Accurate Fatty Liver Disease Diagnosis with a Multi-Source Feature Fusion Model on the Segmented Tongue Image Dataset

Jie Gao, Tao Chen, Yong Xu, Yijie Wu, Kunhong Liu, *Member, IEEE*, Weihong Qiu, Weimin Ye

**Supplementary Material**

# Section A

**Statistical Analysis on Dataset**

The Segmented Tongue Image Dataset (STID) comprised 5,717 samples, encompassing tongue images of 3,690 Non-FLD participants and 2,027 FLD patients, resulting in an imbalance ratio of 1.82. Additionally, each participant’s data included eight physiological indicators: sex, age, height, waist circumference, hip circumference, weight, systolic blood pressure (SBP), and diastolic blood pressure (DBP). The samples were categorized based on their labels, and a statistical analysis was conducted to compare the indicators between the two groups. We calculated the *P*-values for the distribution differences of each indicator within the Non-FLD and FLD samples, respectively, and found that all indicators satisfied *P*-values < 0.001 [1]. This indicates that these indicators possess a high level of statistical significance. Table A1 illustrates statistically significant differences in all eight indicators between the two groups, with the FLD group exhibiting higher values for age, height, waist circumference, hip circumference, weight, SBP, and DBP. Furthermore, the Cohen's d values for waist circumference, hip circumference and weight surpassed 0.8, indicating substantial effects of these indicators [2].

TABLE A1

Differences in Physiological Indicators between Non-FLD and FLD Groups

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Non-FLD**  **()** | **FLD**  **()** | **Cohen's** |
| **Sex (male/female)** | 1219/2471 | 918/1109 | 0.255 |
| **Age (years)** | 54.7(10.3) | 55.6(9.6) | 0.095 |
| **Height (cm)** | 159.8(7.6) | 161.5(8.3) | 0.218 |
| **Waist (cm)** | 77.6(8.3) | 88.1(8.2) | 1.265 |
| **Hipline (cm)** | 91.7(5.8) | 97.0(6.4) | 0.892 |
| **Weight (kg)** | 58.2(8.4) | 69.2(10.2) | 1.214 |
| **SBP (mmHg)** | 126.1(19.2) | 134.9(18.8) | 0.468 |
| **DBP (mmHg)** | 80.8(10.3) | 86.4(11.1) | 0.529 |
| Note: Continuous variables are described as mean (standard deviation). | | | |

TABLE A2

Counts of Labels Across Different Age Ranges and Sex Groups

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Age** | **Sex** | **Num** | **Non-FLD** | **FLD** | **Non-FLD /FLD** |
| **35–54** | Male | 886 | 421 | 465 | 0.91 |
| Female | 1951 | 1463 | 488 | 3.00 |
| **55–75** | Male | 1251 | 798 | 453 | 1.76 |
| Female | 1629 | 1008 | 621 | 1.62 |
| **Total** | / | 5717 | 3690 | 2027 | 1.82 |

Moreover, statistically significant differences in tongue characteristics and prevalence of FLD were observed across different age groups and sexes [3]. Consequently, we conducted a statistical analysis of age, sex, and label distribution in the published dataset. As shown in Table A2, the dataset adequately encompassed diverse age groups, sexes, and labels. However, there was a higher proportion of females than males, particularly within the 35–54 age group. Additionally, there was a skew in label distribution, with the number of the non-FLD samples generally exceeding those labeled FLD, especially prominent among the females aged 35–54. Therefore, it could be advisable to implement appropriate balancing measures during model training to mitigate this imbalance.

# Section B

**Implement Details of** **Comparison Algorithms**

We conducted the training and validation for all comparison models following the configurations outlined in Table A3. All the models underwent training using the SGD optimizer with a batch size of 64 and a cosine decay learning rate scheduler. Because UNet and DeepLabV3+ were primarily utilized for instance segmentation, their adaptation for image classification tasks involved the initial generation of a tensor serving as a mask annotation during training. This tensor utilized for calculating the loss function with the model output assigned the label values to the tongue area. For the predictions, the classification results were determined by computing the global average across all channels of the outputs from UNet and DeepLabV3+.

TABLE A3

Training Configuration for Comparison Models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Model** | **Input** | **Output** | **Epochs** | **Learning Rate** | **Weight Decay** | **Pretrained** |
| **01** | ResNet-34 | Image | Logits | 70 | 1e-2 | 1e-3 | True |
| **02** | ResNet-50 | Image | Logits | 50 | 1e-2 | 1e-3 | True |
| **03** | VGG-16 | Image | Logits | 90 | 1e-2 | 1e-3 | True |
| **04** | DenseNet-201 | Image | Logits | 60 | 1e-3 | 1e-4 | True |
| **05** | MobileNetV3 | Image | Logits | 50 | 1e-2 | 1e-3 | True |
| **06** | ShuffleNetV2 | Image | Logits | 50 | 1e-3 | 1e-4 | True |
| **07** | ConvNeXt | Image | Logits | 80 | 1e-2 | 1e-3 | True |
| **08** | UNet | Image | Segmentation | 45 | 1e-3 | 1e-4 | False |
| **09** | DeepLabV3+ | Image | Segmentation | 45 | 1e-3 | 1e-4 | True |
| **10** | TransFG | Patches | Logits | 70 | 5e-3 | 5e-4 | True |
| **11** | TIPNet | Image | Logits | 150 | 1e-2 | 1e-3 | False |

TABLE A4

Results of Models with Various Branch Counts and Kernel Sizes

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Kernel Size** | **Channels per Branch** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **AUC** |
|  | {3,5} | 768 | 0.624 | 0.740 | 0.676 | 0.749 | 0.852 |
|  | {3,5,7,9} | 384 | 0.634 | 0.714 | 0.671 | 0.751 | 0.851 |
|  | {5,5,5} | 512 | **0.646** | 0.708 | 0.673 | **0.758** | **0.855** |
| **TIPNet** | {3,5,7} | 512 | 0.623 | **0.769** | **0.688** | 0.754 | **0.855** |

# Section C

**Ablation Studies**

In this section, we describe detailed ablation studies conducted to assess the efficacy of different architectural designs within TIPNet for extracting image features relevant to the tongue diagnosis of FLD. These encompassed the influence of multi-branch and multi-scale convolutional kernels, the disparities in the feature extraction between shallow and deep networks, and the impacts of various attention mechanisms. By substituting TIPNet, utilized for multi-scale feature extraction in our proposed method, with corresponding variants, we indirectly compared the feature extraction and processing capabilities of each network structure through their predictive performance for tongue diagnosis of FLD. To focus on the feature extraction from tongue images, only images were utilized as inputs in these experiments.

**1) Ablation of Branch Counts and Kernel Sizes**

To assess the effectiveness of the multi-branch and multi-scale kernel mechanisms, we introduced two types of variants around TIPNet. The first type, denoted as and , maintained the same number of output channels as TIPNet in the last layer, but varied the number of branches by either reducing or increasing it by one. The second type, , comprised three branches, each utilizing the same convolutional kernel size . Table A4 presents the results of the five-fold cross-validation for the TIPNet and the three variant networks. The variant underperformed compared to TIPNet, with a decrease of 1.2% in F1-score, attributed to the absence of a branch with a kernel size of 7, leading to an insufficient capability to handle large-scale features. Meanwhile, the variant, with more branches and a richer variety of kernel sizes, experienced a further decrease of 0.5% in F1-score. This reduction may result from the convolution kernel size of nine, which contained too many parameters, making parameter initialization and training challenging. For the variant, which also utilized three branches, its F1-score was 1.5% lower than that of TIPNet, and recall decreased by 6.1%, as its overly uniform kernel size failed to adequately extract the features of different scales. These comparisons demonstrated that our proposed three-branch network structure with convolution kernel sizes of 3, 5, and 7 better balances the diversity of feature extraction and training complexity, thus yielding superior performance in tongue diagnosis for FLD.

**2) Ablation of Model Depth**

In this section of the experiment, we compared the efficacy of tongue characteristic extraction across various depths of TIPNet variants. Specifically, we modified the number of sequentially connecting TCB-Blocks and TCN-Blocks in each branch of TIPNet. These variants were denoted according to the number of convolutional layers as TIPNet-{6, 9, 11, 14}, with TIPNet-9 representing the network proposed in Section III. The quantitative results of the five-fold cross-validation of each model are presented in Table A5. Notably, TIPNet-9, comprising one TCB-Block and two TCN-Blocks, achieved the highest predictive performance. Conversely, TIPNet-6 performed the poorest because of its insufficient depth, hindering its ability to effectively handle the rich color and texture information in tongue images. Meanwhile, TIPNet-11

TABLE A5

Results of Models with Various Depths

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **TCB-Block** | **TCN-Block** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **AUC** |
| **TIPNet-6** | 1 | 1 | 0.629 | 0.726 | 0.673 | 0.751 | 0.848 |
| **TIPNet-9** | 1 | 2 | 0.623 | **0.769** | **0.688** | 0.753 | **0.855** |
| **TIPNet-11** | 2 | 2 | 0.644 | 0.716 | 0.678 | 0.758 | 0.853 |
| **TIPNet-14** | 2 | 3 | **0.652** | 0.707 | 0.676 | **0.760** | 0.849 |

and -14, with greater depths and more learnable parameters, exhibited a decline in the predictive performance, possibly because of their tendency to over-extract less relevant shape and contour information from tongue images as depth increased, inhibiting the transmission of critical color and texture information. Moreover, as the network depth increased, training shallower layers via back-propagation became more challenging. Consequently, both excessively shallow and deep network structures could lead to the performance degradation. Our proposed TIPNet-9, compared to other depth variants, is more suitable for tongue diagnosis for FLD.

**3) Ablation of Attention Mechanism**

Within TIPNet, we incorporated the attention mechanisms to regulate channel and spatial weights, aiding the network in prioritizing the extraction of features pertinent to tongue diagnosis. In this ablation study, we investigated the impact of various attention mechanisms on the proposed model. Compared variants included certain models employing Efficient Channel Attention (ECA) [4], Style-based Recalibration Module (SRM) [5], Convolutional Block Attention Module (CBAM) [6], and Multidimensional Collaborative Attention (MCA) [7] as attention modules, alongside a model without any attention mechanism (labeled “None”). The ECA utilized the 1D convolution to compute the inter-channel relationships with the minimal computational cost, thereby enhancing the efficiency of the network in the feature utilization. The SRM introduced the “style pooling”, computed through mean and variance pooling, to extract the style information from each channel of feature maps, thereby enhancing the feature representation. The CBAM integrated the spatial and channel attention mechanisms within a unified module to improve the ability of the model to highlight the key information. The MCA stood out by simultaneously inferring the attention in the channel, height, and width dimensions, thus directing the network’s focus towards critical image features more effectively.

Table A6 presents the performance metrics derived from the five-fold cross-validation for each model. The findings indicated that integrating the proposed attention mechanism enhanced the model performance, with increases in Recall and F1-score of 4.2% and 1.2%, respectively. The models augmented with MCA exhibited slight improvements in all metrics, except Precision. Interestingly, the inclusion of ECA, SRM, and CBAM resulted in a decreased model performance. This decline may be attributed to the ECA and SRM adjusting weights solely through channels, neglecting the spatial distribution of the feature information. The sequential weight adjustment of CBAM, prioritizing channels before spatial aspects, poses challenges in preserving the spatial feature distribution after the channel adjustment. In contrast, the attention mechanism proposed in this study concurrently adjusted the image features from both the channel and spatial perspectives in a parallel configuration which could align more closely with the requirements for extracting image features in tongue diagnosis tasks for FLD.

# Section D

**Predictive Performance in Subgroups**

The age and sex affected the distribution of tongue characteristics. Furthermore, the label distributions across different age ranges and sex groups within the dataset exhibited

TABLE A6

Results of Models with Various Attention Mechanisms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attention** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **AUC** |
| **None** | 0.633 | 0.727 | 0.676 | 0.753 | 0.846 |
| **ECA** | 0.625 | 0.735 | 0.675 | 0.750 | **0.857** |
| **SRM** | 0.637 | 0.705 | 0.664 | 0.748 | 0.853 |
| **CBAM** | 0.637 | 0.706 | 0.669 | 0.753 | 0.843 |
| **MCA** | **0.643** | 0.713 | 0.676 | **0.758** | 0.852 |
| **TIPNet** | 0.623 | **0.769** | **0.688** | 0.753 | 0.855 |

TABLE A7

Results of the Proposed Method Across Age and Sex Subgroups

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | **Sex** | **Num** | **Non-FLD**  **/FLD** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **AUC** |
| **35–54** | All | 2289 | 2.127 | **0.785** | **0.853** | 0.817 | **0.878** | **0.941** |
| **55–75** | All | 3234 | 1.660 | 0.736 | 0.846 | 0.787 | 0.828 | 0.912 |
| **35–75** | Male | 2137 | 1.328 | 0.763 | 0.892 | **0.822** | 0.835 | 0.921 |
| **35–75** | Female | 3578 | 2.229 | 0.743 | 0.811 | 0.775 | 0.855 | 0.924 |

subgroups delineated by age range and sex. To this end, we segmented the dataset into two age subsets, including 35–54 and 55–75 years, through employing the median age of 55 years as the demarcation, and two sex subsets, male and female. Subsequently, we separately evaluated the performance of the proposed method using the fusion strategy within each subset, as presented in Table A7. In terms of age groups, the proposed method demonstrated the superior performance in the 35–54 age group across all metrics compared with the 55–75 age group, with a notable increase in the F1-score by 3.0%. This disparity may arise from the presence of multiple health abnormalities among older participants, potentially disrupting the normal mapping relationship between tongue characteristics and FLD labels, increasing the diagnostic complexity. Regarding the comparison of the sex groups, the predictive results of the proposed method in the female subset were notably inferior to those in the male subset. This discrepancy was attributed to the larger sample imbalance ratio of 2.229 in the female subset, posing a greater challenge for accurate FLD prediction. Nevertheless, the proposed method still achieved an F1-score of at least 0.775 in the female subset, demonstrating its ability to maintain high diagnostic effectiveness in tongue diagnosis for FLD across diverse subpopulations.

References

[1] S. Greenland, "Valid *P*-values behave exactly as they should: Some misleading criticisms of *P*-values and their resolution with *S*-values," *Am. Statistician,* vol. 73, pp. 106-114, 2019, doi: 10.1080/00031305.2018.1529625.

[2] M. A. Kraft, "Interpreting Effect Sizes of Education Interventions," *Educ. Res.,* vol. 49, no. 4, pp. 241-253, May. 2020, doi: 10.3102/0013189x20912798.

[3] P. C. Hsu *et al.*, "Gender- and age-dependent tongue features in a community-based population," *Medicine,* vol. 98, no. 51, Dec. 2019, doi: 10.1097/md.0000000000018350.

[4] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," presented at the *2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Virtual, Jun. 13-19, 2020.

[5] H. Lee, H. E. Kim, H. Nam, and Ieee, "SRM : A style-based recalibration module for convolutional neural networks," presented at the *IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Seoul, South Korea, Oct. 27-Nov. 02, 2019.

[6] S. H. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," presented at the *15th Eur. Conf. Computer Vision (ECCV)*, Munich, Germany, Sep. 08-14, 2018.

[7] Y. Yu, Y. Zhang, Z. Y. Cheng, Z. Song, and C. K. Tang, "MCA: Multidimensional collaborative attention in deep convolutional neural networks for image recognition," *Eng. Appl. Artif. Intell.,* vol. 126, Art. no. 107079, Nov. 2023, doi: 10.1016/j.engappai.2023.107079.