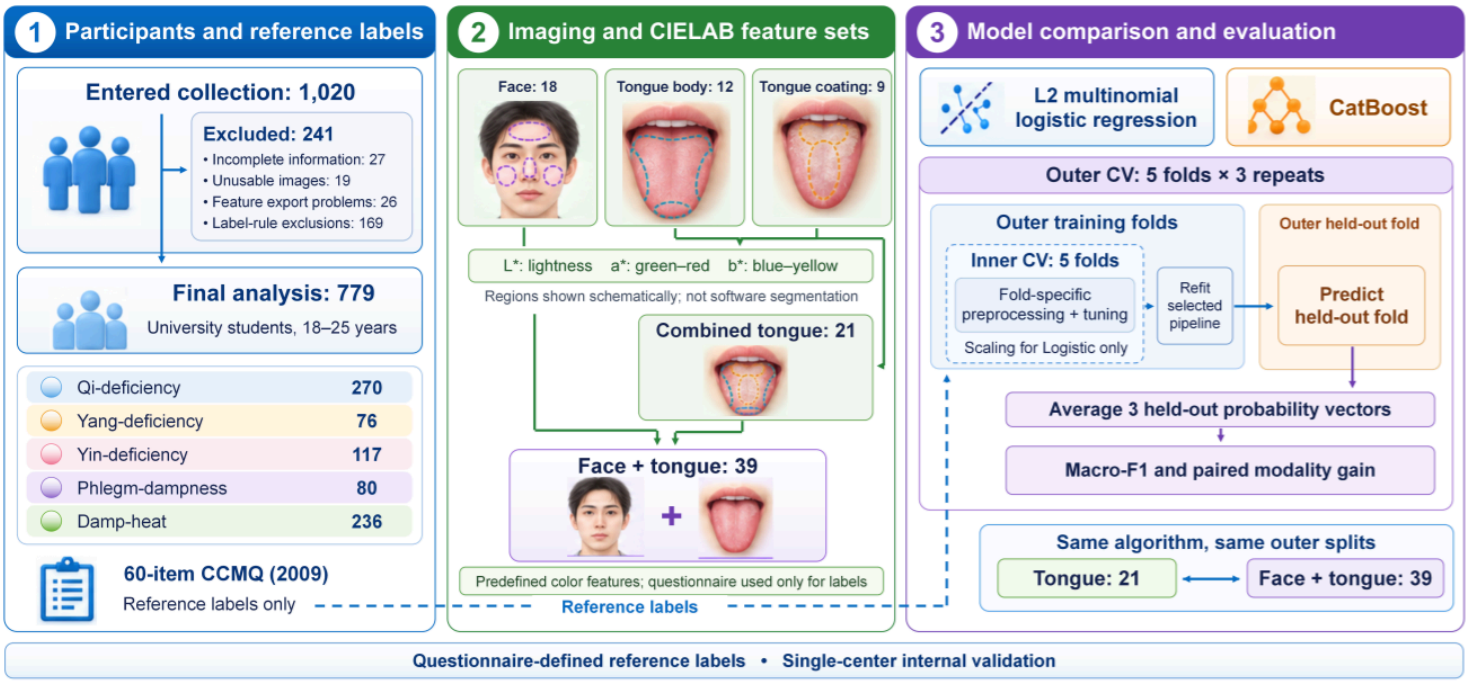


Figure 1

Study workflow.

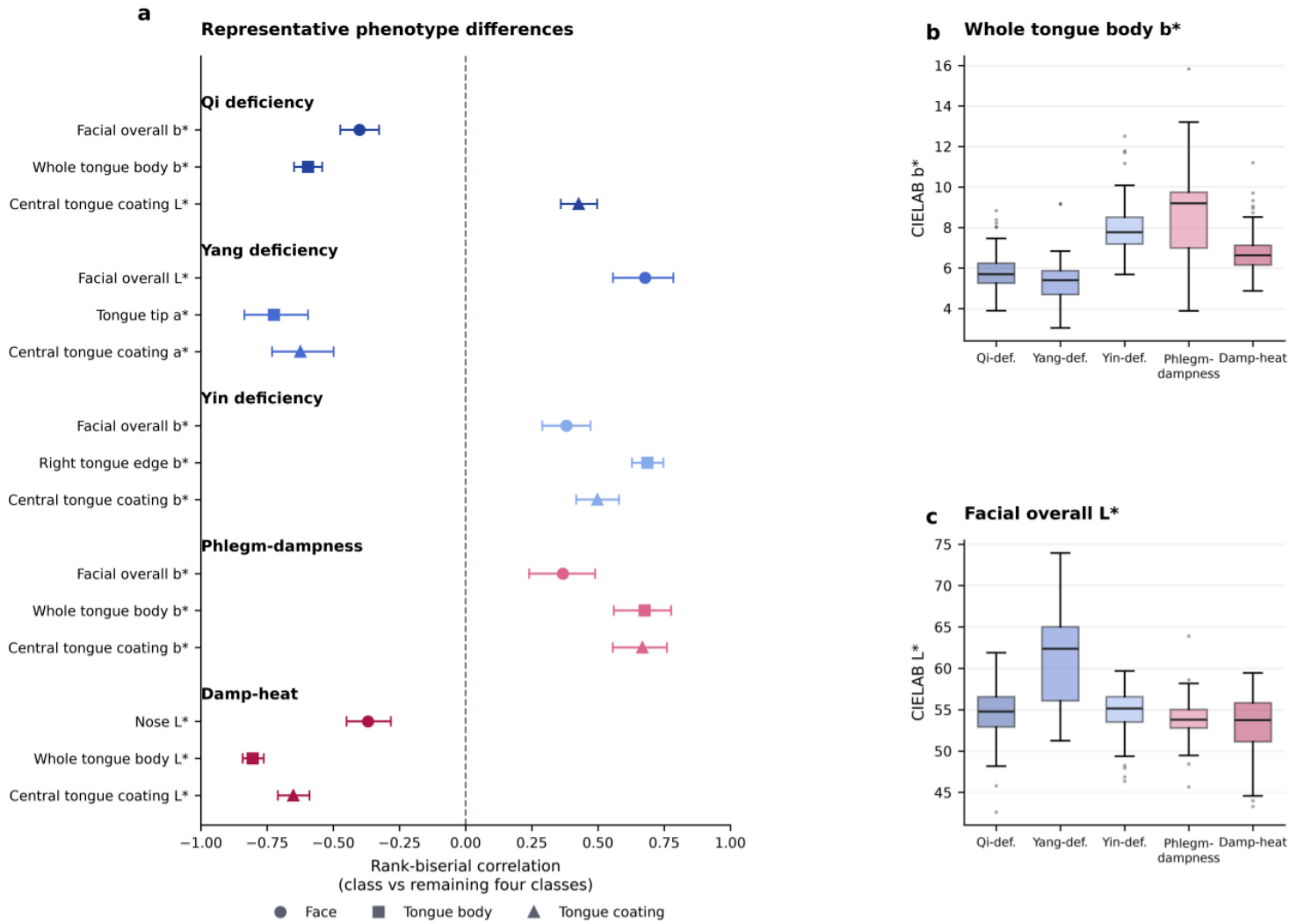


Figure 2

Facial and tongue CIELAB differences across constitution categories.

Note: (a) Rank-biserial correlation coefficients comparing each category with the pooled participants from the other four categories, with 95% confidence intervals estimated from 2,000 stratified participant-level bootstrap resamples. Marker shapes distinguish facial, tongue body, and tongue coating features. Positive values indicate higher feature values in the corresponding category, whereas negative values indicate lower values. (b) Distributions of whole tongue body b* and (c) overall facial L* across the five categories. Boxes span the interquartile range (IQR), and the central lines indicate medians. Whiskers extend to the most extreme observations within 1.5 times the IQR below the first quartile or above the third quartile; points represent observations beyond the whiskers. The asterisks in L*, a*, and b* are part of the color coordinate notation and do not indicate statistical significance.

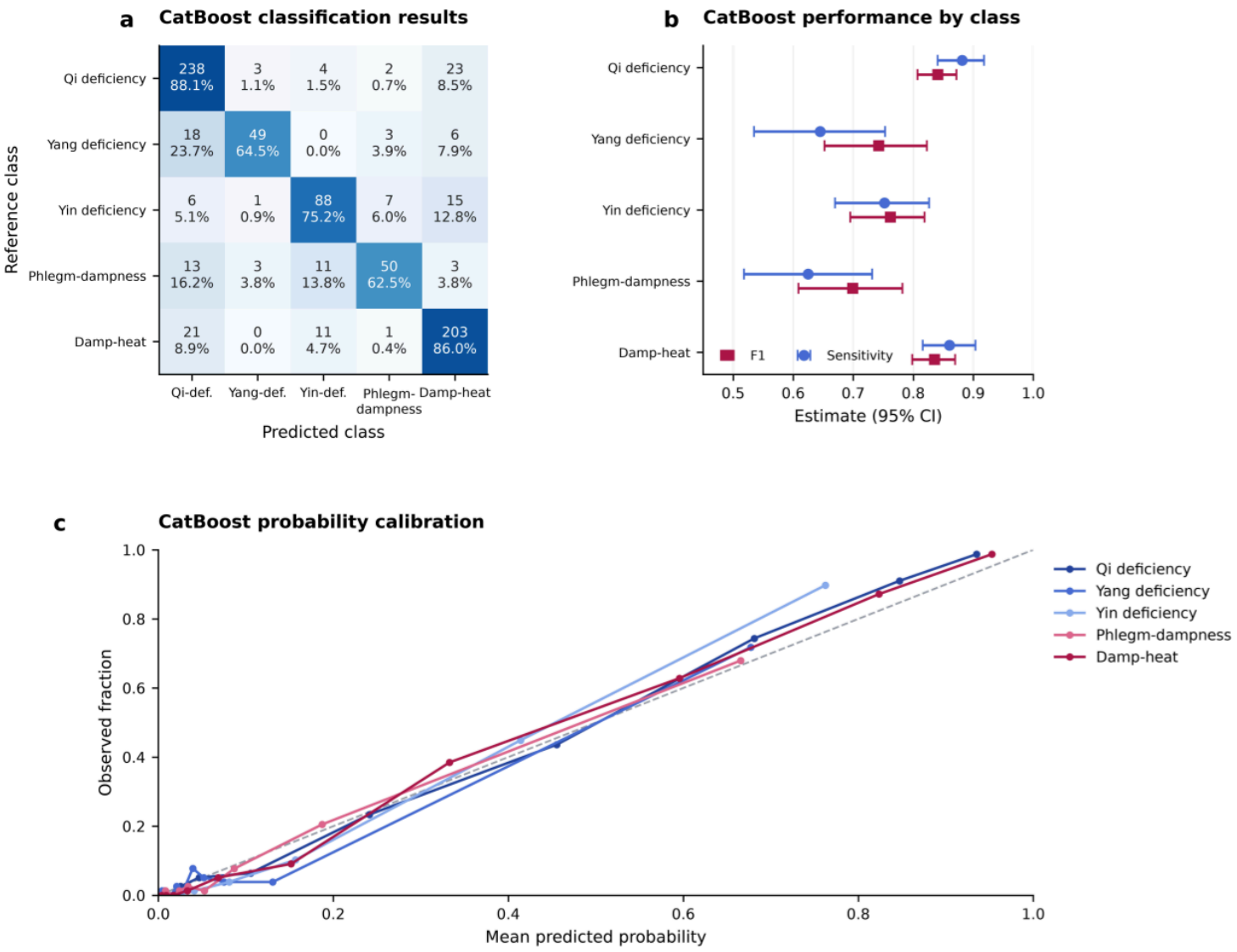


Figure 3

Internal classification performance of CatBoost.

Note: (a) Confusion matrix showing participant counts and row percentages. (b) Class-specific F1 scores and sensitivities with participant-level bootstrap 95% confidence intervals. (c) One-versus-rest calibration curves based on 10 quantile bins of predicted probability; the dashed line indicates ideal calibration. Results use each participant's outer held-out probabilities averaged across three repetitions.

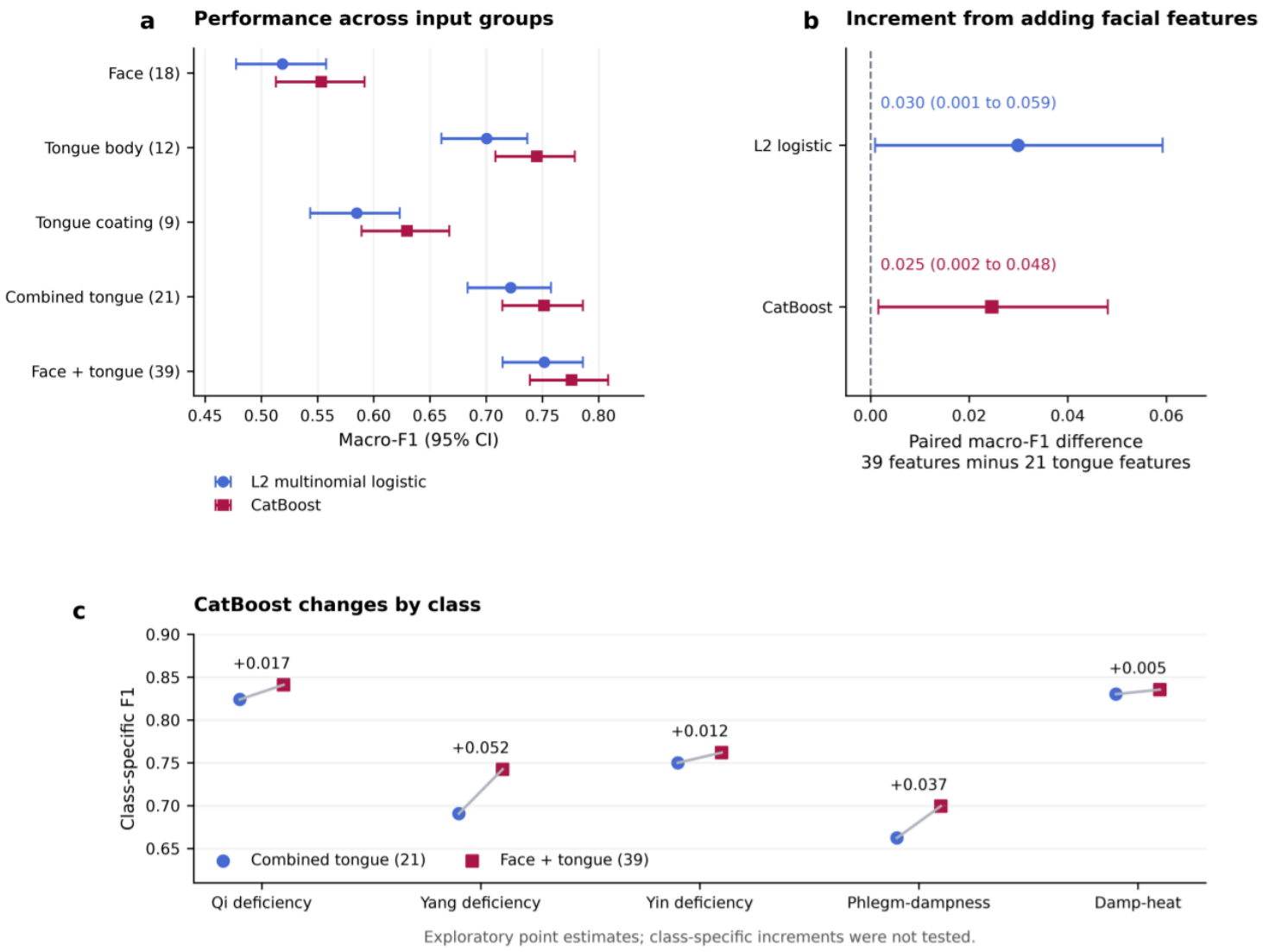


Figure 4

Modality ablation and incremental value of facial features.

Note: (a) Macro-averaged F1 scores with 95% confidence intervals by algorithm and feature set. (b) Paired within-algorithm F1 increments from adding facial features to combined tongue features; the dashed line indicates zero increment. (c) Class-specific F1 changes in CatBoost, shown as exploratory estimates without significance testing. Confidence intervals were estimated by participant-level bootstrap. Models used identical outer splits with separate inner tuning.

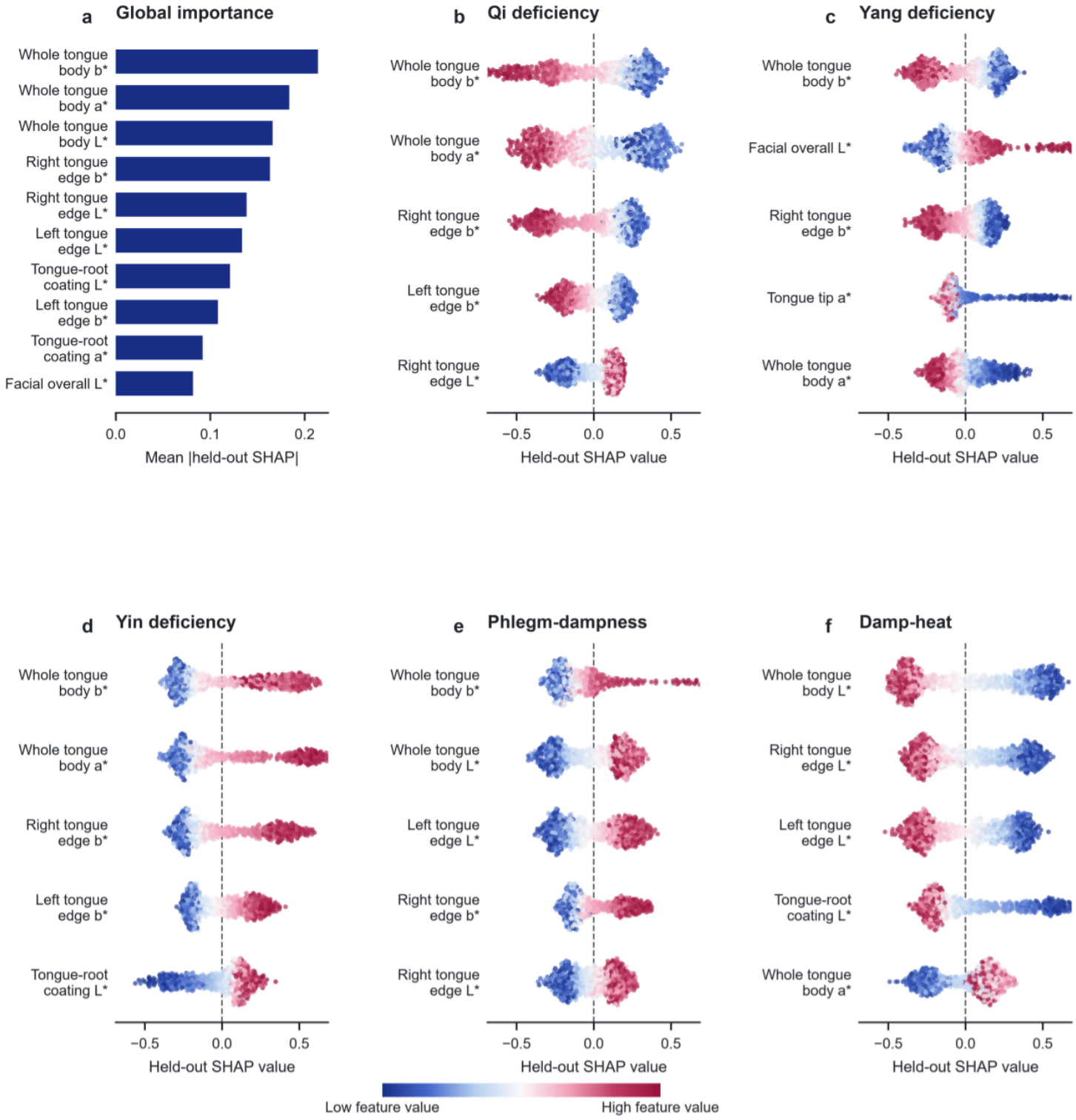


Figure 5

Global and class-specific SHAP explanations for CatBoost.

Note: (a) Global feature importance measured by mean absolute SHAP values. (b–f) SHAP distributions for qi deficiency, yang deficiency, yin deficiency, phlegm-dampness, and damp-heat, respectively. Positive SHAP values indicate contributions toward the corresponding class. Blue-to-red colors indicate low-to-

high feature values. SHAP values explain raw class scores rather than probability changes. Each class panel includes all 779 participants.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx