Supplementary Materials for:

Tongue Diagnosis Framework for Fatty Liver Disease Severity Classification Using Kolmogorov-Arnold Network

Section A. Data Quality Control

After segmenting facial photographs to obtain tongue images, we obtained an initial set of 6,298 images. A subsequent quality review identified several suboptimal cases. Specifically, some participants did not fully protrude their tongues, resulting in incomplete visualization of the dorsal surface. To address this issue, for each segmented tongue image we defined h as the vertical height (in pixels) of its bounding box. We then excluded images with $h < 420$ pixels, thereby removing 146 samples. In addition, motion blur was observed in a subset of images. To detect low local contrast consistent with blur, we computed s as the mean absolute difference of pixel intensities within a 20×20-pixel patch centered in the same bounding box. A further 435 images with $s < 2.60$ were excluded. In total, 581 images were removed, yielding 5,717 quality-controlled images suitable for subsequent tongue diagnosis modeling. Representative examples of excluded and retained cases after quality control are shown in Fig. A.1.

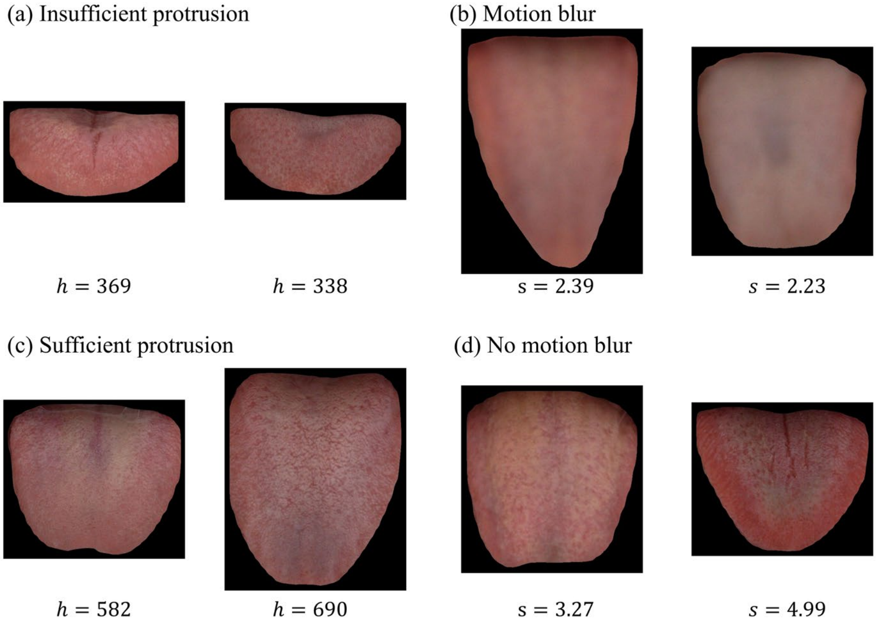


Fig. A.1 Representative tongue images: Excluded (a, b) and Retained cases (c, d).

Since both thresholds are label-agnostic by design, this quality control (QC) step is unlikely to introduce class-dependent selection bias. To formally evaluate whether QC altered the class distribution of FLD severity, we performed a Pearson's χ^2 test of independence on the 2×3 contingency table comparing sample counts pre- versus post-QC, as shown in Table A.1. Under the null hypothesis that class proportions are identical pre- and post-QC, the test yielded $\chi^2(2) = 8.95$, $p = 0.0126$. As an effect-size measure, Cramer's $V = 0.027$, indicating a negligible association [1]. In absolute terms, the post-QC changes in class proportions were modest: $+1.64$, -2.32, and $+0.69$ percentage points for Non-FLD, Mild, and Moderate/Severe, respectively. Taken together, the QC step led to a statistically detectable but practically minor shift in class distribution.

Table A.1 Distribution of samples across FLD severity categories before and after image quality control.

Quality control	**Total**	**Non-FLD**	**Mild**	**Moderate/Severe**
Pre-QC	6298	3962(62.91%)	1812(28.77%)	524(8.32%)
Post-QC	5717	3690(64.54%)	1512(26.45%)	515(9.01%)

Values are presented as n (%), calculated row-wise within each dataset.

Section B. Dataset Description

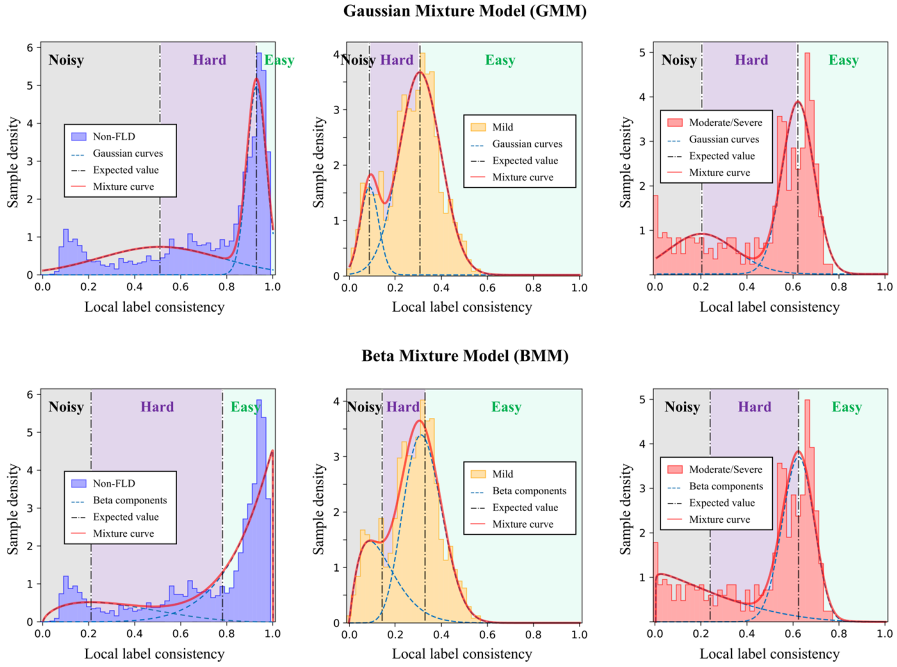
The newly constructed medium-sized tongue diagnosis dataset for FLD severity classification, named Tongue-FLD, is available at *https://github.com/MLDMXM2017/HSM-TDF*. The physiological indicators include Gender, Age, Height, Waist circumference (WC), hip circumference (HC), Weight, systolic blood pressure (SBP), and diastolic blood pressure (DBP). We performed statistical analysis of these indicators to examine their differences across groups with varying FLD severity, as shown in Table B.1. The *t-test* revealed significant differences ($P < 0.001$) between all paired categories for indicators, except for Gender, Age, and Height. Due to variations in tongue characteristics across different Genders and Age groups, these two factors are retained. Height is included because of its prior association with other indicators, such as Weight and blood pressure.

Table B.1 Differences in physiological indicators among FLD severity groups.

Indicator		**Non-FLD** **(n=3690)**	**Mild** **(n=1512)**	**Moderate/Severe** **(n=515)**
Gender	Male	1219 (33.04%) ^{a}	681 (45.04%) ^{b}	237 (46.02%) ^{b}
	Female	2471 (66.96%) ^{a}	831 (54.96%) ^{a}	278 (53.98%) ^{a}
Age (years)		54.66 (10.35) ^{a}	55.59 (9.46) ^{b}	55.75 (9.92) ^{b}
Height (cm)		159.82 (7.63) ^{a}	161.45 (8.14) ^{b}	161.77 (8.69) ^{b}
WC (cm)		77.63 (8.33) ^{a}	86.67 (7.64) ^{b}	92.35 (8.36) ^{c}
HC (cm)		91.66 (5.84) ^{a}	96.19 (5.89) ^{b}	99.55 (7.04) ^{c}
Weight (kg)		58.21 (8.35) ^{a}	67.54 (9.19) ^{b}	73.96 (11.26) ^{c}
SBP (mmHg)		126.06 (18.77) ^{a}	133.20 (18.52) ^{b}	139.70 (18.87) ^{c}
DBP (mmHg)		80.76 (10.31) ^{a}	85.32 (10.70) ^{b}	89.38 (11.52) ^{c}

Data are presented as mean (standard deviation), or frequency (percentage). The alphabetic superscripts indicate differences among categories. Identical superscripts or an inclusion relationship denote non-significant differences ($P < 0.001$), whereas different superscripts or no intersection indicate significant differences ($P \geq 0.001$)

Section C. Comparison of Fitting Models


Gaussian Mixture Model (GMM)

Beta Mixture Model (BMM)

Fig. C.1 Visual comparison of the Gaussian mixture model and beta mixture model fits.

To model the LLC distributions of noisy and true labels, prior work has commonly used Gaussian mixture models (GMMs). However, LLC is bounded on the unit interval $[0, 1]$, whereas Gaussian components have unbounded support. When fit directly, GMMs therefore place non-negligible probability mass outside the domain—for example, at values exceeding one for Non-FLD and at values below zero for Moderate/Severe. Moreover, with a limited number of components, the symmetry of individual Gaussians hinders accurate modeling of empirically skewed densities, such as the true-label distribution for Non-FLD and the noisy-label distribution for Moderate/Severe. In contrast, a beta mixture model (BMM) respects the bounded support and captures skewness parsimoniously. Consequently, the GMM fit distorts the expected values, thereby shifting decision thresholds: for example, many Non-FLD cluster-center samples with LLC exceeding 0.8 are incorrectly flagged as "hard," and approximately half of the noisy samples in Mild are not identified. Visual comparisons in Fig. C.1 corroborate these effects.

Section D. Implementation Details

Our framework is implemented using PyTorch on a NVIDIA GeForce RTX 4090 GPU with 24GB

RAM. We apply data augmentation techniques, including random rotation and horizontal flipping, followed by resizing the images to a standardized size of 448×448 before feeding them into the network. During training, we use the AdamW optimizer with a cosine decay schedule. The epochs for each training step are [80, 15, 20], with initial learning rates set to [0.001, 0.02, 0.02], respectively. Progressively-balanced sampling [2] is employed to mitigate the adverse effects of class imbalance. For a more comprehensive implementation, the code is available at *https://github.com/MLDMXM2017/HSM-TDF*.

Section E. Metrics Definition

In the context of multi-class classification tasks, we denote the number of samples labeled as k-th class and correctly predicted as such is denoted as True Positive (TP_k). The number of samples labeled as k-th class but incorrectly predicted as another class is denoted as False Negative (FN_k). Conversely, the number of samples incorrectly predicted as k-th class when they belong to another class is denoted as False Positive (FP_k). Metrics based on macro-averaging are given in Eq. (E.1) to Eq. (E.5):

$$Accuracy = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} (TP_k + FN_k)}, \tag{E.1}$$

$$Recall = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}, \tag{E.2}$$

$$Precision = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k}, \tag{E.3}$$

$$F1_{score} = \frac{1}{K}\sum_{k=1}^{K} \frac{2 \times Precision_k \times Recall_k}{Precision_k + Recall_k}, \tag{E.4}$$

$$AUC = \frac{1}{K}\sum_{k=1}^{K} AUC_k, \tag{E.5}$$

where K represents the total number of classes, and AUC_k is calculated following the one-vs-rest approach [3]. The metrics Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are calculated using the following equations:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \left|k_i - \hat{k}_i\right|, \tag{E.6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(k_i - \hat{k}_i\right)^2}, \qquad (\mathrm{E}.7)$$

where N the total number of samples, k_i represents the index of the true class for the i-th sample, while $\hat{k}_i$ denotes the predicted class index. The class indices 0, 1, and 2 correspond to Non-FLD, Mild FLD, and Moderate/Severe FLD, respectively.

Section F. Additional Performance Comparison Experiments

To investigate the impact of class imbalance in Tongue-FLD, we report per-class and overall metrics for HSM-TDF and the two strongest baselines (Li et al. [4]; ConvNeXt), as shown in Table F.1. In the "Overall" volumes, precision, recall, F1, and AUC are macro-averaged across the three classes, whereas MAE and RMSE are computed at the sample level. HSM-TDF achieves the highest macro-precision (0.6486) and macro-F1 (0.6207), the highest overall AUC (0.8694), and the lowest ordinal errors (MAE = 0.2682; RMSE = 0.5659). Importantly, it delivers the strongest performance on the minority classes: for Moderate/Severe, HSM-TDF attains the best F1 (0.4576) and recall (0.4402), the highest AUC (0.8967), and the lowest MAE/RMSE (0.6329/0.9381); for Mild, it achieves the best F1 (0.5475). By contrast, the method with the highest overall accuracy (Li et al., Acc = 0.7498) concentrates performance on the dominant Non-FLD class (F1 = 0.8421) while underperforming on the minority classes (Moderate/Severe F1 = 0.3495). ConvNeXt offers a competitive macro-F1 (0.6047) and improves upon Li et al. on the Moderate/Severe class (F1 = 0.3927 versus 0.3495), yet it still trails HSM-TDF on minority-class F1. Collectively, these results indicate that HSM-TDF mitigates majority-class bias and provides a more balanced performance profile.

Table F.1 Comparison of tongue diagnosis methods in FLD severity classification.

Method	Class	Acc	Pre	Rec	F1	AUC	MAE↓	RMSE↓
Li et al.	Non-FLD		0.8622	0.8483	0.8421	0.8895	0.1228	0.4339
	Mild		0.5766	0.4950	0.5217	0.8112	0.4602	0.7167
	Moderate/Severe		0.3980	0.3315	0.3495	0.8433	0.7934	1.1005
	Overall	0.7498	0.6123	0.5583	0.5711	0.8480	0.2758	0.5716
ConvNeXt	Non-FLD		0.8885	0.8830	0.8795	0.9192	0.1387	0.4707
	Mild		0.5660	0.5331	0.5418	0.8301	0.4612	0.6959
	Moderate/Severe		0.3854	0.4172	0.3927	0.8586	0.7728	1.0847
	Overall	0.7463	0.6133	0.6111	0.6047	0.8693	0.2843	0.5879
HSM-TDF	Non-FLD		0.9090	0.8151	0.8570	0.8987	0.1593	0.4759
	Mild		0.5529	0.5543	0.5475	0.8129	0.4007	0.6519
	Moderate/Severe		0.4839	0.4402	0.4576	0.8967	0.6329	0.9381
	Overall	0.7428	0.6486	0.6032	0.6207	0.8694	0.2682	0.5659

Section G. Additional Ablation Studies

Tongue diagnosis relies primarily on shallow, fine-grained textures and color cues. Accordingly, the image encoder must preserve low-level information. Although Conv-KAN offers strong nonlinear representational capacity, its per-node computational complexity makes the use of large kernels impractical. Residual blocks (Res-Blocks) provide an effective compromise by reducing spatial resolution while retaining salient low-level details. By contrast, standard convolutional neural network backbones such as ResNet-50 and ConvNeXt tend to prioritize deep semantic features at the expense of shallow textures, which can lead to insufficient extraction of tongue-specific cues.

To validate the effectiveness of the Res-Blocks + Conv-KAN design, we replace the image encoder (IE) with three alternatives: (i) a ResNet, (ii) a ConvNeXt, and (iii) a pure Conv-KAN extractor, while keeping the downstream decoder (DE), FFC, and MEC unchanged. As reported in Table G.1, the proposed IE + DE configuration achieves the strongest overall results (F1 = 0.6207; MAE = 0.2682), outperforming ResNet + DE (F1 = 0.5830; MAE = 0.2973), ConvNeXt + DE (F1 = 0.5793; MAE = 0.3079), and Conv-KAN + DE (F1 = 0.5714; MAE = 0.2989). These results demonstrate that coupling Res-Blocks with Conv-KAN preserves critical shallow textures for tongue diagnosis while maintaining computational efficiency, thereby yielding the most favorable accuracy–efficiency trade-off.

Table G.1 Ablation results for replacing the image encoder in the proposed method.

Encoder	Classifier	Acc	Pre	Rec	F1	AUC	MAE↓	RMSE↓
ResNet50+DE	FFC+MEC	0.7272	0.6215	0.5687	0.5830	0.8360	0.2973	0.5879
ConvNeXt+DE	FFC+MEC	0.7174	0.5986	0.5539	0.5793	0.8328	0.3079	0.5969
Conv-KAN+DE	FFC+MEC	0.7258	0.6059	0.5552	0.5714	0.8397	0.2989	0.5897
IE+DE	FFC+MEC	**0.7428**	**0.6486**	**0.6032**	**0.6207**	**0.8694**	**0.2682**	**0.5659**

To assess the necessity of using KAN at the classification stage, we replace all Linear-KAN classifiers with multilayer perceptrons (MLP). We match network depths and layer widths across architectures and hold all training settings constant. The results (Table G.2) indicate that substituting MLP degrades all metrics, most notably, F1 decreases from 0.6207 to 0.5862 (−0.0345) and MAE increases from 0.2682 to 0.3064 (+0.0382). These findings support the conclusion that the concatenated image and indicator features exhibit strong nonlinear interactions that MLP fails to capture. In contrast, KAN's spline-based nonlinearities more effectively model higher-order dependencies, yielding superior performance.

Table G.2 Ablation results for replacing the classifier architecture in the proposed method.

Encoder	Classifier	Acc	Pre	Rec	F1	AUC	MAE↓	RMSE↓
IE+DE	MLP, MLP	0.7228	0.6012	0.5807	0.5862	0.8335	0.3064	0.6035
IE+DE	FFC, MEC	**0.7428**	**0.6486**	**0.6032**	**0.6207**	**0.8694**	**0.2682**	**0.5659**

To quantify the contribution of each modality and to substantiate the value of fusion, we conduct an ablation study in which the model was trained (i) on tongue images only using the Image Encoder (IE), (ii) on physiological indicators only using the Dense Encoder (DE), and (iii) on both modalities using IE+DE. Detailed results are reported in Table G.3. The indicators-only model outperforms the images-only model (F1 increases from 0.4313 to 0.5451, +0.1138; AUC increases from 0.6783 to 0.8212, +0.1429; MAE decreases from 0.4747 to 0.3178, −0.1569). Multimodal fusion further improves performance (F1 increases from 0.5451 to 0.6207, +0.0756; AUC increases from 0.8212 to 0.8694, +0.0482; MAE decreases from 0.3178 to 0.2682, −0.0496). Together, these results show that fusion enhances discrimination while reducing ordinal error, indicating that tongue images and physiological indicators provide complementary information. Moreover, our method achieves effective cross-modal integration of these modalities.

Table G.3 Ablation results for different input modalities.

Imput	Encoder	Acc	Pre	Rec	F1	AUC	MAE↓	RMSE↓
Image	IE	0.5990	0.4430	0.4447	0.4313	0.6783	0.4747	0.7862
Indicators	DE	0.7073	0.5958	0.5365	0.5451	0.8212	0.3178	0.6056
Multimodal	IE+DE	**0.7428**	**0.6486**	**0.6032**	**0.6207**	**0.8694**	**0.2682**	**0.5659**

Section H. Hyperparameters Studies

We perform tuning on key hyperparameters of the proposed MFF-KAN to observe their impact on evaluation metrics, as shown in Table H.1. The adjusted hyperparameters include B-spline grid size $s_\mathrm{g} \in \{3,5,7\}$, order $k_\mathrm{B} \in \{2,3,4,5\}$ and DE width $w_\mathrm{DE} \in \{64,128,512\}$. We evaluate the overall impact of hyperparameter variations on performance by analyzing changes in average rank. As shown in the table, a grid size $s_\mathrm{g} = 5$ is appropriate. A smaller $s_\mathrm{g} = 3$ weakens the fitting capability of the B-spline, while a larger $s_\mathrm{g} = 7$ may lead to overly complex activation functions, which could degrade performance. Both scenarios lead to a decrease in average rank. In B-splines, the order k_B determines the influence range of each knot and the overall smoothness of the curve. The poor performance at $k_\mathrm{B} = 2$ suggests that a limited influence range hinders the activation function's ability to learn overall patterns. Orders of

3 and 4 yield similar superior performance, with average rankings of 2.4 and 2.0, respectively. Considering that increasing the order results in higher computational overhead, we adopt $k_{\mathrm{B}} = 3$ in the proposed method to maintain a balance between efficiency and performance. However, an excessively large k_{B}, such as 5, causes changes at each knot to affect the entire spline, which not only degrades performance but also increases computational cost. Additionally, w_{DE} determines the number of nodes and paths in the network. It can be observed that increasing w_{DE} from 64 to 128 improves the average rank by 2.3, similar to how an MLP gains stronger representational capacity by increasing layer width. However, an excessive number of nodes increases the difficulty of selecting critical nodes and paths through the regularization loss function $\mathcal{L}_{\mathrm{Reg}}^{l}$, leading to a decline in performance, as observed at $w_{\mathrm{DE}} = 512$. In summary, hyperparameters may influence the network's fitting capability and learning process from different perspectives, and should be appropriately selected within a reasonable range.

Table H.1 The ablation study results for the hyperparameters in the proposed method

Grid size	Order	KAN width	Accuracy	Precision	Recall	F1-score	AUC	MAE↓	RMSE↓	AvgRank↓
3	3	128	0.7253$_{(7)}$	0.6376$_{(5)}$	0.6226$_{(2)}$	0.6239$_{(2)}$	0.8444$_{(2)}$	0.2860$_{(6)}$	0.5773$_{(5.5)}$	4.2
7	3	128	0.7313$_{(5)}$	0.6320$_{(7)}$	0.6064$_{(5)}$	0.6102$_{(5)}$	0.8427$_{(3)}$	0.2800$_{(3)}$	0.5727$_{(3)}$	4.4
5	2	128	0.7317$_{(4)}$	0.6310$_{(8)}$	0.6087$_{(4)}$	0.6093$_{(7)}$	0.8384$_{(5.5)}$	0.2865$_{(8)}$	0.5831$_{(8)}$	6.4
5	4	128	0.7381$_{(2)}$	**0.6495**$_{(1)}$	0.6221$_{(3)}$	**0.6279**$_{(1)}$	0.8387$_{(4)}$	0.2741$_{(2)}$	**0.5650**$_{(1)}$	**2.0**
5	5	128	0.7325$_{(3)}$	0.6448$_{(3)}$	0.5927$_{(7)}$	0.6095$_{(6)}$	0.8381$_{(8)}$	0.2863$_{(7)}$	0.5773$_{(5.5)}$	5.6
5	3	64	0.7311$_{(6)}$	0.6321$_{(6)}$	**0.6292**$_{(1)}$	0.6218$_{(3)}$	0.8384$_{(5.5)}$	0.2852$_{(4.5)}$	0.5819$_{(7)}$	4.7
5	3	512	0.7251$_{(8)}$	0.6446$_{(4)}$	0.5884$_{(8)}$	0.6043$_{(8)}$	0.8383$_{(7)}$	0.2852$_{(4.5)}$	0.5764$_{(4)}$	6.2
5	**3**	**128**	**0.7428**$_{(1)}$	0.6486$_{(2)}$	0.6032$_{(6)}$	0.6207$_{(4)}$	**0.8694**$_{(1)}$	**0.2682**$_{(1)}$	0.5659$_{(2)}$	2.4

Section I. Representative Misclassified Cases

To further elucidate the sources of misclassification, we examine one cross-validation fold and, for each combination of true and predicted classes, selecte representative samples for illustration. Low-uncertainty (< 0.3) cases are shown in Fig. I.1, whereas high-uncertainty (< 0.7) cases are shown in Fig. I.2. Samples are arranged in the corresponding cells of the confusion matrix, using the label encoding 0 = Non-FLD, 1 = Mild, and 2 = Moderate/Severe. Empty cells indicate that no examples were available for the corresponding confusion-matrix entry.

Among low-uncertainty cases, those predicted as Non-FLD (class 0) commonly exhibit a thin white tongue coating, even when the ground-truth label indicates FLD. Cases predicted as Mild (class 1) typically show a purplish tongue body with mild fissuring. Among cases predicted as Moderate/Severe (class 2), a subset displays dark-purplish discoloration and, in some instances, erythema at the tongue tip.

By contrast, most high-uncertainty cases simultaneously exhibit multiple features, including a thin white coating, a pale red tongue body, localized purplish regions, and mild fissuring. As noted above, each of these characteristics is informative for FLD severity classification. When they co-occur, they provide conflicting cues, thereby confounding the model's decision and elevating predictive uncertainty.

These findings indicate that specific tongue characteristics are primary determinants of the model's decisions. When characteristics associated with a single severity class co-occur, the model produces low predictive uncertainty. By contrast, the simultaneous presence of characteristics linked to multiple classes increases predictive uncertainty. Notably, misclassifications can still occur even when predictive uncertainty is low, reflecting the non-deterministic relationship between tongue phenotypes and FLD severity. Accordingly, the proposed approach is intended for broad preliminary screening rather than as a replacement for definitive clinical diagnosis.

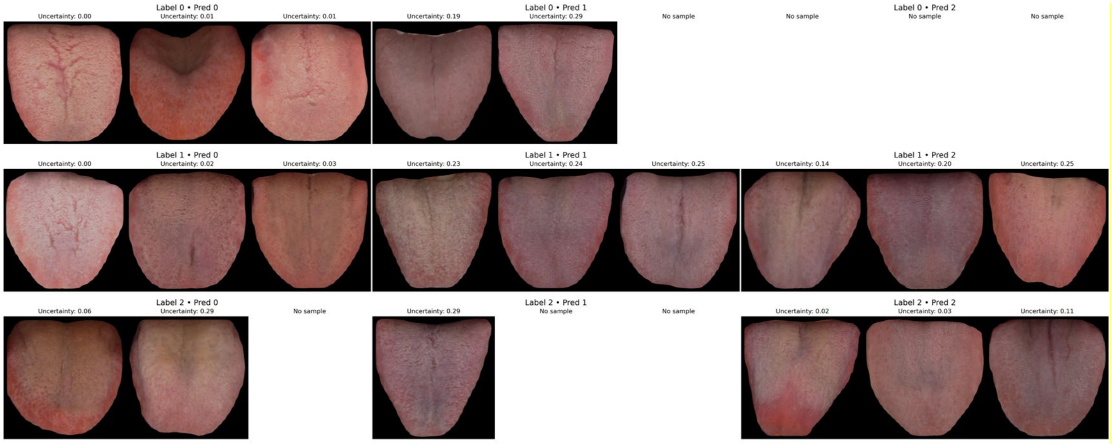


Fig. I.1 Representative low-uncertainty tongue images arranged by confusion-matrix position (Label vs. Prediction), including correct and misclassified cases. Samples are arranged in the corresponding cells of the confusion matrix, using the label encoding 0 = Non-FLD, 1 = Mild, and 2 = Moderate/Severe.

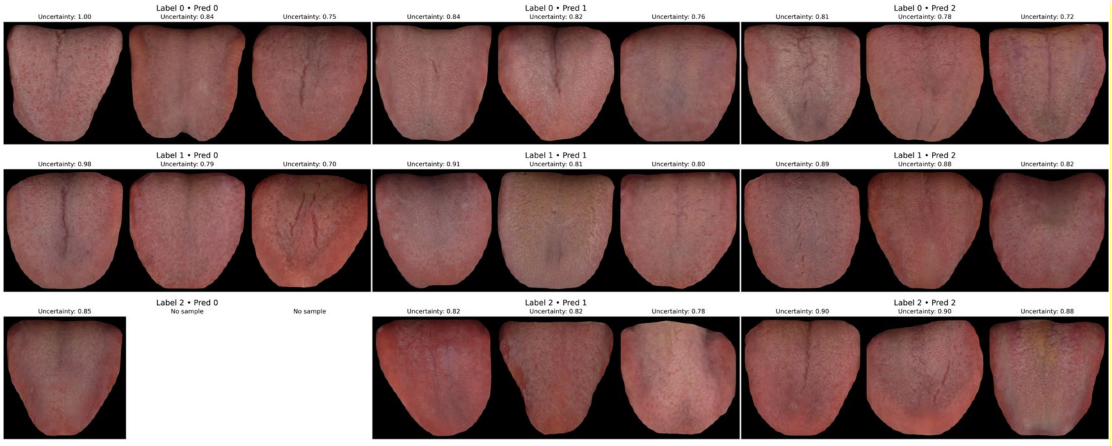


Fig. I.2 Representative high-uncertainty tongue images arranged by confusion-matrix position (Label vs. Prediction), including correct and misclassified cases. Samples are arranged in the corresponding cells of the confusion matrix, using the label encoding 0 = Non-FLD, 1 = Mild, and 2 = Moderate/Severe.

Section J. Crucial Tongue Characteristics

We consider the tongue characteristics closely associated with key nodes as crucial tongue characteristics. To determine which tongue characteristic is closely linked to a key node, we identify tongue images with high activation values at that node as representative samples, which are presented in Fig. J.1. Subsequently, the common tongue characteristics are summarized from these representative samples. As indicated in the figure, the samples with high activation values at nodes 90, 210, 274, 275, and 391 exhibit the characteristics of red tongue, white coating, thin coating, purple tongue, and fissured tongue, respectively. This indicates that tongue color and coating play a significant role in classifying the severity of FLD.

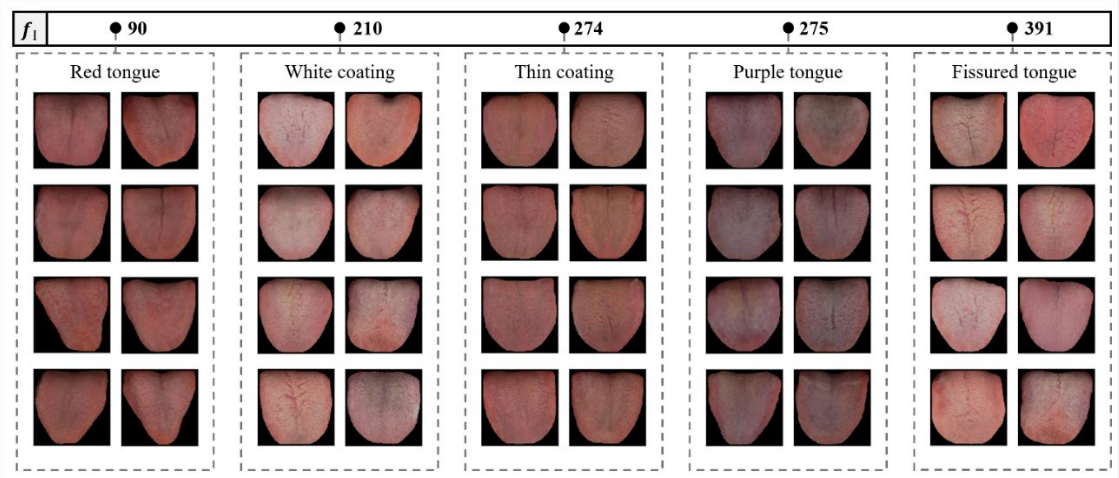


Fig. J.1 Representative tongue images corresponding to key nodes in the image feature vector extracted by the IE module.

Reference

[1] M.L. McHugh, The chi-square test of independence, Biochemia medica 23 (2013) 143-149.

doi:10.11613/bm.2013.018.

[2] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, et al., Decoupling Representation and Classifier for Long-Tailed Recognition, 2019, arXiv preprint arXiv:1910.09217.

[3] A.C.J.W. Janssens, F.K. Martens, Reflection on modern methods: Revisiting the area under the ROC Curve, Int. J. Epidemiol. 49 (2020) 1397–1403. doi:10.1093/ije/dyz274.

[4] J. Li, P. Yuan, X. Hu, J. Huang, L. Cui, J. Cui, et al., A tongue features fusion approach to predicting prediabetes and diabetes with machine learning, J. Biomed. Inform. 115 (2021) 103693. doi:10.1016/j.jbi.2021.103693.