CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

DualSS: A Dual-Diffusion Synthetic Sampling Framework for Screening Fatty Liver Severity With Class-Imbalanced Tongue Images

Tao Chen[1] | Jie Gao[2,3] | Yong Xu[4] | Weihong Qiu[2,3] | Kunhong Liu[5,6] | Yijie Wu[1] | Weimin Ye[2,7]

[1]State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, China | [2]NHC Key Laboratory of Etiological Epidemiology of Chronic Diseases With High Incidence in Fujian-Taiwan Area (Co-Construction), Department of Epidemiology and Health Statistics, School of Public Health, Fujian Medical University, Fuzhou, China | [3]Fujian Provincial Key Laboratory of Environment Factors and Cancer, Department of Epidemiology and Health Statistics, School of Public Health, Fujian Medical University, Fuzhou, China | [4]Xiamen Key Laboratory of Intelligent Fishery, Xiamen Ocean Vocational College, Xiamen, China | [5]Department of Digital Media, School of Film, Xiamen University, Xiamen, China | [6]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China | [7]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Correspondence: Kunhong Liu (lkhqz@xmu.edu.cn) | Yijie Wu (wyj1116@zju.edu.cn) | Weimin Ye (ywm@fjmu.edu.cn)

Received: 6 March 2026 | **Revised:** 5 June 2026 | **Accepted:** 5 August 2026

Keywords: fatty liver disease | latent diffusion model | physics-informed neural networks | synthetic sampling | tongue diagnosis

ABSTRACT

Fatty liver disease (FLD) is a prevalent chronic condition that can progress to clinically significant liver injury if left untreated. Given its high prevalence, FLD poses a substantial public health burden. Traditional Chinese medicine suggests that the appearance of tongue reflects hepatic and metabolic status, enabling artificial intelligence-based tongue diagnosis to provide rapid low-cost prescreening for FLD severity. In practice, the development of high-quality models is hindered by costly disease labelling, scarcity of positive cases, and severe class imbalance. In this work, we propose DualSS, a dual-diffusion synthetic sampling framework to mitigate these limitations. DualSS first trains a disease-agnostic dual-diffusion generator (DDG) on tongue-image data, leveraging abundant negative samples to learn unified tongue representations. Building on DDG, we design a latent space synthetic sampling pipeline that synthesises realistic positive-class images to construct a more balanced augmented training dataset. We then couple this pipeline with a noise-robust training strategy on the augmented dataset to further enhance the learning of downstream diagnostic models. Extensive experiments on FLD severity classification using tongue images and physiological indicators show that DualSS consistently outperforms existing rebalancing baselines, offering an effective and generalisable solution to alleviate positive sample scarcity in tongue image-based medical diagnosis.

1 | Introduction

The prevalence of fatty liver disease (FLD) among adults worldwide is estimated to be 25%–30% [1]. FLD may progress from simple steatosis to steatohepatitis and fibrosis, and in advanced cases to cirrhosis and hepatocellular carcinoma [2]. Greater FLD severity is associated with a higher risk of adverse outcomes. Thus, timely detection and intervention for moderate-to-severe FLD are critical [3]. Current clinical staging of FLD primarily relies on liver biopsy, ultrasound, and magnetic resonance imaging (MRI) [4]. However, their widespread deployment in

Tao Chen and Jie Gao contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

underserved regions is constrained by high equipment costs and the need for specialists [5]. From a population health perspective, broad preliminary screening facilitates earlier detection and increases the yield of subsequent clinical testing. Accordingly, there is an urgent need for a low-cost streamlined approach that supports rapid screening for FLD severity in primary care settings.

Tongue diagnosis, a core technique in traditional Chinese medicine (TCM), provides a conceptual basis for developing efficient screening strategies for FLD. It is a noninvasive method that infers systemic health status by examining tongue colour, coating, and morphology [6]. Population-based studies have associated tongue characteristics, including fissuring, tooth marks, and greasy coating, with metabolic abnormalities such as dyslipidemia, hypertension, and FLD. In FLD cohorts, tongue-coating colour and morphology have also been linked to oral-gut microbial profiles and syndrome-specific metabolic phenotypes [7]. Consistent with these findings, studies of FLD cohorts have reported a higher prevalence of purplish discolouration, greasy coating [8], red spots [9], fissuring, and tooth marks [10] in individuals with FLD. Because tongue images can be acquired inexpensively and conveniently, AI approaches grounded in tongue diagnosis are promising for scalable FLD risk assessment and severity prescreening.

Motivated by these needs, prior studies have incorporated tongue images into machine learning frameworks for binary FLD prescreening. Early efforts mainly focused on binary discrimination between Non-FLD and FLD using handcrafted tongue-image feature descriptors, conventional machine learning classifiers, or deep neural representations [6]. Recent studies have further advanced this research direction. Wang et al. [11] developed noninvasive NAFLD diagnostic models based on tongue-image features and anthropometric indicators. Gao et al. [12] proposed an explainable multi-source feature-fusion framework for FLD diagnosis. More recently, Chen et al. [13] extended tongue-image-based FLD modelling from binary diagnosis to multi-class severity classification using hard-sample mining. Collectively, these studies demonstrate the feasibility of intelligent tongue diagnosis for FLD screening.

However, because clinically verified annotations are costly, most prior studies have relied on datasets of limited size. Positive cases with pronounced pathological features are particularly scarce in clinical cohorts, which further limits model performance. Nevertheless, existing studies on intelligent tongue diagnosis have paid insufficient attention to this imbalance problem. This limitation becomes particularly critical when the target task shifts from binary FLD detection to FLD severity classification. Moderate-to-severe FLD cases often exhibit characteristic disease-related tongue manifestations, but they are substantially underrepresented in available tongue-image datasets [14]. Consequently, supervised classifiers tend to overfit to the limited positive cases. Conventional class-balancing strategies typically adjust sampling rates or loss-function weights, but they do not fundamentally alleviate the paucity of positive-class patterns. Existing GAN- or diffusion-based generative methods also remain inadequate because they generally require a sufficiently representative positive sample set to learn a stable class-conditional distribution. This requirement is inherently difficult to satisfy in settings characterised by positive-class scarcity. Therefore, a clear methodological gap remains between current intelligent tongue diagnosis methods and the need for high-fidelity positive-class-targeted data synthesis.

In this study, we treat modelling FLD severity from tongue images and physiological indicators as a representative case to address the above methodological gap. Figure 1 illustrates the task setting and the overall conceptual framework. Our goal is to mitigate positive-class scarcity and the resulting class imbalance by generating supplementary minority-class samples. However, as commonly encountered by existing generative methods, the scarcity of positive cases also prevents generative models from learning a reliable prior over positive-class phenotypes.

To address this limitation, we propose a dual-diffusion synthetic sampling framework (DualSS) for tongue diagnosis. Unlike purely class-conditional generative methods, DualSS decouples disease-agnostic tongue-pattern learning from positive-class-targeted synthetic sampling. Specifically, the generator is trained using all available tongue images, including both the limited positive samples and abundant negative samples. This design enables the model to learn shared tongue anatomy, coating patterns, and fine-grained surface textures from the entire dataset. We then perform intraclass synthesis to generate tongue images whose fine-grained texture patterns are aligned with the positive-class distribution. Through this design, DualSS leverages negative samples to compensate for the insufficient representation of positive-class patterns. It further constructs a more balanced and diagnostically informative augmented training set. After quality control, the generated positive samples are incorporated into the training set to rebalance the class distribution and improve predictive performance. The primary contributions of this work are as follows:

1. DualSS framework. We propose a synthetic sampling strategy specifically tailored to addressing class imbalance in tongue diagnosis datasets. Its key advantage is the ability to exploit abundant negative tongue images to learn a disease-agnostic generative prior, thereby supporting high-quality

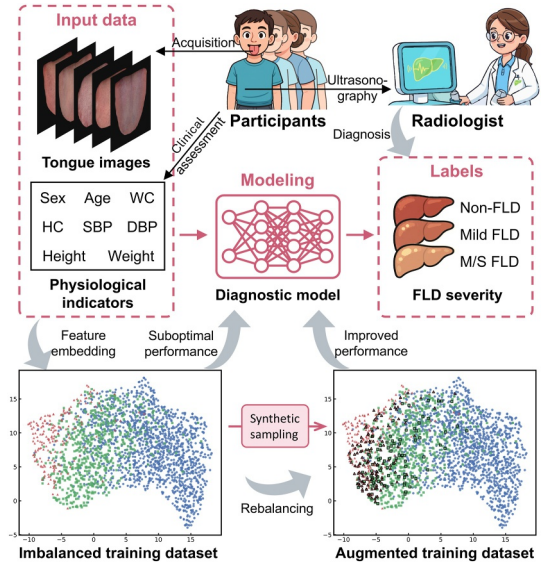


FIGURE 1 | Conceptual framework for FLD severity modelling from tongue images and physiological indicators, with synthetic sampling for class rebalancing.

positive-class synthesis even when clinically verified positive samples are limited.

2. Dual-diffusion generator. The model introduces physics-informed priors by imposing reaction-diffusion constraints in pixel space. It then employs context-aware diffusion in latent space to strengthen semantic consistency. Together, these design choices yield high-fidelity biologically plausible tongue-surface textures.
3. Latent space synthetic sampling. Building on the generator, we develop a synthetic sampling and quality control pipeline. The method synthesises positive samples that are perceptually realistic and distributionally aligned with disease-associated features of the target class, yielding a more balanced augmented dataset.
4. Noise-robust training strategy. Leveraging the augmented dataset, we implement a noise-robust training regimen that consistently improves diagnostic performance in FLD severity classification.

2 | Related Work

2.1 | Intelligent Tongue Diagnosis

Existing research on AI-based tongue diagnosis has made measurable progress on the binary task of detecting the presence or absence of FLD. However, because the correspondence between hepatic status and tongue characteristics is not one-to-one, models trained solely on tongue images under direct supervision often exhibit limited robustness [6].

To enhance diagnostic performance, prior studies have progressed along three principal directions: (i) integrating domain-specific medical priors to guide model design; (ii) combining manually engineered descriptors with deep neural representations; and (iii) fusing multi-source data for joint diagnosis. For example, guided by medical priors, Wang et al. [11] defined 16 structured labels for the tongue body and coating to discriminate FLD, achieving competitive performance while preserving interpretability. Dai et al. [15] introduced Tongue-GAN for tongue-image feature extraction and further combined the extracted features with 16S rRNA microbiome profiles from the tongue coating and the gut, yielding additional performance gains. Gao et al. [12] developed TIPNet, a neural network tailored for tongue-image feature extraction, and integrated it with basic physiological indicators to achieve high diagnostic accuracy.

In parallel, AI-assisted TCM diagnostics have also advanced rapidly beyond FLD. Recent studies have explored standardized tongue-image acquisition and automated analysis systems [16], multimodal and multi-label tongue-diagnosis models aligned with TCM theory [17], and home-based tongue-image monitoring for liver fibrosis [18]. These developments indicate that AI-based tongue diagnosis is moving towards standardized imaging, multimodal fusion, and clinically deployable systems.

Despite these advances, most existing methods focus on improving diagnostic performance through better representation learning or modality fusion. They do not directly address the lack of positive-class phenotypic diversity in the training data. This gap is particularly critical for FLD severity classification, where moderate-to-severe cases are much rarer than mild FLD or Non-FLD cases. Under such class imbalance, classifiers may learn decision boundaries dominated by majority classes. Meanwhile, the limited positive samples may be insufficient to represent the tongue characteristics associated with more severe FLD. Therefore, beyond improving discriminative modelling, a data-level strategy is needed to generate reliable positive-class samples and enable balanced learning while preserving the low-cost and noninvasive nature of tongue diagnosis.

2.2 | Imbalanced Classification and Synthetic Sampling

In the biomedical domain, class imbalance is pervasive and constitutes a central methodological challenge [19]. The scarcity of minority-class examples renders models more prone to majority-class bias, thereby compromising predictive performance and reliability [20]. This issue is particularly pronounced in medical image analysis. Positive cases are scarce and costly to obtain, yet they carry a disproportionate fraction of disease-relevant signal [21]. Existing approaches can be broadly divided into algorithm-level and data-level strategies. Algorithm-level methods, such as Focal Loss [22] and Class-Balanced Loss [23], modify the training objective by imposing greater penalties on minority-class errors. However, these methods do not increase the morphological diversity of minority-class samples. When only a few positive tongue images are available, loss reweighting alone may amplify the influence of these limited samples and thereby increase the risk of overfitting, rather than providing new positive-class patterns for model learning.

Data-level rebalancing methods attempt to alter the training distribution. Common approaches include minority oversampling, majority undersampling, and hybrid schemes that combine both [19]. Majority undersampling may discard informative negative examples and is therefore undesirable in medical settings where data collection is costly. Naive minority oversampling increases the probability of selecting positive samples but can easily lead to memorisation and overfitting. To mitigate this issue, the Synthetic Minority Oversampling Technique (SMOTE) [24] and its variants, such as Borderline-SMOTE [25] and Global-Local-based Oversampling [26], generate synthetic minority-class samples via local linear interpolation between minority instances and their nearest neighbours. Although such methods can smooth decision boundaries in low-dimensional tabular data, they are poorly suited to high-dimensional tongue images because linear interpolation may produce anatomically implausible patterns and ambiguous labels. Notably, DeepSMOTE [27] extends this idea by sampling in a learnt latent space, but fine-grained tongue-surface details and biologically plausible texture patterns remain difficult to preserve.

Generative augmentation has therefore attracted increasing attention [28]. GAN-based methods have been widely used for medical image synthesis, but their adversarial training is unstable and susceptible to mode collapse, especially when the target class contains only a few examples [29]. Diffusion models generally provide better sample fidelity and diversity [30], and recent studies have applied latent diffusion-based augmentation to imbalanced medical image classification [31]. However, class-conditional diffusion usually requires sufficient labelled samples

from each class to learn reliable conditional distributions [32]. Prior work has shown that minority-class generation may suffer from degraded fidelity and diversity when the training data are highly imbalanced [33]. In medical settings, this limitation is amplified because positive cases are not only rare but also clinically heterogeneous.

These limitations are particularly relevant to tongue-image-based FLD severity classification. Many tongue-surface textures, such as coating granularity, colour transitions, and local fissure patterns, are shared across severity levels, whereas the discriminative signal often lies in the global composition and co-occurrence of multiple tongue characteristics. Directly training a positive-class generator on the scarce moderate-to-severe cases is therefore likely to overfit the few seed samples. To address this problem, our method first learns a disease-agnostic tongue-image prior from all available images, including both positive and negative samples, to capture shared anatomical and textural regularities. It then performs controlled intraclass synthesis guided by positive seed exemplars. This design provides a clearer advantage conventional oversampling, SMOTE-like interpolation, GAN-based augmentation, and standard class-conditional diffusion by explicitly exploiting negative-class data to support positive-class synthesis under extreme scarcity.

2.3 | Physics-Informed Neural Networks

Physics-informed neural networks (PINNs) incorporate partial differential equations (PDEs) into the loss function to enforce physical consistency, thereby improving generalisation even under limited data conditions [34]. This property is particularly important in the biomedical domain, where obtaining high-quality annotated data is often prohibitively expensive [35]. Prior studies have shown that incorporating a linear-elastic model as a constraint in prostate image registration is effective [36]. Even with a limited number of training samples, this approach enables accurate cross-patient alignment. Moreover, physics-informed constraints can facilitate high-resolution reconstruction and preserve fine details. For example, in the super-resolution reconstruction of 4D-flow MRI, researchers have incorporated the Navier–Stokes equations together with boundary conditions, substantially mitigating the loss of fine details caused by low-resolution sampling and noise [37, 38]. Similarly, in brain haemodynamic prediction [39] and in multi-b-value diffusion MRI analysis [40], incorporating flow and diffusion equations markedly enhanced the fine-grained structural details and improved physiological consistency. Collectively, these studies demonstrate that PINNs provide a dual benefit in medical image analysis by simultaneously improving generalisation performance and physical fidelity.

Motivated by these insights, this study introduces reaction–diffusion (RD) [41] equations as a physical prior for tongue image generation. RD equations have long been used to model morphogenetic pattern formation in biological organisms. The proposed design aims to simulate the physiological mechanisms underlying tongue texture formation, thereby generating more realistic and detailed tongue images. This supports the synthesis of high-quality diagnostic training datasets.

3 | Material and Methods

3.1 | Data Collection and Preprocessing

This study uses Tongue-FLD, the largest publicly available tongue diagnosis dataset for assessing the severity of FLD. Data were obtained from a cohort study on etiological factors of cancer and other chronic diseases. The study was approved by the Ethics Review Committee of Fujian Medical University (approval numbers [2017-07] and [2020-58]). Participants were residents of Fuqing, Fujian Province, China.

During data acquisition, tongue images were collected by trained field investigators rather than by TCM clinicians. Before data collection, all operators underwent standardized training in participant positioning, tongue-protrusion instructions, and device operation. Each participant was asked to extend the tongue naturally, and a facial photograph containing the tongue region was captured using a standardized 'Four Diagnostic Device'. The device was equipped with an internal light source and enclosed housing, which together formed a controlled imaging chamber and minimised interference from ambient illumination. A built-in support structure was used to stabilise the participant's face during image capture. Together, the device-based controls and standardized operating procedures helped ensure consistency in illumination, acquisition angle, and facial positioning across participants.

The tongue region was then segmented from the facial photograph. The non-tongue background was masked and set to black, yielding standardized tongue images. Concurrently, eight basic physiological indicators were recorded: sex, age, height, weight, waist circumference (WC), hip circumference (HC), systolic blood pressure (SBP), and diastolic blood pressure (DBP). Subsequently, participants underwent abdominal ultrasonography, and radiologists assessed FLD severity based on the imaging findings.

In the final dataset, each sample comprises a standardized tongue image and eight basic physiological indicators as inputs, with FLD severity determined by abdominal ultrasonography serving as the label. The dataset contains 5717 samples: 3690 Non-FLD, 1512 Mild FLD, and 515 Moderate/Severe FLD (M/S FLD). This corresponds to a class-imbalance ratio of 7.17:2.94:1.00.

3.2 | Method

3.2.1 | *Task Description and Framework Overview*

This section formulates the problem and introduces the proposed dual-diffusion synthetic sampling framework (DualSS) for tongue diagnosis, as illustrated in Figure 2. We consider the Tongue-FLD dataset $\mathcal{D}_{\mathrm{FLD}} = \{\boldsymbol{x}_{\mathrm{I}}^{(i)}, \boldsymbol{x}_{\mathrm{P}}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{N}$. Here, $\boldsymbol{x}_{\mathrm{I}}^{(i)} \in \mathbb{R}^{H_0 \times W_0 \times 3}$ denotes the standardized tongue image, $\boldsymbol{x}_{\mathrm{P}}^{(i)} \in \mathbb{R}^{1 \times d}$ denotes the vector of basic physiological indicators, and $\boldsymbol{y}^{(i)} \in \{0, 1, 2\}$ is the FLD severity label (0: Non-FLD, 1: Mild FLD, 2: M/S FLD). N denotes the total number of samples. H_0 and W_0 denote the image height and width, respectively. d is the dimensionality of the physiological-indicator vector. In this study, the final diagnostic model is a multimodal model that jointly uses the standardized tongue image and the eight basic physiological indicators for FLD

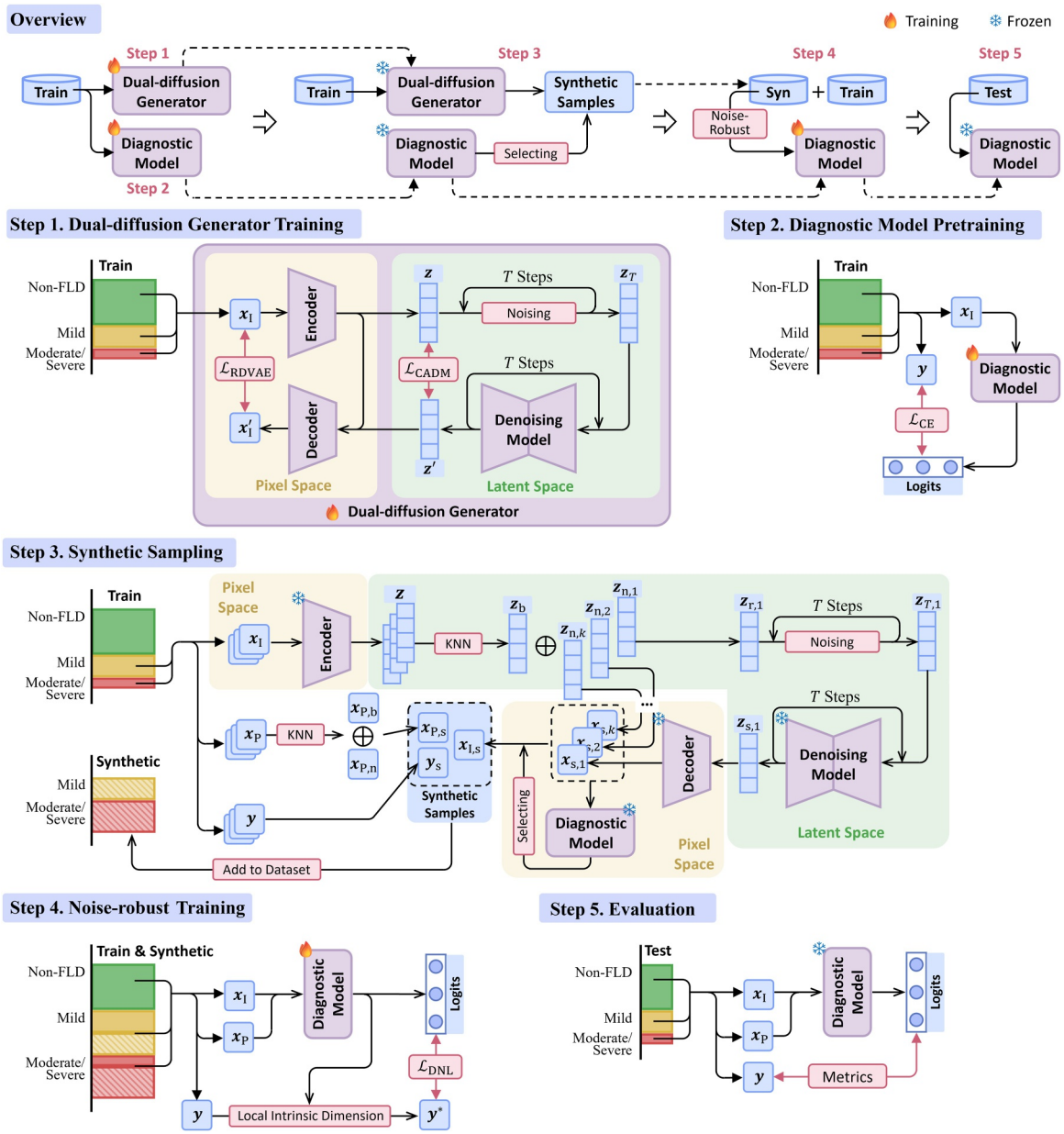


FIGURE 2 | The proposed synthetic sampling-based framework for tongue diagnosis.

severity prediction. We perform five-fold cross-validation. In each fold, $\mathcal{D}_{\text{FLD}}$ is split 4:1 into a training set $\mathcal{D}_{\text{Train}}$ and a test set $\mathcal{D}_{\text{Test}}$. The diagnostic model is trained on $\mathcal{D}_{\text{Train}}$ and evaluated on $\mathcal{D}_{\text{Test}}$ under single-pass inference, with the objective of maximising predictive performance.

Because the number of positive cases (Mild and M/S FLD) is substantially smaller than the number of Non-FLD cases, the dataset is severely imbalanced. Such skewed class prevalence is typical in population-level screening and is a major cause of performance degradation in conventional modelling methods. To alleviate this issue, we propose a Dual-diffusion Synthetic Sampling framework (DualSS). The framework employs a dual-diffusion generator (DDG) to synthesise samples for the underrepresented positive classes and construct an augmented dataset. This enables more balanced learning and improves diagnostic performance. Overall workflow of DualSS consists of the following steps.

1. DDG training. Train the DDG to learn a class-agnostic latent distribution on $\mathcal{D}_{\text{Train}}$ and reconstruction mapping for tongue images. To this end, all available tongue images from both positive and negative classes in $\mathcal{D}_{\text{Train}}$ are used for DDG training with ignoring the class labels of images. And images in $\mathcal{D}_{\text{Test}}$ are kept from the training process.
2. Diagnostic model pretraining. Pretrain a diagnostic model on tongue images to learn representations related to FLD severity and to produce class-probability estimates. The diagnostic model is instantiated as an image classifier that has demonstrated strong performance in prior tongue diagnosis studies.
3. Synthetic sampling. Use the DDG to synthesise candidate samples for the minority classes in $\mathcal{D}_{\text{Train}}$. Filter these candidates according to the confidence scores of the pretrained diagnostic model to construct the synthetic dataset $\mathcal{D}_{\text{Syn}}$.

4. Noise-robust training. Combine $\mathcal{D}_{\text{Syn}}$ with $\mathcal{D}_{\text{Train}}$ to form an augmented dataset and train the final diagnostic model. Because synthesis may introduce mislabelled samples, we assess label reliability using the local intrinsic dimension and adopt a noise-robust training strategy accordingly.
5. Evaluation. Evaluate the trained diagnostic model on $\mathcal{D}_{\text{Test}}$ for FLD severity prediction.

The following sections detail the DDG architecture and training configuration, the synthetic sampling strategy, and the implementation of the noise-robust training procedure. Additional implementation-oriented pseudocode for the complete DualSS workflow can be found in Supporting Information S1: Section SA.

3.2.2 | *Dual-Diffusion Generator*

The proposed dual-diffusion generator (DDG) comprises two main modules: a Reaction–Diffusion Variational Autoencoder (RD-VAE) and a Context-Aware Diffusion Model (CA-DM), as illustrated in Figure 3. The RD-VAE maps each input image $\boldsymbol{x}_{\text{I}}$ to a latent feature $\boldsymbol{z} \in \mathbb{R}^{h \times w \times c}$ via an encoder $\mathcal{E}$. Here, h, w, and c denote the height, width, and number of channels of the latent feature, respectively. The latent feature is vector-quantized [42] using a discrete codebook $\{e_j\}_{j=1}^{K_{\text{code}}}$, where each code vector $e_j \in \mathbb{R}^c$ and K_{code} is the codebook size. This quantization maps continuous representations to a finite set of discrete codes, thereby reducing redundancy in the latent space. A decoder $\mathcal{D}$ then reconstructs $\hat{\boldsymbol{x}}_{\text{I}}$ from $\boldsymbol{z}$, with reaction–diffusion constraints imposed to regularise the reconstruction. The CA-DM takes noisy latent features as input and progressively removes noise through a context-aware diffusion process.

As a generative model, the DDG must achieve both high-fidelity reconstruction and strong sampling capability. However, these two objectives are inherently in tension [43], reflecting a trade-off between the information capacity of the latent representation and its alignment with the prior distribution. This trade-off is particularly sensitive to the choice of latent dimensionality. If $\boldsymbol{z}$ is too low in dimension, excessive compression leads to information loss and degraded reconstruction quality; if it is too high in dimension, enforcing a well-structured latent space becomes difficult, which in turn impairs sampling efficiency and quality. To overcome this, following prior work [44], we apply three

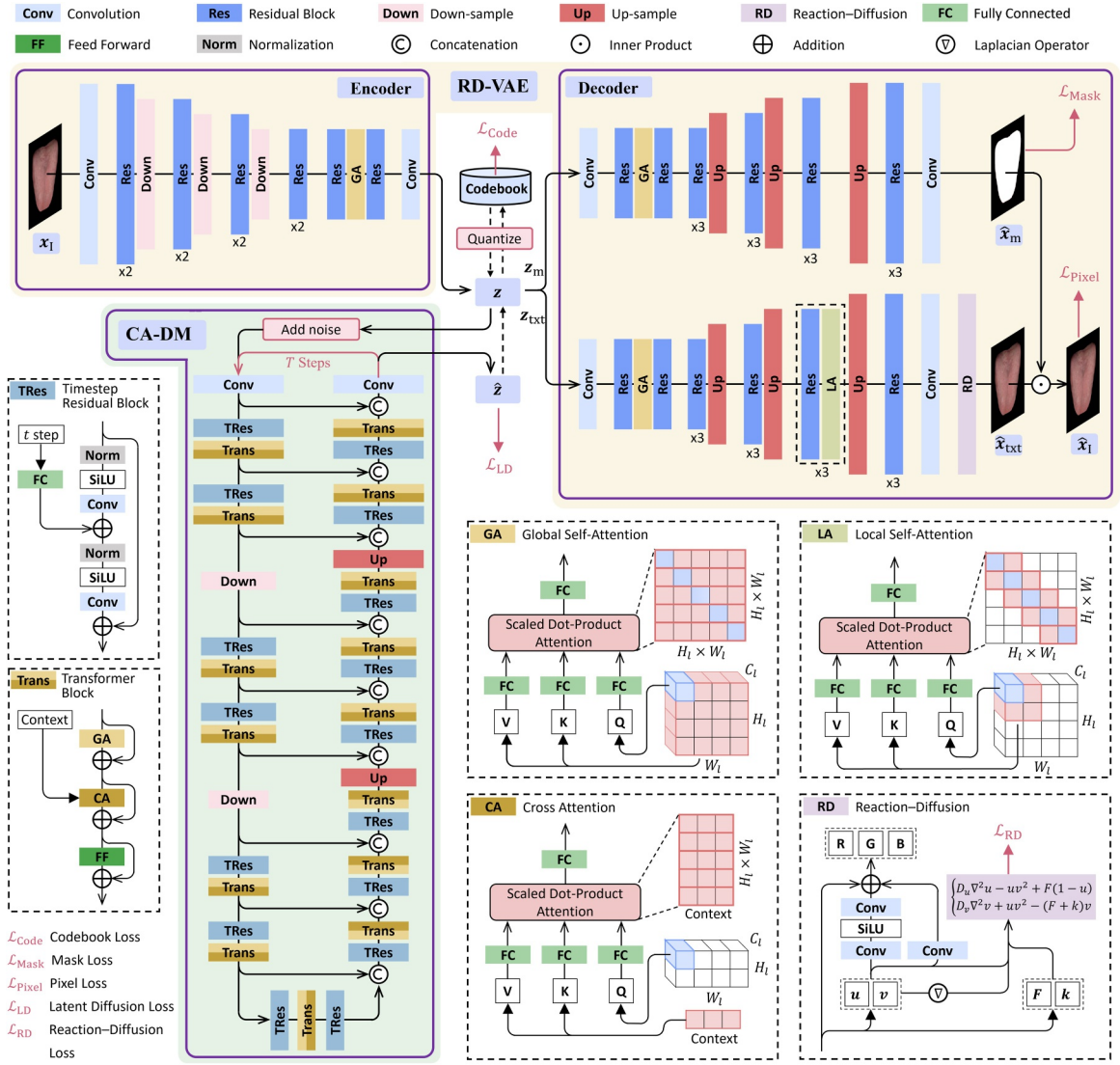


FIGURE 3 | The detailed structure of the proposed dual-diffusion generator.

downsampling stages in the encoder. This design helps maintain a well-structured latent space while avoiding excessive information loss. We then incorporate several architectural enhancements into the RD-VAE to improve tongue image reconstruction quality. In parallel, we enhance sampling quality through context-aware diffusion within the CA-DM. The following subsections present the architectural details of the RD-VAE and the CA-DM.

3.2.2.1 | Texture–Shape Decoupled Generation Mechanism. We incorporate medical prior knowledge into the design of the RD-VAE. Studies have shown that tongue coating, colour, size, and contour are important indicators of disease [6, 45]. These tongue characteristics can be grouped into two categories: local texture and global shape. FLD-associated tongue signs, such as greasy coating, fissures, tooth marks, and localised colour changes, involve both fine-grained surface patterns and global morphological variations. Preserving these cues requires the generator to model subtle local textures while preventing global foreground structures from dominating the reconstruction process.

A conventional single-decoder generator must learn local texture and global shape through the same reconstruction pathway. This coupled learning process may be suboptimal for FLD-related tongue images because diagnostically relevant texture patterns are often subtle and spatially heterogeneous. To address this limitation, the proposed RD-VAE adopts a dual-decoder architecture that explicitly decouples the generation of texture and shape. Specifically, the latent feature is partitioned along the channel dimension into two components, $\boldsymbol{z}_{\mathrm{m}} \in \mathbb{R}^{h\times w\times c_{\mathrm{m}}}$ and $\boldsymbol{z}_{\mathrm{txt}} \in \mathbb{R}^{h\times w\times c_{\mathrm{txt}}}$. These components are independently decoded by the shape decoder $\mathcal{D}_{\mathrm{m}}$ and the texture decoder $\mathcal{D}_{\mathrm{txt}}$ to produce a shape mask $\widehat{\boldsymbol{x}}_{\mathrm{m}} \in \mathbb{R}^{H_0\times W_0\times 1}$ and a texture map $\widehat{\boldsymbol{x}}_{\mathrm{txt}} \in \mathbb{R}^{H_0\times W_0\times 3}$, respectively. The final reconstruction is obtained as

$$\widehat{\boldsymbol{x}}_{\mathrm{I}} = \widehat{\boldsymbol{x}}_{\mathrm{txt}} \odot \widehat{\boldsymbol{x}}_{\mathrm{m}}, \tag{1}$$

where $\odot$ denotes element-wise multiplication.

This explicit texture–shape disentanglement mechanism offers important advantages for standardized tongue images. In our preprocessing pipeline, the non-tongue background is masked and filled with black. As a result, the boundary between the tongue foreground and the background exhibits abrupt intensity transitions. In a single-decoder generator, these high-gradient boundary regions must be reconstructed together with fine-grained surface textures. The resulting shape-related gradients may interfere with local texture learning, leading to artefacts and blurred coating patterns.

By contrast, the proposed dual-decoder design separates these two learning objectives. During training, the shape mask blocks gradients from the background region. As a result, the texture decoder $\mathcal{D}_{\mathrm{txt}}$ focuses on modelling fine-grained details within the tongue region rather than fitting abrupt pixel transitions at the tongue-background boundary. This improves training stability and enhances the textural quality of the generated images.

3.2.2.2 | Global and Local Self-Attention. Attention mechanisms are powerful tools for enhancing representational capacity and have been widely adopted in autoencoder architectures. Among these, global self-attention (GA) is one of the most commonly used formulations. GA treats the input feature map $\boldsymbol{F}_l \in \mathbb{R}^{H_l\times W_l\times C_l}$ as H_lW_l tokens, each of dimension C_l. Global information exchange among all tokens is then achieved through the standard query-key-value (QKV) attention mechanism [46]. Ignoring the number of attention heads and constant factors, its computational complexity is $\mathcal{O}\big((H_lW_l)^2C_l\big)$. In deeper layers near the latent bottleneck, H_lW_l is relatively small, so the cost of GA is acceptable and high-level semantic representations benefit from global information exchange. However, in shallow decoder layers close to the output, H_lW_l becomes substantially larger, and the computational and memory costs of GA increase sharply. Moreover, fine-grained texture details primarily depend on local spatial correlations, so the benefit of full global interaction in these layers is limited.

Therefore, the proposed RD-VAE applies GA in deep layers to enhance semantic representations. In the shallow layers of the texture decoder $\mathcal{D}_{\mathrm{txt}}$, it employs local self-attention (LA) to improve texture synthesis quality while reducing computational cost. Similar to GA, LA treats $\boldsymbol{F}_l$ as H_lW_l tokens of dimension C_l. However, each token attends only to tokens within a local window of size $K_l \times K_l$ centred on itself, where $K_l \ll H_l$ and $K_l \ll W_l$. In this case, the computational complexity is reduced to $\mathcal{O}\big(K_l^2H_lW_lC_l\big)$, which is substantially lower than that of GA. This hierarchical attention scheme effectively balances global semantic modelling in deeper layers and local texture synthesis in shallower layers, while maintaining low computational and memory costs.

3.2.2.3 | Reaction–Diffusion Driven Physics-Informed Modelling. Inspired by mechanisms of biological surface pattern formation, we hypothesise that tongue body and coating textures can be approximated by the Grey–Scott reaction–diffusion model (GSRD) [47]. The GSRD has been widely used to simulate diverse biological surface patterns, including spots, stripes, and fissures. Its governing equations are given by

$$\begin{cases} \dfrac{\partial u}{\partial t} = D_u\nabla^2 u - uv^2 + F(1-u) \\ \dfrac{\partial v}{\partial t} = D_v\nabla^2 v + uv^2 - (F+k)v \end{cases}, \tag{2}$$

where $u(x,t)$ and $v(x,t)$ denote the concentrations of two chemical species. D_u and D_v are the corresponding diffusion coefficients. ∇^2 is the Laplacian operator. F represents the feed rate of species u. k denotes the removal rate of species v due to decay and reaction.

To further improve the quality of the texture decoder $\mathcal{D}_{\mathrm{txt}}$, we incorporate the Grey–Scott model as an additional physics-informed constraint. Specifically, we design a reaction–diffusion (RD) block and embed it into $\mathcal{D}_{\mathrm{txt}}$. Unlike conventional decoders that directly predict RGB images, $\mathcal{D}_{\mathrm{txt}}$ first outputs four spatial fields $[u, v, F, k]$ with the same resolution as the target image. It then maps the predicted u and v fields to RGB space using residual convolutional layers, whereas F and k are primarily used to impose physics-informed constraints. Because tongue texture is quasi-static, we assume a steady state condition with $\partial u/\partial t = \partial v/\partial t = 0$. Under this assumption, the right-hand sides of Equation (2) should be close to zero at every spatial location,

imposing an additional constraint on the predicted fields $[u, v, F, k]$. This constraint encourages the predicted physical fields to satisfy reaction–diffusion equilibrium, thereby yielding textures with improved physical plausibility and perceptual fidelity.

3.2.2.4 | *Context-Aware Diffusion in Latent Space.* To enhance the global consistency and semantic alignment of the latent features, we incorporate a Context-Aware Diffusion Model (CA-DM) into the DDG. Here, 'context-aware' refers to a conditional latent diffusion process guided by a learnable context vector. Formally, CA-DM is defined as a conditional denoising function that estimates the noise added during the forward diffusion process, given a noisy latent feature, the diffusion step, and a global context anchor. The model follows the standard forward and reverse diffusion processes of the denoising diffusion probabilistic model (DDPM) [30]. In the forward process, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is gradually added to the latent feature:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_{t-1} + (1-\alpha_t)\boldsymbol{\epsilon}, t = 1, 2, ..., T, \tag{3}$$

where t denotes the diffusion step, T is the total number of diffusion steps, $\mathbf{z}_t$ is the latent feature at step t, α_t is a predefined noise decay coefficient, and $\boldsymbol{I}$ is the identity matrix. During the reverse diffusion process, CA-DM predicts the noise conditioned on the diffusion step t:

$$\widehat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_{\mathrm{CADM}}(\mathbf{z}_t, t, \boldsymbol{c}), \tag{4}$$

here, the context vector $\boldsymbol{c} \in \mathbb{R}^{512}$ serves as a trainable global semantic anchor that is shared across samples and optimised jointly with the denoising network. This vector imposes a global contextual constraint, which encourages latent features to remain semantically aligned during reverse diffusion. Starting from $\mathbf{z}_T$, we apply T reverse diffusion steps using the predicted noise $\widehat{\boldsymbol{\epsilon}}_t$, following the update rule in Equation (5), to obtain the reconstructed latent feature $\widehat{\mathbf{z}}$:

$$\widehat{\mathbf{z}} = \mathrm{Denoise}_{\mathrm{CADM}}(\mathbf{z}_T, T, \mathbf{c}). \tag{5}$$

In implementation, the CA-DM adopts a U-Net backbone and alternates Timestep Residual (TRes) Blocks and Transformer (Trans) Blocks across its layers, as illustrated in Figure 3. Each TRes block embeds the timestep t via a fully connected layer and injects this embedding into the intermediate features of the residual block. Each Trans block incorporates global self-attention for long-range interactions and cross-attention to integrate the context information. With this design, the CA-DM exploits global dependencies and contextual constraints throughout the reverse diffusion process, thereby enhancing the structural and semantic consistency of the generated latent features.

3.2.2.5 | *Training Configuration.* The DDG is trained in two stages. We first optimise the RD-VAE and then the CA-DM.

We train the RD-VAE in an adversarial manner with a combination of loss terms. During optimization, the PatchGAN [48] discriminator and the RD-VAE are updated alternately. The PatchGAN discriminator adopts a fully convolutional architecture and is trained to distinguish real images from those reconstructed by the RD-VAE. It is optimised with the following adversarial loss:

$$\mathcal{L}_{\mathrm{Adv,Dis}} = \frac{1}{2}\mathbb{E}_{\widehat{\boldsymbol{x}}_{\mathrm{I}}}\left[(\boldsymbol{\varphi}(\widehat{\boldsymbol{x}}_{\mathrm{I}}))^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{\mathrm{I}}}\left[(\boldsymbol{\varphi}(\boldsymbol{x}_{\mathrm{I}}) - 1)^2\right], \tag{6}$$

here, $\boldsymbol{\varphi}(\cdot)$ denotes the patch-wise score map produced by PatchGAN, where each element is regressed towards 1 for real images and 0 for reconstructed images. The RD-VAE is optimised with the following adversarial loss:

$$\mathcal{L}_{\mathrm{Adv}} = \frac{1}{2}\mathbb{E}_{\widehat{\boldsymbol{x}}_{\mathrm{I}}}\left[(\boldsymbol{\varphi}(\widehat{\boldsymbol{x}}_{\mathrm{I}}) - 1)^2\right], \tag{7}$$

which encourages reconstructed images to be classified as real by PatchGAN and thereby enhances the realism of fine-grained texture details. In addition, RD-VAE training incorporates multiple auxiliary loss terms. For latent feature quantization, we adopt the standard codebook and commitment loss defined as

$$\mathcal{L}_{\mathrm{Code}} = \|\mathrm{sg}[\boldsymbol{z}] - \boldsymbol{e}_{k^*}\|_2^2 + \beta\|\boldsymbol{z} - \mathrm{sg}[\boldsymbol{e}_{k^*}]\|_2^2, \tag{8}$$

where $\mathrm{sg}[\cdot]$ denotes the stop gradient operator, β is a weighting coefficient set to 0.25, and $\boldsymbol{e}_{k^*}$ is the nearest code vector in the codebook with index k^*. The shape mask is learnt using an L2 loss that minimises the mean squared error between the predicted mask $\widehat{\boldsymbol{x}}_{\mathrm{m}}$ and the ground truth mask:

$$\mathcal{L}_{\mathrm{Mask}} = \mathbb{E}_{\boldsymbol{x}_{\mathrm{m}}, \widehat{\boldsymbol{x}}_{\mathrm{m}}}\left[\|\boldsymbol{x}_{\mathrm{m}} - \widehat{\boldsymbol{x}}_{\mathrm{m}}\|_2^2\right], \tag{9}$$

where $\boldsymbol{x}_{\mathrm{m}}$ is the binarised mask of the input image $\boldsymbol{x}_{\mathrm{I}}$ with background pixels set to 0 and tongue-region pixels set to 1. Texture generation is supervised by a pixel-wise reconstruction loss and a perceptual consistency loss:

$$\mathcal{L}_{\mathrm{Pixel}} = \mathbb{E}_{\boldsymbol{x}_{\mathrm{I}}, \widehat{\boldsymbol{x}}_{\mathrm{I}}}\left[\|\boldsymbol{x}_{\mathrm{I}} - \widehat{\boldsymbol{x}}_{\mathrm{I}}\|_2^2\right] + \mathcal{L}_{\mathrm{per}}(\boldsymbol{x}_{\mathrm{I}}, \widehat{\boldsymbol{x}}_{\mathrm{I}}), \tag{10}$$

where

$$\mathcal{L}_{\mathrm{per}} = \mathbb{E}_{\boldsymbol{x}_{\mathrm{I}}, \widehat{\boldsymbol{x}}_{\mathrm{I}}}\left[\sum_i \left\|\boldsymbol{\psi}_i(\boldsymbol{x}_{\mathrm{I}}) - \boldsymbol{\psi}_i(\widehat{\boldsymbol{x}}_{\mathrm{I}})\right\|_2^2\right], \tag{11}$$

and $\boldsymbol{\psi}_i$ denotes the feature map from the i-th pooling layer of a pretrained VGG16 network. This term encourages semantic consistency and improves the fidelity of the generated textures. To incorporate the Grey–Scott physics prior, we define the reaction–diffusion loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{RD}} = \mathbb{E}_{u,v,F,k}\Big[&\left\|D_u\nabla^2 u - uv^2 + F(1-u)\right\|^2 \\ &+ \left\|D_v\nabla^2 v + uv^2 - (F+k)v\right\|^2\Big],\end{aligned} \tag{12}$$

where the Laplacians $\nabla^2 u$ and $\nabla^2 v$ are computed using second-order finite differences, and D_u and D_v are globally shared learnable diffusion coefficients. The overall RD-VAE training objective is

$$\mathcal{L}_{\mathrm{RDVAE}} = \mathcal{L}_{\mathrm{Adv}} + \mathcal{L}_{\mathrm{Code}} + \mathcal{L}_{\mathrm{Mask}} + \mathcal{L}_{\mathrm{Pixel}} + \mathcal{L}_{\mathrm{RD}}. \tag{13}$$

The CA-DM is trained with a mean squared error objective for noise prediction:

$$\mathcal{L}_{\mathrm{CADM}} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{\mathrm{CADM}}(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{14}$$

where $\boldsymbol{\epsilon}_t$ denotes the true Gaussian noise added at diffusion step t in the forward process.

Together, these objectives guide the DDG to learn physically grounded and semantically consistent tongue-image patterns that support high-quality synthetic sample generation.

3.2.3 | *Latent Space Synthetic Sampling*

Leveraging the class-agnostic tongue-image patterns learnt by the DDG, we propose a latent space synthetic sampling method to generate additional minority-class samples. Because the original dataset is multimodal and the final diagnostic model jointly uses tongue images and physiological indicators, each synthetic training instance is constructed as a multimodal sample. Specifically, it consists of a synthetic tongue image, a paired indicator vector, and the corresponding class label. In this procedure, the DDG is used only for tongue-image synthesis, whereas the paired indicator vector is generated through local interpolation between real samples from the same class.

We first describe the synthetic sampling procedure for tongue images. Let $\{\boldsymbol{x}_{\mathrm{I}}\}$ denote the set of tongue images from the target class, and let $\{\boldsymbol{z}\}$ denote their latent features obtained by mapping $\{\boldsymbol{x}_{\mathrm{I}}\}$ through the encoder $\mathcal{E}$. For a selected base sample with latent feature $\boldsymbol{z}_b$, we search its K nearest neighbours in the latent space, denoted $\{\boldsymbol{z}_{n,k}\}_{k=1}^{K}$. Each candidate latent feature is obtained by fusing $\boldsymbol{z}_b$ with its kth neighbour:

$$\boldsymbol{z}_{r,k} = \rho_{\mathrm{I}}\boldsymbol{z}_{\mathrm{b}} + (1-\rho_{\mathrm{I}})\boldsymbol{z}_{n,k}, \tag{15}$$

where $\rho_{\mathrm{I}} \sim \mathcal{U}(0,1)$ is a random mixing coefficient. Because the latent space may exhibit nonlinear distortions, direct linear fusion can introduce artefacts. We treat this deviation as noise and refine each candidate via the CA-DM diffusion process:

$$\boldsymbol{z}_{\mathrm{s},k} = \mathrm{Denoise}_{\mathrm{CADM}}(\boldsymbol{z}_{r,k}, T_{\mathrm{s}}, \boldsymbol{c}), \tag{16}$$

where T_{s} denotes the number of diffusion steps used for sampling. This process is approximately equivalent to searching for latent features with higher posterior probability in the neighbourhood of $\boldsymbol{z}_{r,k}$. It thereby reduces the artefacts introduced by linear fusion. Figure 4 visualises the latent feature fusion and denoising refinement procedures in both the latent space and the pixel space. The refined latent feature $\boldsymbol{z}_{s,k}$ is then decoded to obtain the corresponding synthetic tongue image $\boldsymbol{x}_{s,k}$. Next, we feed each $\boldsymbol{x}_{s,k}$ into the pretrained diagnostic model and, for each base sample, select the candidate with the highest predicted probability for the target class as the final synthetic tongue image $\boldsymbol{x}_{I,s}$. This selection step ensures that $\boldsymbol{x}_{I,s}$ exhibits tongue characteristics strongly indicative of the corresponding FLD severity.

When synthesising the basic physiological indicators, continuous variables are first normalised using the statistics of the training set, and categorical variables are converted into one-hot encoded representations. For each selected base sample with indicator vector $\boldsymbol{x}_{P,b}$, we then identify its same-class neighbours and randomly select one neighbour $\boldsymbol{x}_{P,n}$. A new indicator vector is then generated by linear interpolation:

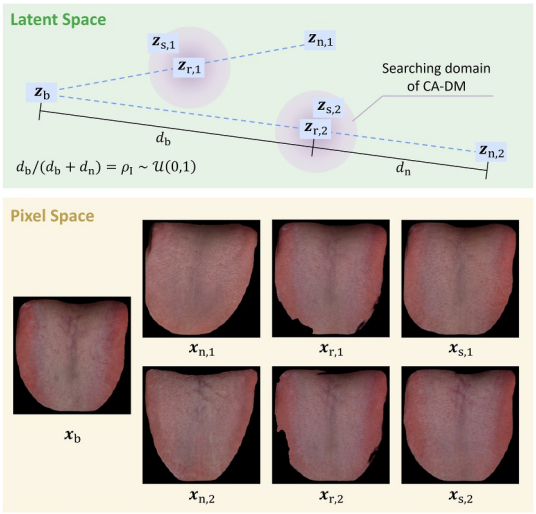


FIGURE 4 | Visualisation of the image synthetic sampling process in latent and pixel spaces.

$$\boldsymbol{x}_{P,s} = \rho_{\mathrm{P}}\boldsymbol{x}_{P,b} + (1-\rho_P)\boldsymbol{x}_{P,n}, \tag{17}$$

where $\rho_P \sim \mathcal{U}(0,1)$ is a random mixing coefficient. Each synthetic sample is then constructed by pairing the generated tongue image $\boldsymbol{x}_{I,s}$ and the synthetic physiological indicator vector $\boldsymbol{x}_{P,s}$ with the corresponding class label. To mitigate class imbalance, we perform synthetic sampling only for classes other than class 0 (Non-FLD), which is the majority class. For each minority class j, the number of synthetic samples is given by

$$n_{\mathrm{s},j} = \gamma(n_0 - n_j), \tag{18}$$

here, n_0 and n_j denote the original training data sizes of class 0 and class j, respectively, and γ denotes the sampling rate coefficient that governs the extent of rebalancing. When $\gamma = 0$, no additional samples are generated. When $\gamma = 1$, sampling continues until each minority class is fully balanced with class 0. The resulting synthetic samples form the synthetic dataset $\mathcal{D}_{\mathrm{Syn}}$.

One advantage of this method is that interpolation is performed only between samples of the same class, which encourages the synthetic samples to inherit class-consistent features. In addition, screening with the pretrained diagnostic model further reduces the risk of assigning incorrect labels. Moreover, the CA-DM-based diffusion refinement reduces artefacts caused by latent space nonlinearities, thereby producing perceptually reliable synthetic tongue images.

3.2.4 | *Noise-Robust Training With the Augmented Dataset*

We construct the augmented dataset $\mathcal{D}_{\mathrm{Aug}}$ by combining the original training set and the synthetic set:

$$\mathcal{D}_{\mathrm{Aug}} = \mathcal{D}_{\mathrm{Train}} \cup \mathcal{D}_{\mathrm{Syn}}, \tag{19}$$

We then train a diagnostic model on $\mathcal{D}_{\mathrm{Aug}}$ to predict FLD severity by jointly exploiting tongue images and basic physiological indicators. Although candidate samples are filtered by predicted confidence during synthetic sampling, some synthesised samples may still be mislabelled. Therefore, we treat training as learning

with noisy labels and do not fully trust the assigned labels. Specifically, let $\boldsymbol{y}$ denote the one-hot encoded label and $\hat{\boldsymbol{y}}$ the current model prediction. At training epoch i, the mixed label is defined as

$$\boldsymbol{y}^* = \alpha_i \boldsymbol{y} + (\mathbf{1} - \alpha_i)\hat{\boldsymbol{y}}, \tag{20}$$

here, the weight α_i is dynamically updated during training:

$$\alpha_i = \begin{cases} \exp\left(-\dfrac{i}{T_e} \dfrac{\widehat{\mathrm{LID}}_i}{\min_{\tau=0}^{i-1} \widehat{\mathrm{LID}}_\tau} \right), & \text{for samples from } \mathcal{D}_{\mathrm{Syn}} \\ 1, & \text{for samples from } \mathcal{D}_{\mathrm{Train}} \end{cases}, \tag{21}$$

where T_e is the total number of training epochs and $\widehat{\mathrm{LID}}_i$ is the local intrinsic dimension [49] at epoch i, which characterises the local complexity of the intraclass data distribution. In the early training phase, samples from the same class become more compact in the latent feature space, $\widehat{\mathrm{LID}}_i$ remains close to its minimum, and α_i stays near 1. Consequently, $\boldsymbol{y}^*$ depends primarily on the assigned labels. As training progresses and the model begins to fit noisy samples, $\widehat{\mathrm{LID}}_i$ increases, α_i decreases, and the weight of potentially noisy labels is correspondingly reduced. The final optimization objective for the diagnostic model is the cross-entropy loss computed with the mixed labels:

$$\mathcal{L}_{\mathrm{DNL}} = -\sum_{j=0}^{C-1} y_j^* \log P_j(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{P}}), \tag{22}$$

where C is the number of classes and $P_j(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{P}})$ denotes the predicted probability of class j given the tongue image $\boldsymbol{x}_{\mathrm{I}}$ and physiological indicators $\boldsymbol{x}_{\mathrm{P}}$.

This training strategy enables a dynamic balance between reliance on the assigned labels and the model predictions throughout training. In the early stages, it fully exploits the available annotations to guide representation learning. As training progresses, it gradually suppresses the influence of samples that are likely to carry incorrect labels and reduces their impact on the optimization. As a result, the adverse impact of mislabelled synthetic samples is largely mitigated, leading to robust learning under noisy labels and improved diagnostic performance.

4 | Experiments

4.1 | Diagnostic Performance

4.1.1 | *Comparison Methods*

To evaluate the effectiveness of the proposed DualSS framework in promoting balanced learning for FLD severity assessment, we compare it against representative state-of-the-art class-imbalance learning methods from three categories: algorithm-level rebalancing, sampling-based rebalancing, and generative augmentation. For algorithm-level rebalancing, we include Focal Loss [22] and Class-Balanced Loss (CBL) [23], Label-Distribution-Aware Margin Loss (LDAM) [50], and Learning to Reweight Examples (L2RW) [51]. We further include two sampling-based rebalancing strategies: Class-Balanced Sampling (CBS) and Progressively Balanced Sampling (PBS) [52]. In addition, we compare against four generative approaches: Synthetic Minority Oversampling Technique (SMOTE) [24], DeepSMOTE [27], CellGAN [29], and Latent Diffusion Models (LDM) [44]. For all generative baselines, each synthetic sample is required to include both a tongue image and the corresponding physiological indicators. For interpolation-based methods such as SMOTE and DeepSMOTE, we first select a base sample and then generate a synthetic tongue image and physiological indicators separately from this sample. For prior-based models such as CellGAN and LDM, we first synthesise tongue images and then, for each generated image, identify the most similar real tongue image from the same class. The matched real sample is treated as the base sample, and its physiological indicators and neighbours are used with SMOTE to generate the corresponding synthetic physiological indicators. All methods are implemented using their default parameter settings.

4.1.2 | *Diagnostic Backbone Models*

We combine each class-balancing method with multiple diagnostic backbone models and evaluate their performance. This strategy reduces potential coupling between any single model and a specific balancing technique, thereby strengthening the reliability of the experimental conclusions. Specifically, following prior work on intelligent tongue diagnosis, we adopt four validated tongue image feature extractors as backbones: ResNet [53], MobileNet [54], ConvNeXt [55], and TIPNet [12]. For each backbone, the output representation is projected by a fully connected (FC) layer into a 512-dimensional feature vector to form a unified embedding space. The physiological indicators are embedded using a two-layer multilayer perceptron (MLP) composed of FC layers with ReLU activations. The hidden dimension is set to 128, and the output dimension is set to 512. The image and indicator features are concatenated to form a fused representation, which is then fed into a two-layer MLP. The MLP uses a hidden dimension of 128 and an output dimension of 3, yielding the final FLD severity predictions.

4.1.3 | *Evaluation Metrics*

The following metrics are used to evaluate the performance of class-balanced methods: average class-specific accuracy (ACSA), macro-averaged F1 score (F1), and G-mean for multiclass (GMM). For a C-class classification problem, we construct a one-vs-rest confusion matrix for each class $j \in \{0, \ldots, C-1\}$, with the numbers of true positive (TP_j), false positive (FP_j), and false negative (FN_j). Based on these quantities, the ACSA is defined as

$$\mathrm{ACSA} = \frac{1}{C} \sum_{j=0}^{C-1} \frac{TP_j}{TP_j + FN_j}. \tag{23}$$

ACSA averages per-class accuracies and thereby reduces the dominance of the majority class. The precision and recall for class j are given by

$$\mathrm{Prec}_j = \frac{TP_j}{TP_j + FP_j}, \mathrm{Rec}_j = \frac{TP_j}{TP_j + FN_j}. \tag{24}$$

The F1 score for class j is

$$F1_j = \frac{2Prec_j Rec_j}{Prec_j + Rec_j}. \quad (25)$$

The macro-averaged F1 is then defined as

$$\text{Macro F1} = \frac{1}{C}\sum_{j=0}^{C-1} F1_j. \quad (26)$$

Macro F1 averages per-class F1 scores and assigns equal weight to every class, including minority classes. GMM is computed as

$$\text{GMM} = \left(\prod_{j=0}^{C-1} Rec_j\right)^{1/C}, \quad (27)$$

which takes the geometric mean of per-class recalls and strongly penalises methods that perform poorly on any single class. Together, these metrics provide a comprehensive evaluation under class imbalance, capturing overall accuracy while maintaining sensitivity to minority classes [20]. All experiments are conducted using five-fold cross-validation, and the final value for each metric is reported as the mean across folds.

4.1.4 | *Implementation Details*

All experiments were conducted in PyTorch on an NVIDIA GeForce RTX 4090 GPU. All tongue images were resized to 256 × 256 pixels.

In the RD-VAE, the encoder performs three downsampling stages, producing feature maps at spatial resolutions of [256, 128, 64, 32]. The decoder mirrors this hierarchy with symmetric upsampling. The latent feature comprises 16 channels, with 12 assigned to texture decoding and 4 to the shape mask. The codebook size was set to 8192. For RD-VAE training, we used the Adam optimiser with a learning rate of 5×10^{-6} and a cosine annealing schedule for 200 epochs.

For CA-DM training, we used the AdamW optimiser with a learning rate of 1×10^{-6} for 200 epochs, and the maximum number of diffusion steps was 1000.

In the synthetic sampling stage, each base sample used $K = 5$ nearest neighbours. To avoid duplicating near-neighbour pairs, we maintained a class-specific exclusion list and removed anchor-neighbour pairs that had already been used in previous synthetic sampling iterations. The diffusion process was run for 125 steps. The sampling rate was set to $\gamma = 0.1$. This default value was selected through validation-based optimization within the training data. Specifically, candidate values were compared on a validation split derived only from the training set in each fold, and the final value was chosen by prioritising balanced validation performance, as measured by macro F1, ACSA, and GMM.

During noise-robust training on the augmented dataset, the diagnostic model configurations were as summarised in Table 1. To ensure a fair comparison, all competing methods were evaluated using the same diagnostic model configurations.

TABLE 1 | Hyperparameters for diagnostic models.

Settings	ResNet	MobileNet	ConvNeXt	TIPNet
Version	−34	V2 small	Tiny	−9
Image size	256 × 256	256 × 256	256 × 256	256 × 256
Learning rate	0.1	0.01	0.02	0.02
Weight decay	0.01	0.001	0.002	0.002
Epochs	50	50	50	50
Pretrained	True	True	True	False

4.1.5 | *Overall Diagnostic Performance*

Table 2 summarises the performance of all competing methods paired with the four diagnostic models. Across all four backbones, the proposed DualSS consistently achieves the highest macro-F1. Under the ResNet backbone, DualSS achieves the best ACSA and GMM among all competing methods. It also attains the highest G-mean in combination with ResNet. For all other metric-backbone combinations, when DualSS is not the top performer, it always ranks within the top three. Taken together, these findings indicate that DualSS delivers the strongest overall performance in all diagnostic models, demonstrating robust behaviour and good generalisation. Among cost-sensitive learning methods, CBL shows a clear advantage for the GMM metric. This suggests that directly increasing the cost of misclassifying minority classes is effective in improving overall recall. However, it also elevates the false-positive rate, which explains why CBL is comparatively weaker in F1 performance, particularly when paired with ResNet. Notably, the generative baselines exhibit uniformly weak performance. This likely reflects the adverse impact of low-quality synthetic samples on diagnostic training. In the next section, we examine the quality of these synthesised images and their feature distribution in greater detail.

4.2 | Quality of Synthetic Tongue Images

Medical priors in tongue diagnosis indicate that tongue-surface textures carry essential pathological information. Accordingly, synthetic tongue images can enhance diagnostic performance only when they exhibit fine-grained texture and align with the target class-specific feature distribution. Images with blurred textures, aberrant shapes, or visible artefacts can reduce the model's sensitivity to disease-relevant signals and thereby degrade diagnostic performance. Therefore, we further evaluate the quality of synthetic tongue images produced by the generative baselines and compare them with those generated by DualSS.

4.2.1 | *Quantitative Assessment*

To quantitatively compare the quality of synthetic tongue images produced by different generative methods, we adopt the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). These metrics are computed by feeding real and synthetic images into a pretrained Inception network and comparing the resulting feature distributions. Together, they

TABLE 2 | Performance of competing rebalancing methods paired with diagnostic models on Tongue-FLD dataset.

	ResNet			MobileNet			ConvNeXt			TIPNet		
	F1	**ACSA**	**GMM**	**F1**	**ACSA**	**GMM**	**F1**	**ACSA**	**GMM**	**F1**	**ACSA**	**GMM**
Focal	48.32	48.56	26.65	56.90	55.72	44.08	56.11	55.41	41.22	55.74	54.44	41.10
CBL	54.52	57.59	52.10	58.81	60.65	**60.59**	59.95	61.61	**61.38**	61.32	62.27	**62.29**
LDAM	59.73	57.71	49.17	59.40	58.15	50.97	59.70	57.67	51.53	61.71	60.14	53.20
L2RW	57.19	54.91	45.51	59.95	58.65	51.92	56.01	53.97	46.62	60.35	58.33	50.73
CBS	56.02	56.22	50.76	60.22	**60.81**	57.94	58.98	**63.15**	60.46	60.64	62.49	58.52
PBS	58.28	58.27	53.97	60.87	59.91	58.25	61.95	62.53	59.66	61.11	62.21	59.46
SMOTE	54.98	53.77	46.40	56.68	57.06	52.03	60.67	61.17	57.57	60.34	59.11	53.50
DeepSMOTE	57.26	56.63	47.64	59.94	58.88	54.25	60.01	59.91	55.15	60.24	59.31	53.30
CellGAN	56.04	54.41	44.05	58.65	56.79	49.05	60.22	58.54	52.14	60.64	58.99	50.97
LDM	54.10	53.20	46.23	57.34	56.90	51.90	61.86	61.29	57.86	62.24	60.88	56.02
DualSS	**62.29**	**62.65**	**59.37**	**61.02**	60.33	59.61	**62.30**	61.91	60.10	**64.30**	**64.16**	61.23

Note: The best-performing metric in each column is shown in bold.

characterise both the perceptual realism of the synthetic tongue images and their distributional alignment with the real data manifold. For each minority class, we compute these metrics using intraclass synthetic and real samples. In particular, we report separate scores for the Mild FLD and M/S FLD classes. As shown in Table 3, the proposed DualSS achieves the best performance across FID and KID for both minority classes. This result indicates that the synthetic tongue images generated by DualSS exhibit feature distributions that are most similar to those of real samples, and are therefore better suited for balancing the dataset. Additional expert-based evaluation of synthetic-image realism and FLD-related tongue characteristics can be found in Supporting Information S1: Section SC.

TABLE 3 | Class-wise image-quality metrics for synthetic tongue images.

	Mild FLD		M/S FLD	
	FID ↓	**KID ↓**	**FID ↓**	**KID ↓**
SMOTE	56.62	0.0744	62.69	0.0729
DeepSMOTE	123.56	0.1640	131.73	0.1648
CellGAN	73.20	0.0849	70.45	0.0755
LDM	45.59	0.0563	52.92	0.0576
DualSS	**37.26**	**0.0429**	**43.00**	**0.0425**

Note: The best-performing metric in each column is shown in bold.

4.2.2 | *Qualitative Visual Assessment*

Representative examples of synthetic tongue images from each method are shown in Figure 5. Notable differences can be observed among the tongue images generated by different methods in terms of texture sharpness, morphological plausibility, and artefact suppression. SMOTE yields a blurred appearance and visible artefacts caused by pixel-level superposition. Although nearest neighbours share broadly similar shapes and textures, subtle misalignments in fine-scale patterns lead SMOTE's pixel-space interpolation to produce degraded images. DeepSMOTE introduces variability and robustness by varying the reconstruction order during training. In our experiments, however, this strategy often produces blurred images, with fine tongue-surface textures largely lost. We also observe occasional spurious artefacts near the tongue boundary. CellGAN employs adversarial training, which yields sufficiently sharp images and reduces blurring. However, adversarial training is prone to mode collapse, resulting in synthesised images with highly similar characteristics. Consistent with this, the tongue images produced by CellGAN exhibit lower diversity than those generated by the other methods. LDM initialises from random noise and then applies denoising diffusion, which yields high sample diversity and strong textural realism. Nevertheless, some synthesised tongue images exhibit anomalous shapes or sizes. This likely stems from stochastic initialisation, where certain starting states fall into unfavourable regions of the latent space and are not fully corrected during the subsequent reverse diffusion process. By contrast, images synthesised by the proposed DualSS exhibit high-fidelity, biologically plausible tongue-surface textures. They capture realistic tongue coating, spots, fissures, and other fine details. This improvement stems from the synergy of multiple mechanisms in the proposed DDG, including attention mechanisms and physics-informed guidance. In addition, the latent space synthetic sampling procedure increases the diversity of the synthesised images.

Figure 6 further illustrates how DualSS synthesises new tongue images from real samples within the same class. In each panel, the first two columns display real images (the base image and the selected neighbour), and the third column shows the corresponding synthesised image. The latent space synthetic sampling procedure produces images that inherit characteristics from both the base image and the selected neighbour. This helps ensure that the synthesised images are aligned with the feature distribution of the target class. In addition, the CA-DM's denoising stage promotes natural feature fusion and suppresses artefacts. Taken together, these observations indicate that the synthetic samples produced by DualSS exhibit high fidelity and preserve intraclass characteristics, making them well suited for training diagnostic models.

4.3 | Feature Distribution of Synthetic Samples

We further assess whether the synthetic samples generated by DualSS are distributionally aligned with real data in the learnt representation space. To this end, we extract embeddings from

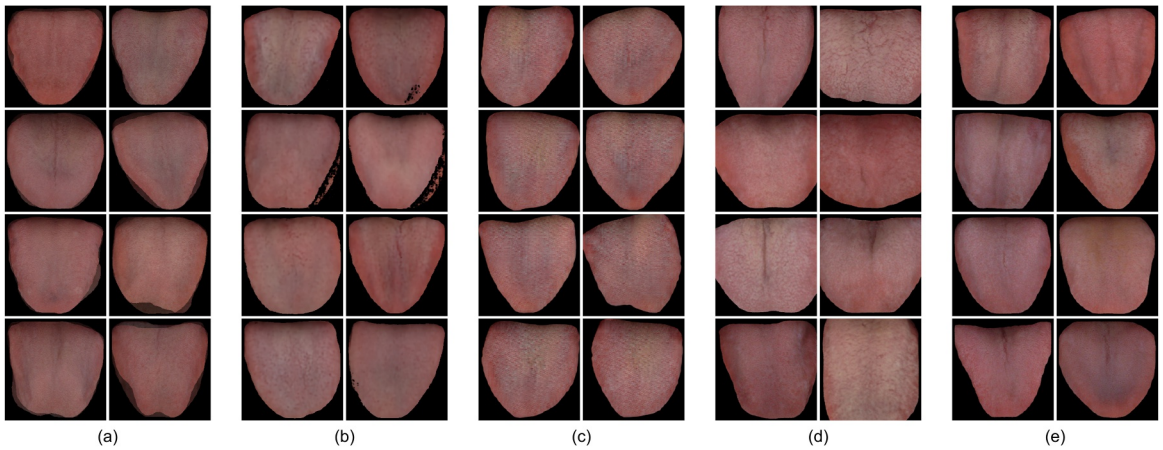


FIGURE 5 | Synthetic tongue images for the positive cases. (a) SMOTE. (b) DeepSMOTE. (c) CellGAN. (d) LDM. (e) DualSS.

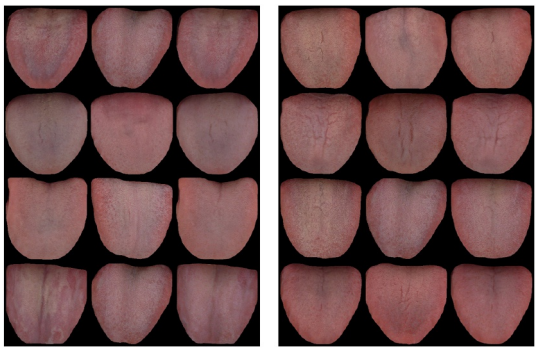


FIGURE 6 | Illustration of tongue images synthesised by DualSS. From left to right: base image, selected neighbour, and synthesised image.

the feature-fusion layer of four diagnostic models trained with DualSS and project them into two dimensions using Uniform Manifold Approximation and Projection (UMAP) [56]. Figure 7 depicts the distributions of real samples and minority-class samples synthesised by DualSS on a two-dimensional manifold. The UMAP projection was fit using only real samples to prevent synthesised data from influencing manifold construction, which could otherwise overstate their alignment with the real distribution. Across the four diagnostic models, the real samples form class-specific clusters, although substantial overlap persists. This pattern indicates that tongue-image features are indicative of FLD severity rather than in a one-to-one correspondence with it. The embeddings also display a pronounced directional trend, consistent with the ordinal nature of disease severity. The synthesised samples (black markers) are not arbitrarily dispersed in the embedding space. Instead, they concentrate within or immediately adjacent to the manifold of the corresponding minority-class real samples, effectively filling sparsely populated local regions. These observations indicate that DualSS effectively enriches the local neighborhoods of real minority-class samples. This effect is consistent across all four architecturally diverse backbones.

Because the distributions of real samples overlap, some instances are embedded outside their nominal class regions. When such out-of-region samples are used as bases for synthesis, the generated samples may fall within dense clusters of other classes in the embedding space. Such placements are legitimate and reflect genuine interclass overlap. They do not indicate spurious outliers.

These findings further corroborate that DualSS produces synthetic samples with high feature-level fidelity and strong distributional alignment. They are therefore well suited for constructing class-balanced training sets for diagnostic models.

4.4 | Ablation Study

We conduct ablation studies on the key mechanisms of the proposed dual-diffusion generator to evaluate their effectiveness in improving image synthesis quality. Because the model learns the tongue-image distribution through autoencoding, each reconstruction can be compared with its corresponding real image. This comparison provides a direct measure of the generator's fidelity to tongue morphology and disease-associated textures. Accordingly, we quantify generation quality using performance metrics for the tongue image reconstruction task, with results reported in Table 4. The mean absolute error (MAE) computes the average absolute per-pixel difference between the reconstruction and the ground-truth image. The root mean squared error (RMSE) computes the square root of the mean of squared per-pixel differences and is more sensitive than MAE to large errors. The perceptual loss (PerLoss) is computed in the feature space of a pretrained VGG-16 network. It quantifies the discrepancy between a reconstructed image and its ground-truth counterpart and captures semantic similarity at a perceptual level. The peak signal-to-noise ratio (PSNR) is defined as the ratio of the squared maximum pixel value to the mean squared error between the reconstructed and reference images, expressed in decibels, where higher PSNR values indicate lower distortion. The structural similarity index (SSIM) assesses the similarity between two images by jointly considering luminance, contrast, and structural information, aligning well with human visual perception. Multi-scale SSIM (MS-SSIM) extends SSIM by evaluating these components across multiple spatial scales, offering a more comprehensive measure of structural fidelity at different resolutions.

We construct a baseline model by removing all new mechanisms while keeping the overall architecture fixed. In particular, we do not change the latent-space dimensionality or the number of encoder and decoder layers. Building on this baseline, we then

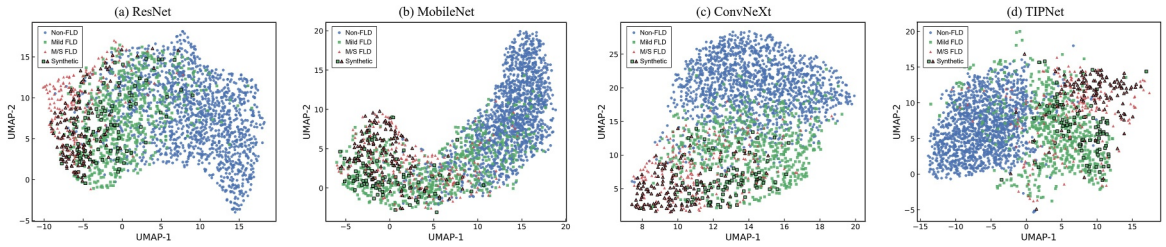


FIGURE 7 | Feature distribution of real and synthetic samples. Embeddings from the feature-fusion layer of four diagnostic models are projected to two dimensions using Uniform Manifold Approximation and Projection (UMAP).

TABLE 4 | Ablation study of the dual-diffusion generator.

Shape mask	Separated decoders	Local attention	GSRD	MAE ↓	RMSE ↓	PerLoss ↓	PSNR	SSIM	MS-SSIM
–	–	–	–	3.161	4.780	0.1287	32.33	0.8315	0.9363
√	–	–	–	3.079	4.798	0.1369	34.55	0.8281	0.9328
√	√	–	–	2.628	4.080	0.0971	36.00	0.8601	0.9511
√	√	√	–	2.556	3.972	0.0907	36.22	0.8706	0.9563
√	√	√	√	**2.309**	**3.596**	**0.0797**	**37.09**	**0.8849**	**0.9616**

Note: The best-performing metric in each column is shown in bold.

add components one at a time to form a series of comparable configurations: (i) a shape mask while retaining a single decoder, (ii) a dual-decoder design that explicitly decouples texture and shape, (iii) a local self-attention mechanism, and (iv) a physics-informed constraint based on the Grey–Scott reaction–diffusion model (GSRD). Table 4 shows that, as the proposed modules are incrementally incorporated, all evaluation metrics of DDG exhibit a steadily improving trend. The full configuration, which integrates all proposed mechanisms, achieves the best performance across all metrics and clearly outperforms the baseline model that excludes these mechanisms. It is noteworthy that, under the configuration that introduces only the shape mask while still employing a single decoder, the RMSE increases slightly, yet both SSIM and MS-SSIM exhibit a moderate decline. This finding suggests that a single decoder experiences representational entanglement and task competition when it is required to learn both texture maps and shape masks. As a result, it struggles to model both types of information adequately. After introducing the dual-decoder architecture that decouples texture and shape, this adverse effect is alleviated and the reconstruction quality improves accordingly. Subsequent incorporation of the local attention mechanism and the GSRD module yields additional gains across all metrics. Taken together, these components exert complementary effects that enhance both tongue-surface texture fidelity and structural plausibility. Additional expert-based realism scoring for the ablation variants can be found in Supporting Information S1: Section SD.

4.5 | Sensitivity to the Sampling Rate

The sampling rate coefficient γ is a key hyperparameter that controls the degree of synthetic oversampling applied to the minority class in DualSS. When γ is too low, only a limited number of minority-class samples are synthesised, and the class imbalance is only mildly alleviated. When γ is too high, the synthesised samples tend to be highly similar to one another and can occupy a large proportion of the training set. This situation reduces the effective diversity of the training data, increases the risk that the model overfits to non–disease-related cues, and drives the overall training distribution away from the true data distribution. To quantitatively characterise this trade-off, we vary the sampling rate coefficient γ in DualSS and evaluate the sensitivity of the final diagnostic performance to these changes. Figure 8 presents three metrics for the four diagnostic models under different sampling rates. To more finely resolve the behaviour in the low sampling rate regime near zero, we use a logarithmic scale and plot the metrics at $\gamma = [0.01, 0.02, 0.04, 0.07, 0.1, 0.2, 0.4, 0.7, 1.0]$. Starting from $\gamma = 0.01$, increasing the sampling rate enlarges the number of synthesised minority-class samples and alleviates the class imbalance in the dataset. As a result, all evaluation metrics exhibit a rising trend. Because the class-imbalance ratio between the Non-FLD and M/S FLD classes is 7.17:1, the number of synthesised M/S FLD samples becomes comparable to the number of real samples when γ is set to approximately 0.1. At this setting, the positive class is substantially expanded, yet synthetic samples do not overwhelm the real ones. Consequently, the four diagnostic models reach or approach their respective optimal performance levels. When γ is further increased, synthetic samples gradually dominate the minority classes, and the training distribution deviates more substantially from the true target distribution. Consequently, the performance of all four diagnostic models declines to varying degrees relative to their respective best levels. These results indicate that DualSS achieves a favourable trade-off between minority-class enhancement and distributional fidelity when the sampling rate lies near $\gamma \approx 0.1$. In this regime, DualSS exhibits strong robustness and stability, and this behaviour is consistently observed when it is combined with different diagnostic models.

5 | Discussion and Conclusion

In this work, we address the challenge of mitigating the adverse impact of positive-sample scarcity in public health applications. Such scarcity induces severe class imbalance and degrades the performance of diagnostic models. We treat FLD severity prediction based on tongue images and physiological indicators as a

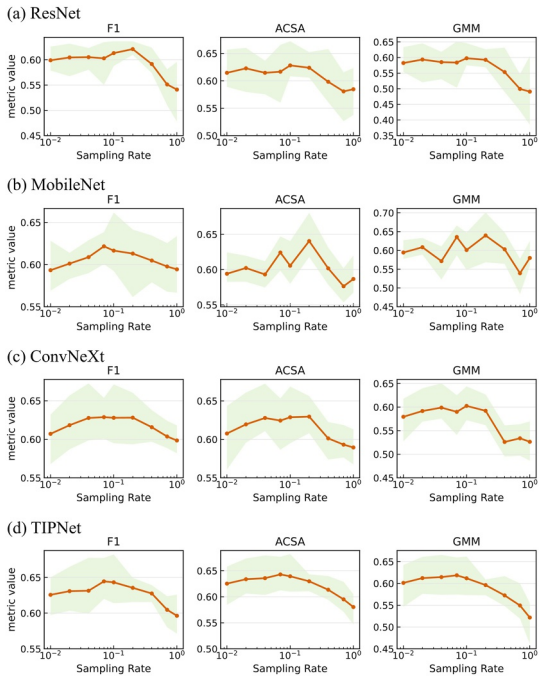


FIGURE 8 | Sensitivity of DualSS performance to sampling rate across diagnostic models with distinct backbone architectures: (a) ResNet, (b) MobileNet, (c) ConvNeXt, and (d) TIPNet. The solid line represents the mean performance from five-fold cross-validation. The shaded region indicates the range between minimum and maximum values across all folds.

representative imbalanced learning task. We propose DualSS, a dual-diffusion synthetic sampling framework that leverages abundant negative and unlabelled data to learn disease-agnostic tongue patterns and then performs intraclass synthesis to generate positive-class samples. Building on a dual-diffusion generator (DDG) with physics-informed reaction–diffusion constraints in pixel space and context-aware diffusion in latent space, DualSS produces high-fidelity biologically plausible tongue-surface textures. On top of DDG, we design a latent space synthetic sampling pipeline and combine it with a noise-robust training strategy for downstream diagnostic models.

Extensive experiments on FLD severity classification demonstrated the effectiveness of this design. Across four tongue diagnosis backbone architectures, DualSS consistently achieved the strongest overall performance. In contrast, existing generative baselines led to degraded performance, likely due to low-quality synthetic images and distributional mismatch. The quantitative quality assessment experiments demonstrated that DualSS outperformed existing generative models in maintaining consistency with the real feature distribution. Detailed visual inspection showed that images synthesised by DualSS preserved fine-grained tongue coating, spots, fissures, and other disease-associated details with high texture fidelity and structural plausibility. UMAP visualizations of the learnt feature space further revealed that DualSS effectively enriched the local neighborhoods of real minority-class samples. Ablation studies confirmed that each component of DDG contributed incrementally to image quality and that the full configuration yielded the best reconstruction metrics. Sensitivity analysis of the sampling rate showed that DualSS achieved a favourable balance between minority-class enhancement and distributional fidelity within an appropriate intermediate sampling regime. This operating range was robust across different diagnostic backbones. Overall, DualSS provides a robust framework that facilitates the deployment of tongue image-based AI systems, thereby enabling population-level screening of FLD severity and promoting public health.

Although DualSS demonstrates substantial benefits in alleviating the scarcity of positive samples, this study still has several limitations that point to promising directions for future work.

First, in this work, intraclass synthesis is used to generate new samples whose statistics align with the target disease classes, and the resulting tongue images exhibit visually realistic coating, fissures, and other disease-related details. However, the current framework does not yet support explicit control over semantic attributes such as coating thickness or fissure morphology when a disease label is specified. Developing attribute-aware tongue image generation mechanisms that can impose such semantic constraints on synthetic images represents an important direction for future research. To further elevate the generation quality of such key clinical features, future efforts could explore adversarial training frameworks or adopt cascaded generation strategies. Furthermore, integrating stronger pathological priors directly into the model architecture may substantially improve the synthesis of highly nuanced diagnostic details.

Second, the current image quality assessment relies primarily on distribution-level statistical metrics that quantify the agreement between the feature distributions of synthetic and real sample sets. For an individual synthetic tongue image, however, constructing clinically interpretable no-reference quality measures remains an open problem because no corresponding real reference image is available.

Third, this study was conducted on a single-centre dataset for assessing FLD through tongue diagnosis. The current data do not yet permit a systematic evaluation of the generalisability of DualSS across multiple centres, populations, and disease entities. Future studies should validate the framework in external cohorts and broader TCM diagnostic tasks. DualSS is designed to transfer knowledge from abundant negative samples to support the high-quality synthesis of positive minority-class samples. This design may be particularly useful in scenarios where positive samples are scarce and where disease-related information is encoded in fine-grained biological patterns. Therefore, DualSS may be applicable to other tongue-image-based disease screening tasks, such as preliminary screening for diabetes, gastrointestinal disorders, and other conditions. TCM facial diagnosis may also represent a potential extension because facial images contain subtle chromatic, textural, and morphological cues that may reflect systemic health status. Beyond TCM diagnostic applications, DualSS may also be extended to medical image analysis tasks in domains characterised by rich biological textures, such as histopathology and dermatology.

Acknowledgements

We gratefully acknowledge Dr. Xiaolan Wang and Dr. Zhibin Wang from Fuzhou University Affiliated Provincial Hospital, and Dr. Wen Tang from Fujian University of Traditional Chinese Medicine, for their valuable guidance on traditional Chinese medicine and their assistance with tongue image scoring and annotation.

Funding

This work was supported by the National Key R&D Program of China (Grants 2023YFC2604400, and 2024YFC3405800), the National Natural Science Foundation of China (Grant 82304086), Fujian Science and Technology Plan Industry-University-Research Cooperation Project of China (Grant 2021H6015), the Public Technology Service Platform Project of Xiamen City of China (Grant 3502Z20231043), the Government Funding of Fuqing City (Grant 2019B003), High-level Talents Research Start-up Project of Fujian Medical University (Grants XRCZX2021025, XRCZX2023030, and XRCZX2025013), the Innovation of Science and Technology, Fujian Province (Grant 2024Y9079), the Public Technology Service Platform Project of Xiamen City (Grant 3502Z20231043), the Open Research Fund of NHC Key Laboratory of Etiological Epidemiology of Chronic Diseases with High Incidence in Fujian-Taiwan Area (Co-construction) (Grant NHCEECD-202601), and Fujian Provincial Key Laboratory of Environment Factors and Cancer (Grant GWSZD-202503).

Ethics Statement

The study was approved by the Ethics Review Committee of Fujian Medical University (approval numbers [2017-07] and [2020-58]).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Code is publicly available at https://github.com/MLDMXM2017/DualSS. Dataset Tongue-FLD is publicly available at https://github.com/MLDMXM2017/HM-TDF.

References

1. M. H. Le, D. M. Le, T. C. Baez, et al., "Global Incidence of Non-Alcoholic Fatty Liver Disease: A Systematic Review and Meta-Analysis of 63 Studies and 1,201,807 Persons," *Journal of Hepatology* 79, no. 2 (2023): 287–295, https://doi.org/10.1016/j.jhep.2023.03.040.

2. A. Borrelli, P. Bonelli, F. M. Tuccillo, et al., "Role of Gut Microbiota and Oxidative Stress in the Progression of Non-Alcoholic Fatty Liver Disease to Hepatocarcinoma: Current and Innovative Therapeutic Approaches," *Redox Biology* 15 (2018): 467–479, https://doi.org/10.1016/j.redox.2018.01.009.

3. Y.-Y. Chen and M. M. Yeh, "Non-Alcoholic Fatty Liver Disease: A Review With Clinical and Pathological Correlation," *Journal of the Formosan Medical Association* 120, no. 1 Part 1 (2021): 68–77, https://doi.org/10.1016/j.jfma.2020.07.006.

4. S. K. Han, S. K. Baik, and M. Y. Kim, "Non-Alcoholic Fatty Liver Disease: Definition and Subtypes," supplement, *Clinical and Molecular Hepatology* 29, no. Suppl (2023): S5–S16, https://doi.org/10.3350/cmh.2022.0424.

5. H. Park, E. L. Yoon, M. Kim, et al., "Cost-Effectiveness of Advanced Hepatic Fibrosis Screening in Individuals With Suspected MASLD Identified by Serologic Noninvasive Tests," *Scientific Reports* 15, no. 1 (2025): 24186, https://doi.org/10.1038/s41598-025-08434-z.

6. Q. Liu, Y. Li, P. Yang, et al., "A Survey of Artificial Intelligence in Tongue Image for Disease Diagnosis and Syndrome Differentiation," *Digital Health* 9 (2023): 20552076231191044, https://doi.org/10.1177/20552076231191044.

7. S. Lee, B. Arefaine, N. Begum, et al., "Oral-Gut Microbiome Interactions in Advanced Cirrhosis: Characterisation of Pathogenic Enterotypes and Salivatypes, Virulence Factors and Antimicrobial Resistance," *Journal of Hepatology* 82, no. 4 (2025): 622–633, https://doi.org/10.1016/j.jhep.2024.09.046.

8. C. Lu, H. Zhu, D. Zhao, et al., "Oral-Gut Microbiome Analysis in Patients With Metabolic-Associated Fatty Liver Disease Having Different Tongue Image Feature," *Frontiers in Cellular and Infection Microbiology* 12 (2022): 787143, https://doi.org/10.3389/fcimb.2022.787143.

9. M.-C. Hu, K.-C. Lan, W.-C. Fang, et al., "Automated Tongue Diagnosis on the Smartphone and Its Applications," *Computer Methods and Programs in Biomedicine* 174 (2019): 51–64, https://doi.org/10.1016/j.cmpb.2017.12.029.

10. T. Jiang, X.-J. Guo, L.-P. Tu, et al., "Application of Computer Tongue Image Analysis Technology in the Diagnosis of NAFLD," *Computers in Biology and Medicine* 135 (2021): 104622, https://doi.org/10.1016/j.compbiomed.2021.104622.

11. R.-R. Wang, J.-L. Chen, S.-J. Duan, et al., "Noninvasive Diagnostic Technique for Nonalcoholic Fatty Liver Disease Based on Features of Tongue Images," *Chinese Journal of Integrative Medicine* 30, no. 3 (2024): 203–212, https://doi.org/10.1007/s11655-023-3616-1.

12. J. Gao, T. Chen, Y. Xu, et al., "Accurate Fatty Liver Disease Diagnosis With a Multi-Source Feature Fusion Model on the Segmented Tongue Image Dataset," *Journal of Advanced Research* 85 (2025): 507–520, (In Press), https://doi.org/10.1016/j.jare.2025.10.003.

13. T. Chen, J. Gao, Y. Xu, et al., "Hard Sample Mining-Based Tongue Diagnosis for Fatty Liver Disease Severity Classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025; Sep 23; Daejeon* (Springer Nature Switzerland, 2025).

14. S. Man, Y. Deng, Y. Ma, et al., "Prevalence of Liver Steatosis and Fibrosis in the General Population and Various High-Risk Populations: A Nationwide Study With 5.7 Million Adults in China," *Gastroenterology* 165, no. 4 (2023): 1025–1040, https://doi.org/10.1053/j.gastro.2023.05.053.

15. S. Dai, X. Guo, S. Liu, et al., "Application of Intelligent Tongue Image Analysis in Conjunction With Microbiomes in the Diagnosis of MAFLD," *Heliyon* 10, no. 7 (2024): e29269, https://doi.org/10.1016/j.heliyon.2024.e29269.

16. T. Chen, Y. Chen, Z. Zhou, Y. Zhu, L. He, and J. Zhang, "Deep Learning-Based Automated Tongue Analysis System for Assisted Chinese Medicine Diagnosis," *Frontiers in Physiology* 16 (2025): 1559389, https://doi.org/10.3389/fphys.2025.1559389.

17. L. Yang, Q. Dong, D. Lin, and X. Lü, "TongueNet: A Multi-Modal Fusion and Multi-Label Classification Model for Traditional Chinese Medicine Tongue Diagnosis," *Frontiers in Physiology* 16 (2025): 1527751, https://doi.org/10.3389/fphys.2025.1527751.

18. X.-Z. Lu, S. Liu, X.-X. Lin, et al., "An AI-Powered Tongue Image Model for Home-Based Monitoring of Liver Fibrosis," *npj Digital Medicine* 9, no. 1 (2025): 67, https://doi.org/10.1038/s41746-025-02246-1.

19. M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling Imbalanced Medical Datasets: Review of a Decade of Research," *Artificial Intelligence Review* 57, no. 10 (2024): 273, https://doi.org/10.1007/s10462-024-10884-2.

20. W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions," *Artificial Intelligence Review* 57, no. 6 (2024): 137, https://doi.org/10.1007/s10462-024-10759-6.

21. E. Pachetti and S. Colantonio, "A Systematic Review of Few-Shot Learning in Medical Imaging," *Artificial Intelligence in Medicine* 156 (2024): 102949, https://doi.org/10.1016/j.artmed.2024.102949.

22. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2017), 2999–3007.

23. Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019), 9260–9269.

24. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research* 16, no. 1 (2002): 321–357, https://doi.org/10.1613/jair.953.

25. H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing* (Springer Berlin Heidelberg, 2005), 878–887.

26. M. Han, H. Guo, J. Li, and W. Wang, "Global-Local Information Based Oversampling for Multi-Class Imbalanced Data," *International Journal of Machine Learning and Cybernetics* 14, no. 6 (2023): 2071–2086, https://doi.org/10.1007/s13042-022-01746-w.

27. D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Transactions on Neural Networks and Learning Systems* 34, no. 9 (2023): 6390–6404, https://doi.org/10.1109/TNNLS.2021.3136503.

28. S. Li, Y. Lin, H. Chen, and K.-T. Cheng, "Iterative Online Image Synthesis via Diffusion Model for Imbalanced Classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2024* (Springer Nature Switzerland, 2024), 371–381.

29. Z. Shen, M. Cao, S. Wang, L. Zhang, and Q. Wang, "CellGAN: Conditional Cervical Cell Synthesis for Augmenting Cytopathological Image Classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2023* (Springer-Verlag, 2023), 487–496.

30. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2020), 574.

31. S. Rajaraman, Z. Liang, Z. Xue, and S. Antani, "Addressing Class Imbalance With Latent Diffusion-Based Data Augmentation for Improving Disease Classification in Pediatric Chest X-Rays," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE, 2024), 5059–5066.

32. Y. Shi, A. Abulizi, H. Wang, et al., "Diffusion Models for Medical Image Computing: A Survey," *Tsinghua Science and Technology* 30, no. 1 (2025): 357–383, https://doi.org/10.26599/TST.2024.9010047.

33. Y. Qin, H. Zheng, J. Yao, M. Zhou, and Y. Zhang, "Class-Balancing Diffusion Models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2023), 18434–18443.

34. M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations," *Journal of Computational Physics* 378 (2019): 686–707, https://doi.org/10.1016/j.jcp.2018.10.045.

35. S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, "Scientific Machine Learning Through Physics–Informed Neural Networks: Where We Are and What's Next," *Journal of Scientific Computing* 92, no. 3 (2022): 88, https://doi.org/10.1007/s10915-022-01939-z.

36. Z. Min, Z. M. C. Baum, S. U. Saeed, et al., "Non-Rigid Medical Image Registration Using Physics-Informed Neural Networks," in *Information Processing in Medical Imaging; San Carlos de Bariloche, Argentina* (Springer Nature Switzerland, 2023), 601–613.

37. F. Shone, N. Ravikumar, T. Lassila, et al., "Deep Physics-Informed Super-Resolution of Cardiac 4D-Flow MRI," in *Information Processing in Medical Imaging* (Springer Nature Switzerland, 2023), 511–522.

38. M. F. Fathi, I. Perez-Raya, A. Baghaie, et al., "Super-Resolution and Denoising of 4D-Flow MRI Using Physics-Informed Deep Neural Nets," *Computer Methods and Programs in Biomedicine* 197 (2020): 105729, https://doi.org/10.1016/j.cmpb.2020.105729.

39. M. Sarabian, H. Babaee, and K. Laksari, "Physics-Informed Neural Networks for Brain Hemodynamic Predictions Using Medical Imaging," *IEEE Transactions on Medical Imaging* 41, no. 9 (2022): 2285–2303, https://doi.org/10.1109/TMI.2022.3161653.

40. P. H. M. Voorter, W. H. Backes, O. J. Gurney-Champion, et al., "Improving Microstructural Integrity, Interstitial Fluid, and Blood Microcirculation Images From Multi-b-Value Diffusion MRI Using Physics-Informed Neural Networks in Cerebrovascular Disease," *Magnetic Resonance in Medicine* 90, no. 4 (2023): 1657–1671, https://doi.org/10.1002/mrm.29753.

41. S. Zakany and M. C. Milinkovitch, "Simple Reaction-Diffusion Modeling Predicts Inconspicuous Neighborhood-Dependent Color Subclustering of Lizard Scales," *Physical Review X* 13, no. 4 (2023): 041011, https://doi.org/10.1103/PhysRevX.13.041011.

42. A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017).

43. Z. Wang and Z. Liu, "Posterior Collapse of a Linear Latent Variable Model," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2022), 37537–37548.

44. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022), 10684–10695.

45. L. Yuan, L. Yang, S. Zhang, et al., "Development of a Tongue Image-Based Machine Learning Tool for the Diagnosis of Gastric Cancer: A Prospective Multicentre Clinical Cohort Study," *EClinicalMedicine* 57 (2023): 101834, https://doi.org/10.1016/j.eclinm.2023.101834.

46. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is all You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2017), 6000–6010.

47. D. L. Gandy and M. R. Nelson, "Analyzing Pattern Formation in the Gray-Scott Model: An XPPAUT Tutorial," *SIAM Review* 64, no. 3 (2022): 728–747, https://doi.org/10.1137/21M1402868.

48. M. Tahmid, M. S. Alam, N. Rao, and K. M. A. Ashrafi, "Image-to-Image Translation With Conditional Adversarial Networks," in *2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, (2023), 1–5s.

49. X. Ma, Y. Wang, M. E. Houle, et al., "Dimensionality-Driven Learning With Noisy Labels," in *Proceedings of the 35th International Conference on Machine Learning; Proceedings of Machine Learning Research* (PMLR, 2018), 3355–3364.

50. K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning Imbalanced Datasets With Label-Distribution-Aware Margin Loss," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019).

51. M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to Reweight Examples for Robust Deep Learning," in *Proceedings of the 35th International Conference on Machine Learning; Proceedings of Machine Learning Research* (PMLR, 2018), 4334-–4343.

52. B. Kang, S. Xie, M. Rohrbach, et al., "Decoupling Representation and Classifier for Long-Tailed Recognition," *arXiv Preprint* (2019), https://doi.org/10.48550/arXiv.1910.09217.

53. X. Wang, X. Wang, Y. Lou, 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): 114905, https://doi.org/10.1016/j.jep.2021.114905.

54. D. Qiu, X. Zhang, X. Wan, et al., "A Novel Tongue Feature Extraction Method on Mobile Devices," *Biomedical Signal Processing and Control* 80 (2023): 104271, https://doi.org/10.1016/j.bspc.2022.104271.

55. X. Sha, Z. Guan, Z. Chen, J. Han, and Y. Wang, "SSC-Net: A Multi-Task Joint Learning Network for Tongue Image Segmentation and Multi-Label Classification," *Digital Health* 11 (2025): 20552076251343696, https://doi.org/10.1177/20552076251343696.

56. P. Castellano-Escuder, D. K. Zachman, K. Han, and M. D. Hirschey, "GAUDI: Interpretable Multi-Omics Integration With UMAP Embeddings and Density-Based Clustering," *Nature Communications* 16, no. 1 (2025): 5771, https://doi.org/10.1038/s41467-025-60822-1.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.
Supporting Information S1: cit270174-sup-0001-suppl-data.docx.