CBF: SEVERITY-AWARE MULTIMODAL FATTY LIVER DISEASE GRADING VIA COLLABORATIVE BRANCH FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Combining tongue images with routine physiological indicators offers a non-invasive means of grading fatty liver disease (FLD). Multimodal models such as HM-TDF fuse both inputs to improve overall accuracy. However, we find that this gain has a clinical cost: the fused model recognizes fewer moderate-to-severe cases than a model that uses the physiological indicators alone. Averaging the predictions of single-input branches at inference raises accuracy but does not recover these cases, because the branches are never trained for the shared decision. To make fusion severity-aware, we propose Collaborative Branch Fine-Tuning (CBF), which keeps tongue-only, physiology-only, and joint predictors and fine-tunes them together with a loss on their averaged prediction and on each branch. We prove that this objective reallocates each record's update toward the branches that carry the correct signal, while the per-branch term keeps every branch a competent predictor on its own. Extensive experiments on Tongue-FLD show that CBF achieves the best accuracy and macro-F1 among all compared methods and restores the moderate-to-severe recall lost by fusion. More importantly, the recovered predictions rely more on the patient's own tongue image.

1 INTRODUCTION

Grading fatty liver disease (FLD) without imaging equipment would support large-scale screening, and the appearance of the tongue combined with routine body measurements has emerged as a promising non-invasive signal (Jiang et al., 2021; Wang et al., 2024; Chen et al., 2025). The two inputs describe complementary aspects of the same person, and fusing such inputs is a central theme in multimodal machine learning (Baltrušaitis et al., 2019; Huang et al., 2020). Among the three grades, moderate-to-severe FLD is the one most likely to prompt further assessment and follow-up, so missing it carries a higher practical cost than confusing mild and normal cases.

Early studies combined quantitative tongue features with body measurements for binary NAFLD detection (Jiang et al., 2021; Wang et al., 2024). HM-TDF (Chen et al., 2025) advanced the task to three-grade classification, released the Tongue-FLD dataset, and introduced a fusion network with a multi-expert classifier for hard samples, reporting higher accuracy than both tongue-diagnosis methods and generic vision backbones. However, ***a higher overall accuracy can come at the cost of the most consequential grade***: the reproduced HM-TDF model is more accurate than a physiology-only model, yet it recognizes markedly fewer moderate-to-severe cases (Figure 1). This mirrors observations in multimodal learning, where joint training can favor one input and underuse the signal of another (Wu et al., 2022; Huang et al., 2022; Peng et al., 2022). Because only 515 of 5,715 records are moderate-to-severe, a gain on the large normal group easily masks such a loss.

In this work, we ask whether the high-grade signal of single-input predictors can be kept without giving up the accuracy of fusion. A natural first step is to keep tongue-only and physiology-only branches next to the joint model and average their predictions. However, ***averaging the branches at inference raises overall accuracy but leaves moderate-to-severe recall almost unchanged***, because each branch is still trained for its own objective rather than for the shared decision. We therefore propose **Collaborative Branch Fine-Tuning (CBF)**, which fine-tunes the three branches together with a loss on their averaged prediction, the output used at deployment, and an auxiliary loss on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

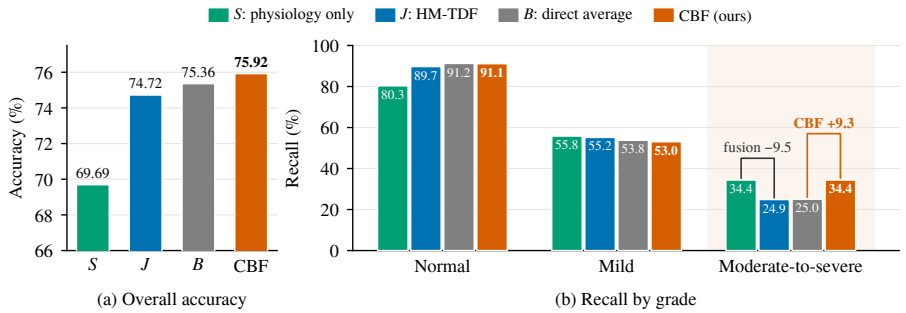


Figure 1: **Fusion can miss severe cases.** Overall accuracy (a) and grade-wise recall (b) for S, J, B, and CBF. Fusion (J) raises accuracy but lowers moderate-to-severe recall from 34.37% (physiology only, S) to 24.85%; direct averaging (B) leaves it at 25.05%, and CBF restores it to 34.37%.

each branch. The design draws on the *joint training of deep ensembles* (Jeffares et al., 2023), where supervising the averaged output couples the members, and adds per-branch supervision to avoid the *learner collusion* that training the average alone can cause.

We further prove that the CBF objective is equivalent to per-record, adaptively weighted cross-entropy on each branch: every branch receives an update in proportion to the probability it already assigns to the true grade, while the per-branch term guarantees it a minimum weight (Proposition 1), so the branches take complementary roles rather than act as interchangeable members.

We evaluate CBF on Tongue-FLD against nine previously evaluated methods and a protocol-matched reproduction of HM-TDF, and complement the comparison with ablations, paired per-grade case accounting with exact significance tests, branch-level analysis, Grad-CAM visualization, and a physiology-matched tongue-replacement test. CBF achieves the best accuracy and macro-F1 among all compared methods and raises moderate-to-severe recall by ***9.52 percentage points*** over HM-TDF, with a gain over direct averaging in every fold. Moreover, ***the recovered predictions rely more on the patient's own tongue image***, as both Grad-CAM and tongue replacement show. Our contributions are summarized as follows:

- We identify a hidden grade-wise cost of fusion in tongue–physiology FLD grading, which averaging at inference does not remove, and expose it with paired per-grade accounting.
- We propose CBF, which trains three branches for their shared decision, and prove that it reallocates each update toward the branches carrying the correct signal.
- Through comparisons, ablations, significance tests, and tongue-reliance tests, we show that CBF achieves the best accuracy and macro-F1 and restores moderate-to-severe recall.

2 Related Work

Tongue-based and multimodal assessment of fatty liver disease. Tongue diagnosis has progressed from hand-crafted features to learned representations, and tongue appearance has been linked to NAFLD. Jiang et al. (2021) combined quantitative tongue features with body measurements or laboratory tests for binary NAFLD detection, and Wang et al. (2024) examined tongue and anthropometric variables with conventional classifiers. HM-TDF (Chen et al., 2025) extended the task to three-grade classification, released the Tongue-FLD dataset, and fused tongue images with physiological indicators through a KAN-based network with a multi-expert classifier for hard samples. More broadly, combining images with tabular clinical records is a well-studied setting in medical deep learning (Huang et al., 2020): Multimodal Lego reuses biomedical encoders across modalities (Hemker et al., 2025), and M4oE learns patient- and task-dependent expert pathways (Wu et al., 2025). These studies report and optimize overall performance; how fusion changes the recognition of each disease grade has not been examined, and this is the problem our work addresses.

Modality imbalance in multimodal learning. Joint training of multimodal networks is known to underuse some inputs, whether through greedy reliance on the fastest-learning modality (Wu et al.,

2022) or through provable competition between modality encoders (Huang et al., 2022). Existing remedies fall into three groups. Reweighting methods adjust the contribution of each modality during training, as in Gradient Blending (Wang et al., 2020) and on-the-fly gradient modulation (Peng et al., 2022). Unimodal-assistance methods add unimodal teachers or ensembles (Du et al., 2023), alternate unimodal updates (Zhang et al., 2024), or reconcile joint and unimodal gradients (Wei & Hu, 2024). Objective-level methods maximize the total correlation between modalities (Yu et al., 2026). Class-balanced losses (Cui et al., 2019) and logit adjustment (Menon et al., 2021) are complementary, as they reweight labels rather than modalities. These methods improve the overall accuracy of a shared model, yet their effect on individual grades of an ordinal disease, where the smallest grade matters most, remains largely unexplored.

Joint training of ensembles. Deep ensembles are usually trained member by member and averaged at inference (Lakshminarayanan et al., 2017), although the averaged output can also be supervised directly. Jeffares et al. (2023) show that optimizing only the averaged output lets members collude, becoming individually degenerate and generalizing worse, and study objectives that interpolate between independent and joint training. Their analysis considers members that receive the same input. How such an interpolation distributes learning among members that see different modalities, and how it affects each disease grade, has not been characterized, which is what our analysis provides.

3 METHOD

3.1 OVERALL FRAMEWORK

We represent the grading data as paired records

$$\mathcal{D} = \{(x_i, s_i, y_i)\}_{i=1}^{N}, \quad x_i \in \mathbb{R}^{3\times H\times W}, \quad s_i \in \mathbb{R}^9, \quad y_i \in \{0, 1, 2\}. \tag{1}$$

Here x_i is a segmented RGB tongue photograph, s_i encodes eight recorded physiological indicators (sex, age, height, waist circumference, hip circumference, weight, and systolic and diastolic blood pressure), and y_i is the ultrasound-assessed grade: no FLD, mild FLD, or moderate-to-severe FLD (Chen et al., 2025). A predictor outputs a three-class probability vector $\mathbf{p}_i \in \Delta^2 = \{\mathbf{p} \in [0, 1]^3 : \sum_k p_k = 1\}$ and predicts $\hat{y}_i = \arg\max_k p_{i,k}$.

Our objective is to train the three branches so that their averaged decision keeps the accuracy gain of fusion while recovering the moderate-to-severe cases that the joint model misses. CBF achieves this in three stages (Figure 2): ***branch prediction*** (a), ***probability averaging*** (b), and ***collaborative fine-tuning*** (c). First, three branch predictors that see different inputs—the tongue image, the indicators, or both—each produce a complete three-grade distribution (Section 3.2). Second, the three distributions are averaged with fixed equal weights, which gives the direct-average predictor B (Section 3.3). Third, the branches are fine-tuned together with a loss on the averaged output and on each branch (Section 3.4). Section 3.5 derives the gradient of this loss, and Section 3.6 summarizes training and inference (Algorithm 1).

3.2 BRANCH PREDICTORS

Joint branch J. The joint branch is HM-TDF (Chen et al., 2025), which reads both the tongue image and the indicators. Its image encoder combines a convolutional network, pre-trained by rotation prediction, with Kolmogorov–Arnold (KAN) convolutions that produce an image feature vector; its indicator encoder is a KAN that maps the indicators to an indicator feature vector. The concatenated features feed two KAN classifiers: a feature-fusion classifier (FFC) with one output per grade, and a multi-expert classifier (MEC) whose binary experts are trained on ordinal groupings of the grades. At test time, HM-TDF uses the FFC prediction when its uncertainty is low and the MEC prediction otherwise. We keep this architecture and routing unchanged, with output $\mathbf{p}_i^J = f_J(x_i, s_i)$.

Single-input branches V and S. The tongue-only branch V uses the image encoder with a classification head and sets the indicator stream to zero; the physiology-only branch S uses the indicator encoder with a classification head and sets the image stream to zero. Both branches are trained separately on the same outer folds as J, and their probability outputs are

$$\mathbf{p}_i^V = f_V(x_i), \qquad \mathbf{p}_i^S = f_S(s_i), \qquad \mathbf{p}_i^V, \mathbf{p}_i^S, \mathbf{p}_i^J \in \Delta^2. \tag{2}$$

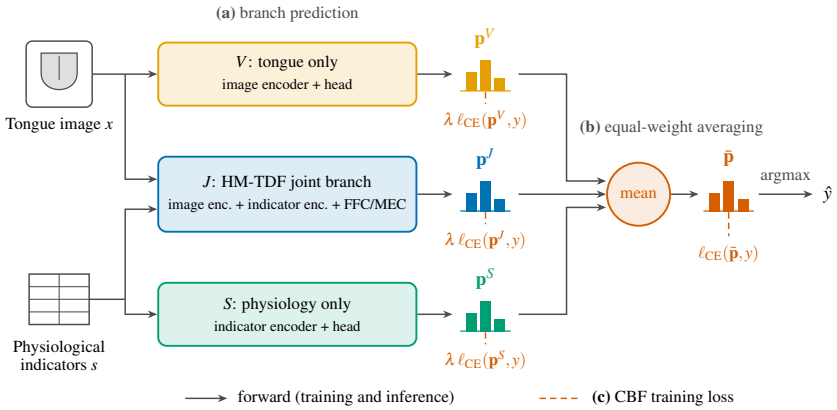


Figure 2: **Overview of CBF.** (a) Tongue and physiological inputs feed the tongue-only branch V, the physiology-only branch S, and the joint branch J (HM-TDF). (b) Their class probabilities are averaged with equal weights. (c) CBF fine-tunes all branches with cross-entropy on the averaged output and a λ-weighted cross-entropy on each branch (dashed red).

Thus V never reads s_i, S never reads x_i, and J reads the original pair. The three branches differ only in the inputs they see, and each produces a complete three-grade distribution rather than an intermediate feature selected for one particular grade.

3.3 Probability Averaging

We combine the three distributions by a fixed classwise mean:

$$\bar{\mathbf{p}}_i = \frac{\mathbf{p}_i^V + \mathbf{p}_i^S + \mathbf{p}_i^J}{3}. \tag{3}$$

The final label follows the argmax rule of Section 3.1. Equal weights keep this composition explicit: there is no learned gate, record-dependent weight, or grade-dependent routing, and the mean is a valid probability distribution because each branch output lies in Δ^2.

We call the direct average of the separately trained branches B and, after CBF training, the predictor F; the joint checkpoint alone is J. $J \to B$ isolates averaging and $B \to F$ isolates CBF training.

3.4 Collaborative Branch Fine-Tuning

Direct averaging produces a final prediction but does not train the branches for that prediction. Starting from the branch checkpoints used in B, CBF fine-tunes all three branches together with a loss on the averaged output and an auxiliary loss on each branch. For a training mini-batch $\mathcal{B}$, let $\ell_{\mathrm{CE}}(\mathbf{p}_i, y_i) = -\log p_{i,y_i}$. The objective is

$$\mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \Big[\ell_{\mathrm{CE}}(\bar{\mathbf{p}}_i, y_i) + \lambda \sum_{r \in \{V,S,J\}} \ell_{\mathrm{CE}}(\mathbf{p}_i^r, y_i) \Big], \qquad \lambda = 0.5. \tag{4}$$

The first term supervises the output used at inference, so each update depends on how the three probability distributions combine. The second term keeps each branch's own output under direct label supervision, so every branch remains a usable predictor on its own. Unlike gradient-modulation methods (Peng et al., 2022; Wei & Hu, 2024), CBF does not alter gradients; it places the loss directly on the decision used at deployment.

3.5 Gradient Analysis

To see how the averaged term couples the branches, let each branch output be a softmax over its logits, $\mathbf{p}_i^r = \mathrm{softmax}(\mathbf{z}_i^r)$ (for J, the logits of the head selected by the HM-TDF routing).

Algorithm 1 Collaborative Branch Fine-Tuning (CBF) for one outer fold

Require: training split $\mathcal{D}_{\text{tr}}$, validation split $\mathcal{D}_{\text{val}}$; trained branches f_V, f_S, f_J; weight λ; epochs E

$B(x,s) \leftarrow \frac{1}{3}\big(f_V(x) + f_S(s) + f_J(x,s)\big)$ ▷ direct average, no update

for $e = 1, \dots, E$ **do**

  for each mini-batch $\mathcal{B} \subset \mathcal{D}_{\text{tr}}$ **do**

    compute $\mathbf{p}_i^V$, $\mathbf{p}_i^S$, $\mathbf{p}_i^J$, and $\bar{\mathbf{p}}_i$ for all $i \in \mathcal{B}$

    update f_V, f_S, f_J jointly with $\nabla \mathcal{L}_{\mathcal{B}}$ from Eq. (4)

  end for

  $\ell_e \leftarrow \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i \in \mathcal{D}_{\text{val}}} \ell_{\text{CE}}(\bar{\mathbf{p}}_i, y_i)$ ▷ validation loss of the average

end for

keep the branches from epoch $\arg\min_e \ell_e$

return $F(x,s) = \arg\max_k \bar{p}_k(x,s)$

Proposition 1. *For a record i in mini-batch $\mathcal{B}$ and any branch $r \in \{V, S, J\}$, the gradient of the CBF objective (4) with respect to the logits of branch r is*

$$\nabla_{\mathbf{z}_i^r} \mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\big(w_i^r + \lambda\big)\big(\mathbf{p}_i^r - \mathbf{e}_{y_i}\big), \qquad w_i^r = \frac{p_{i,y_i}^r}{\sum_{s \in \{V,S,J\}} p_{i,y_i}^s}, \qquad \sum_{r \in \{V,S,J\}} w_i^r = 1. \tag{5}$$

Proof. Write $\bar{p}_{i,y} = \frac{1}{3}\sum_s p_{i,y}^s$ with $y = y_i$. Using the softmax Jacobian $\partial p_y / \partial z_k = p_y(\delta_{ky} - p_k)$, the averaged term gives $\partial(-\log \bar{p}_{i,y})/\partial z_{i,k}^r = -\frac{p_{i,y}^r}{3\bar{p}_{i,y}}(\delta_{ky} - p_{i,k}^r) = w_i^r\,(p_{i,k}^r - \delta_{ky})$, since $3\bar{p}_{i,y} = \sum_s p_{i,y}^s$. The per-branch term contributes $\lambda(p_{i,k}^r - \delta_{ky})$, and the other two branch terms do not depend on $\mathbf{z}_i^r$. Summing and averaging over $\mathcal{B}$ gives Eq. (5); the weights w_i^r sum to one by construction. □

Proposition 1 shows that CBF trains every branch with its ordinary cross-entropy gradient, $\mathbf{p}_i^r - \mathbf{e}_{y_i}$, rescaled per record by $w_i^r + \lambda$. The averaged term distributes a unit budget across the three branches in proportion to the probability each already assigns to the true grade: on a given record, a branch that carries the correct signal receives a larger share of the update than a branch that does not. Independent training corresponds to dropping w_i^r; training only the average corresponds to $\lambda = 0$, where a branch that is consistently wrong on some records can receive vanishing updates there—the learner-collusion failure mode described by Jeffares et al. (2023). The per-branch term guarantees every branch a minimum weight λ on every record, so CBF interpolates between the two regimes.

3.6 Training and Inference

Algorithm 1 summarizes the procedure for one outer fold. At inference, each branch is evaluated once and the three distributions are averaged; no test-time adaptation, class weighting, or grade-specific threshold is used. Hyperparameters are given in Section 4.2.

4 Experiments

We evaluate CBF at five levels of increasing granularity: comparison with existing methods, component ablation, per-grade analysis, branch-level analysis, and case-level evidence from Grad-CAM and tongue replacement (Sections 4.4–4.9).

4.1 Dataset and Evaluation Protocol

Tongue-FLD contains 5,717 paired tongue-image and physiological records with ultrasound-assessed three-grade labels (Chen et al., 2025). We aimed to evaluate on further datasets, but to our knowledge Tongue-FLD is currently the only public dataset that pairs tongue images and physiological indicators with graded FLD labels; all experiments are therefore conducted on it. The available five-fold predictions cover 5,715 records: 3,688 normal, 1,512 mild, and 515 moderate-to-severe. Each outer test fold contains 1,143 records. All protocol-matched methods, including

Table 1: **Comparison with existing methods on Tongue-FLD.** Methods are grouped by type and ordered by year (references in Section 4.2). P, R, F1: macro-averaged; IE/DE: HM-TDF image/indicator encoders. Best and second-best results are marked in red and blue.

	Method	Source	Encoder	Classifier	Acc.↑	P↑	R↑	F1↑	MAE↓	RMSE↓
Classic	SVM	MLJ'95	IE+DE	SVM	73.05	63.26	55.35	57.00	0.2957	0.5901
	XGBoost	KDD'16	IE+DE	XGBT	73.61	61.95	56.46	58.15	0.2899	0.5847
	Deep Forest	IJCAI'17	IE+DE	DF	70.20	58.87	55.33	56.14	0.3330	0.6346
Vision	ViT	ICLR'21	ViT	MLP	71.72	59.19	58.76	58.77	0.3088	0.6008
	Vim	ICML'24	Mamba	MLP	73.05	59.03	59.90	59.32	0.3018	0.6055
Tongue	Li et al.	JBI'21	ResNet50	GA-XGBT	74.98	61.23	55.83	57.11	0.2758	0.5716
	Wang et al.	JEthno'22	ResNet34	MLP	72.60	58.12	56.32	56.70	0.3097	0.6167
	Qiu et al.	BSPC'23	MobileNet	MLP	72.70	59.41	**60.06**	59.06	0.3024	0.6006
	Dai et al.	Heliyon'24	GAN	XGBT	73.05	62.15	55.67	57.38	0.2919	0.5801
	HM-TDF	MICCAI'25	IE+DE	FFC+MEC	74.72	63.44	56.57	58.44	0.2714	0.5554
CBF (ours)		–	IE+DE	3 branches	**75.92**	**64.20**	59.51	**61.30**	**0.2632**	**0.5549**

the reproduced HM-TDF, use aligned folds and the inherited HM-TDF pipeline, so comparisons are paired on the same records; two records absent from the saved predictions are excluded.

4.2 Implementation Details and Comparators

The joint branch J is the reproduced HM-TDF checkpoint, and V and S are the tongue-only and physiology-only branches. SV averages V and S; B averages J, V, and S without further updates. CBF initializes from these checkpoints, optimizes Eq. (4) with $\lambda = 0.5$ fixed for all folds, and yields F. Tongue inputs are 448×448 pixels. CBF uses AdamW (Loshchilov & Hutter, 2019) with learning rate and weight decay 10^{-4} for at most 15 epochs, using six-record microbatches and gradient accumulation toward 60 records. In each outer training fold, 4,114 records fit parameters and 458 select the checkpoint; test folds are never used for selection. We further compare with nine methods evaluated on Tongue-FLD by Chen et al. (2025), all of which use both tongue images and physiological indicators: four tongue-diagnosis methods (Wang et al., 2022; Dai et al., 2024; Li et al., 2021; Qiu et al., 2023), two image backbones (Dosovitskiy et al., 2021; Zhu et al., 2024), and three conventional classifiers (Cortes & Vapnik, 1995; Zhou & Feng, 2017; Chen & Guestrin, 2016).

4.3 Metrics

Following Chen et al. (2025), we report accuracy, macro-averaged precision, recall, and F1, and, because the grades are ordinal, the mean absolute error (MAE) and root mean squared error (RMSE) between predicted and true grade indices. Our metrics are calculated after pooling the five held-out folds, rather than by averaging fold-level values. For grade g, recall is TP_g/N_g; paired changes are *fixes* (wrong to correct) minus *breaks* (correct to wrong) for the same test records. Negative log-likelihood is $-\log p_{i,y_i}$ and is reported for tongue replacement. In all comparison tables, the best and second-best results are highlighted in red and blue, respectively.

4.4 Comparison with Existing Methods

Comparison protocol. Table 1 compares CBF with HM-TDF and nine methods whose results are reported by Chen et al. (2025); their evaluation details may differ slightly from our pooled evaluation. HM-TDF and CBF are run by us with the official code on the same folds. For HM-TDF we obtain 74.72% accuracy and 58.44% macro-F1 (75.24% and 62.37% originally reported).

Results. The existing methods form three tiers. Conventional classifiers on HM-TDF features reach 70.20–73.61% accuracy, and generic image backbones reach 71.72–73.05%. Tongue-diagnosis methods reach 72.60–74.98%, with Li et al. (2021) the most accurate, and HM-TDF attains the highest precision and the lowest MAE and RMSE among them, although its macro-F1 (58.44%) trails Vim (59.32%). CBF improves on all tiers: it achieves the best accuracy, macro-precision, macro-F1,

Table 2: **Ablation over inputs, averaging, and CBF training** on the same 5,715 pooled records.

		Components						Recall (%)		
Idx	Model	Tongue	Indic.	Avg.	CBF	Acc.	Macro-F1	Normal	Mild	Mod.-sev.
1	S: physiology only		✓			69.69	57.58	80.31	**55.82**	**34.37**
2	V: tongue only	✓				69.15	44.97	88.02	45.63	3.11
3	J: HM-TDF	✓	✓			74.72	58.44	89.70	55.16	24.85
4	SV: average of S, V	✓	✓	✓		75.01	58.31	**91.92**	50.93	24.66
5	B: average of S, V, J	✓	✓	✓		75.36	58.95	91.24	53.77	25.05
6	CBF (F, ours)	✓	✓	✓	✓	**75.92**	**61.30**	91.11	53.04	**34.37**

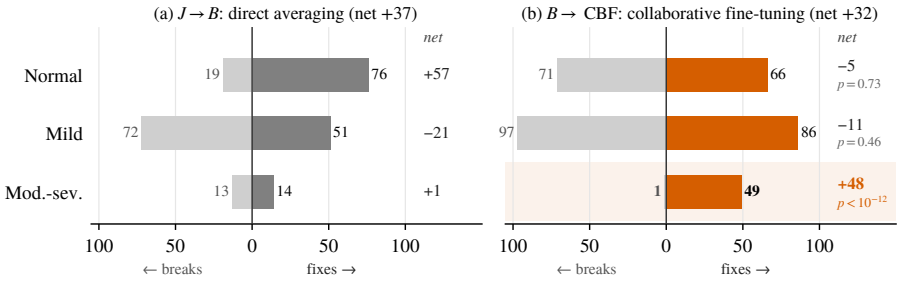


Figure 3: **Per-grade paired fixes and breaks.** Fixes (wrong to correct) point right and breaks (correct to wrong) point left, for $J \to B$ and $B \to F$.

MAE, and RMSE, and exceeds the strongest accuracy and macro-F1 of any existing method. Two methods have slightly higher macro-recall, but each has lower accuracy and macro-F1.

4.5 Ablation Study

Component analysis. Table 2 adds one component at a time. Comparing Idx 1–3 shows that fusion (J) is more accurate than either single-input branch but loses the moderate-to-severe recall of the physiology branch (24.85% versus 34.37%). Idx 4–5 show that averaging raises accuracy and normal-grade recall while leaving moderate-to-severe recall near 25%. Only CBF training (Idx 6) improves both, raising accuracy to 75.92% and macro-F1 to 61.30% while restoring moderate-to-severe recall. Correct predictions increase from 4,270 for J to 4,307 for B and 4,339 for CBF, i.e., 69 additional correctly classified records (+1.21 percentage points over J).

Consistency and significance. CBF exceeds J and B in accuracy in four of five folds. The net change in correctly classified moderate-to-severe records from B to CBF is positive in every fold (+5, +13, +10, +15, +5), including fold 1, where total accuracy drops slightly (fold accuracies of CBF: 76.55, 75.59, 77.17, 75.15, 75.15). The overall $B \to F$ difference is modest ($p = 0.107$, McNemar test) but concentrated in one grade: restricted to true moderate-to-severe records, $B \to F$ has 49 fixes and one break (exact McNemar $p < 10^{-12}$), whereas the changes on normal (66 fixes vs. 71 breaks, $p = 0.73$) and mild records (86 vs. 97, $p = 0.46$) are not distinguishable from chance. Direct averaging itself improves on J overall (141 fixes vs. 104 breaks, $p = 0.021$). Relative to J, CBF has 277 fixes and 208 breaks ($p = 0.002$).

4.6 Per-Grade Analysis

Paired fixes and breaks. Figure 3 locates the gains by grade. Direct averaging ($J \to B$) nets 57 additional correct normal predictions, loses 21 mild predictions, and gains only one moderate-to-severe prediction. CBF ($B \to F$) instead fixes 49 moderate-to-severe cases and breaks one, a net gain of 48. The cost is a net loss of five normal and 11 mild correct predictions.

Confusion matrices. The pooled confusion matrices (Figure 4) show where the high-grade recall returns. Correct moderate-to-severe predictions rise from 128 for J to 129 for B and 177 for CBF,

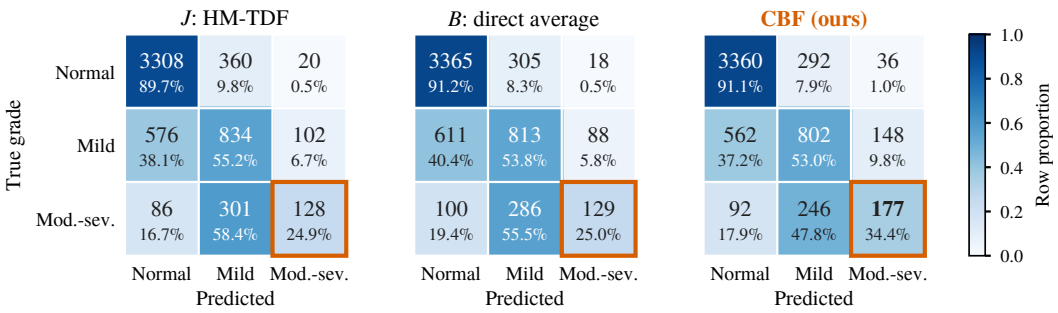


Figure 4: **Pooled confusion matrices.** Record counts for J, B, and CBF; colors show row-normalized proportions, and orange boxes mark correct moderate-to-severe predictions.

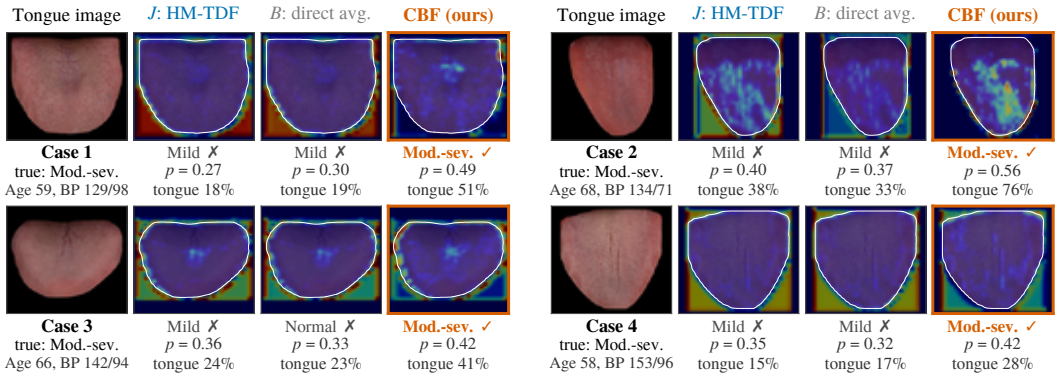


Figure 5: **Grad-CAM on recovered moderate-to-severe cases.** Four moderate-to-severe records that J and B misclassify and CBF classifies correctly. Below each map: predicted grade, moderate-to-severe probability p, and share of attribution inside the tongue (white outline).

while true high-grade cases assigned to mild fall from 301 to 286 to 246. CBF thus recovers many high-grade cases that the earlier models place in the adjacent mild category. For mild cases, predictions of normal fall from 611 (B) to 562 (CBF), consistent with heavier borderline predictions.

4.7 Branch-Level Analysis

Table 3: **Branch-level behavior before and after CBF.** Values are computed from saved per-branch probabilities on the same 5,715 records; mean p(mod.-sev.) is a branch's average moderate-to-severe probability on true moderate-to-severe records.

	Acc. (%)		Mod.-sev. recall (%)		Mean p(mod.-sev.)	
Branch	Before	After	Before	After	Before	After
V (tongue)	69.15	69.96	3.11	6.41	0.224	0.248
S (physiology)	69.69	70.18	34.37	37.48	0.374	0.390
J (joint)	74.72	73.81	24.85	38.45	0.307	0.387
Average (B/CBF)	75.36	75.92	25.05	34.37	0.302	0.341

Branch probabilities. Before fine-tuning (Table 3), on true moderate-to-severe records the joint branch assigns mean probability 0.307 to that grade and the tongue branch 0.224, below the physiology branch's 0.374, so the average carries a diluted 0.302. After CBF, the joint branch changes most: its recall rises from 24.85% to 38.45% and its mean probability to 0.387 (average: 0.341).

Division of labor. After CBF, the joint branch alone is slightly less accurate (73.81%) but more sensitive to the high grade than the average (38.45% vs. 34.37% recall); averaging with V and S tempers this shift into the best overall accuracy.

4.8 VISUALIZATION

We compute Grad-CAM (Selvaraju et al., 2017) for the moderate-to-severe grade on the last spatial layer of the tongue-image encoder; for B and CBF, the map of the averaged probability is summed over the V and J encoders, and each map is scaled to its 99th percentile. Figure 5 shows four of the 49 moderate-to-severe records that B misses and CBF recovers: Cases 1–2 have the largest increase in the averaged moderate-to-severe probability, and Cases 3–4 are drawn at random from the remaining records. Over the 49 records, J and B have nearly identical maps (median correlation 0.95) that place most attribution outside the tongue; after CBF, the median share inside the tongue rises from 22% to 41% and is higher than for J in 41 records.

4.9 PHYSIOLOGY-MATCHED TONGUE REPLACEMENT

We test dependence on the original tongue–indicator pairing by freezing J and F, retaining each recipient's indicators, and replacing their tongue image with a physiology-matched donor image. Donors share recorded sex, are close on the seven standardized continuous indicators, and are chosen without labels or prediction scores. Twenty fixed seeds choose among up to five nearest eligible donors, identically for both models.

Table 4: **Physiology-matched tongue replacement** on the shared 5,273-record eligible subset; replaced-image results are means over 20 donor assignments. Lower NLL is better.

Model	Own acc. (%)	Replaced acc. (%)	Own NLL	Replaced NLL
J	75.16	69.61	0.5882	0.7157
CBF (F)	76.41	68.42	0.6126	0.7110

Both models lose accuracy under replacement (Table 4), so they depend on the patient's own tongue–indicator pairing. The drop is larger for CBF (76.41% $\rightarrow$ 68.42%, -7.99 points) than for J (75.16% $\rightarrow$ 69.61%, -5.55 points). With the indicators held fixed, the larger drop shows that CBF relies more on patient-specific tongue information. At the same time, CBF's NLL increases less under replacement ($+0.098$ versus $+0.128$), so its probabilities degrade more gracefully.

5 DISCUSSION

CBF resolves the trade-off between overall accuracy and moderate-to-severe recall, and the analyses agree on why: the averaged term reallocates each update toward the branches carrying the correct signal (Proposition 1); the joint branch shifts most toward the high grade (Section 4.7), and Grad-CAM and tongue replacement both show greater reliance on the patient's own tongue image. More broadly, multimodal medical grading should report grade-specific recall alongside overall accuracy, so that missed high-grade patients are not hidden.

6 CONCLUSION

We studied tongue–physiology fatty liver grading and showed that a stronger fusion score can co-exist with missed moderate-to-severe cases, which averaging branch predictions at inference does not recover. CBF keeps three complete branch predictors and fine-tunes them for their averaged decision. It reaches 75.92% accuracy and 61.30% macro-F1, the best among all compared methods, restores moderate-to-severe recall to 34.37%, and gains a net 48 correctly classified high-grade cases over direct averaging, with a positive gain in every fold. The key is to train the branches for the decision that is actually deployed rather than for their own predictions alone. CBF needs no new architecture or learned gate and applies whenever separate image, tabular, and joint predictors are available. We plan to evaluate it on further tongue-diagnosis cohorts as they become available and on other image–tabular grading tasks.

Ethics Statement

This work adheres to the ICLR Code of Ethics. We use only the publicly available Tongue-FLD dataset (Chen et al., 2025), strictly for research purposes; no new patient data were collected, and the data contain no personally identifying information. The model is intended for research and is not a substitute for clinical diagnosis.

Reproducibility Statement

The dataset used in our experiments is publicly available. Implementation details and hyperparameters are given in Section 4.2, the evaluation metrics in Section 4.3, the training procedure in Algorithm 1, and the proof of Proposition 1 in Section 3.5. All source code and fold assignments will be released upon publication.

AI Use Statement

Generative AI tools were used to assist with literature searches and language polishing; the research design and experiments were independently conducted by the authors, who take full responsibility for the final manuscript.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Tao Chen, Jie Gao, Yong Xu, Weihong Qiu, Yijie Wu, Weimin Ye, and Kunhong Liu. Hard sample mining-based tongue diagnosis for fatty liver disease severity classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15974, pp. 260–269. Springer Nature Switzerland, 2025.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

S. Dai et al. Application of intelligent tongue image analysis in conjunction with microbiomes in the diagnosis of MAFLD. *Heliyon*, 10, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8632–8656. PMLR, 2023.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Multimodal Lego: Model merging and fine-tuning across topologies and modalities in biomedicine. In *International Conference on Learning Representations*, 2025.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *npj Digital Medicine*, 3(1):136, 2020. doi: 10.1038/s41746-020-00341-z.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259. PMLR, 2022.

Alan Jeffares, Tennison Liu, Jonathan Crabbé, and Mihaela van der Schaar. Joint training of deep ensembles fails due to learner collusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tao Jiang, Xiao-Jing Guo, Li-Ping Tu, Zhou Lu, Ji Cui, Xu-Xiang Ma, Xiao-Juan Hu, Xing-Hua Yao, Long-Tao Cui, Yong-Zhi Li, Jing-Bin Huang, and Jia-Tuo Xu. Application of computer tongue image analysis technology in the diagnosis of NAFLD. *Computers in Biology and Medicine*, 135:104622, 2021. doi: 10.1016/j.compbiomed.2021.104622.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

J. Li et al. A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. *Journal of Biomedical Informatics*, 115, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

D. Qiu et al. A novel tongue feature extraction method on mobile devices. *Biomedical Signal Processing and Control*, 80, 2023.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Rong-Rui Wang, Jia-Liang Chen, Shao-Jie Duan, Ying-Xi Lu, Ping Chen, Yuan-Chen Zhou, and Shu-Kun Yao. Noninvasive diagnostic technique for nonalcoholic fatty liver disease based on features of tongue images. *Chinese Journal of Integrative Medicine*, 30(3):203–212, 2024. doi: 10.1007/s11655-023-3616-1.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

X. Wang et al. Constructing tongue coating recognition model using deep transfer learning to assist syndrome diagnosis and its potential in noninvasive ethnopharmacological evaluation. *Journal of Ethnopharmacology*, 285, 2022.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52559–52572. PMLR, 2024.

Chenwei Wu, Zitao Shuai, Zhengxu Tang, Luning Wang, and Liyue Shen. Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In *International Conference on Learning Representations*, 2025.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055. PMLR, 2022.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4ab56e2e7ea64ff654d28bc0fcaee108-Abstract-Conference.html.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27456–27466, 2024.

Zhi-Hua Zhou and Ji Feng. Deep forest: Towards an alternative to deep neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.