Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

Dual Generation of Medical Dermatology Image-Mask Pairs Based on Fine-Tuned Stable-Diffusion

Zhaobin Xu *

Shandong University, Qingdao, China

* Corresponding Author Email: sea.xuo@gmail.com

Abstract. Medical image analysis plays a pivotal role in the early diagnosis of diseases such as skin lesions. However, the scarcity of data and class imbalance significantly hinder the performance of deep learning models. This paper proposes a novel method that leverages the pre-trained Stable Diffusion-2.0 model to generate high-quality synthetic skin lesion images and corresponding segmentation masks. This approach augments training datasets for classification and segmentation tasks. We adapt Stable Diffusion-2.0 through domain-specific Low-Rank Adaptation (LoRA) fine-tuning and joint optimization of multi-objective loss functions, enabling the model to simultaneously generate clinically relevant images and segmentation masks conditioned on textual descriptions in a single step. Experimental results show that the generated images, validated by FID scores, closely resemble real images in quality. A hybrid dataset combining real and synthetic data markedly enhances the performance of classification and segmentation models, achieving substantial improvements in accuracy and F1-score of 8% to 15%, with additional positive gains in other key metrics such as the Dice coefficient and IoU. Our approach offers a scalable solution to address the challenges of medical imaging data, contributing to improved accuracy and reliability in diagnosing rare diseases.

Keywords: Stable Diffusion; Medical Image Synthesis; Data Augmentation; AI for Skin Lesions; Large Language Model.

1. Introduction

Medical image analysis serves as a cornerstone of modern medicine, particularly in the early detection of skin lesions. Despite its importance, medical image analysis faces two primary challenges: the scarcity of high-quality annotated data and class imbalance. Firstly, acquiring medical images demands specialized equipment and personnel, while the annotation process relies on experienced dermatologists, rendering data collection both costly and time-intensive. Secondly, skin lesion datasets frequently exhibit severe class imbalance, with benign lesions (e.g., nevi) vastly outnumbering malignant ones (e.g., melanoma). This disparity biases deep learning models toward majority classes during training, compromising their ability to detect rare, minority-class lesions accurately.

Conventional data augmentation techniques—such as rotation, flipping, and color adjustments—offer limited diversity and fail to adequately address class imbalance, despite marginally increasing data volume [1]. Consequently, synthetic data generation has emerged as a promising approach to overcome these limitations by producing diverse, high-quality images and annotations. Synthetic data generation has emerged as a promising solution by producing high-quality images and annotations. Koetzier et al. [2] reviewed the utility of generative models in medical imaging, while Ibrahim et al. [3] explored recent advancements in multimodal synthetic data generation. Furthermore, Ktena et al. [4] demonstrated that generative models can enhance classifier fairness, providing novel strategies to mitigate class imbalance. Among these approaches, Generative Adversarial Networks (GANs) have been employed to synthesize medical images; however, their training instability and mode collapse limit both image quality and diversity [5]. In recent years, diffusion models have gained prominence as an advanced image generation technique. Stable Diffusion, a pre-trained diffusion-based model,

effectively generates high-quality images from text prompts, providing robust flexibility and control [6].

This paper introduces a method that utilizes Stable Diffusion to generate synthetic skin lesion images and segmentation masks, thereby enriching training datasets for classification and segmentation tasks. This paper fine-tunes Stable Diffusion-2.0 using Low-Rank Adaptation (LoRA) [7], proposes a one-prompt dual-generation technique, and harnesses large language models to enhance dataset diversity. Experiments conducted on the ISIC-GPT and HAM10000 datasets validate the efficacy of our synthetic data, demonstrating substantial performance improvements. This paper's primary contributions are threefold:

1.1. Domain-Specific Fine-Tuning:

This paper adapts Stable Diffusion-2.0 to the medical imaging domain through domain-specific LoRA fine-tuning and joint optimization of multi-objective loss functions, enabling the generation of high-quality synthetic skin lesion images;

1.2. One-Prompt Dual-Generation:

This paper develops a technique to simultaneously generate image-segmentation mask pairs from a single prompt, ensuring efficiency and consistency in the output;

1.3. Performance Validation:

This paper verifies gains in classification and segmentation tasks using a hybrid dataset of real and synthetic data, evaluated on the ISIC-GPT and HAM10000 datasets.

2. Related Works

2.1. Medical Images Generation

Medical image generation techniques address the challenges of data scarcity and privacy constraints by producing realistic synthetic images.

Variational Autoencoders (VAEs) generate new samples by learning a probabilistic mapping from images to a latent space, optimized through minimizing reconstruction error and Kullback-Leibler (KL) divergence [8]. For example, Baur et al. applied VAEs to generate brain MRI images, improving tumor detection accuracy [9]. However, VAE-generated images often lack sharpness and high-frequency details. Generative Adversarial Networks (GANs) produce high-fidelity images via adversarial training between a generator and a discriminator [10]. Yi et al. surveyed GAN applications in medical imaging, including reconstruction, segmentation, and classification, emphasizing their ability to generate high-resolution CT and MRI images [11]. Despite these strengths, GANs frequently encounter mode collapse and training instability, reducing image diversity and clinical utility.

Diffusion models have recently emerged as a robust alternative to GANs, generating high-quality images through iterative denoising processes. Ho et al.'s Denoising Diffusion Probabilistic Model (DDPM) laid the theoretical groundwork with a forward diffusion process (adding Gaussian noise) and a reverse denoising process (image recovery) [12]. Kazerouni et al. provided an extensive review of diffusion models in medical imaging, highlighting their strengths in synthesis, reconstruction, and augmentation, as well as their ability to enhance downstream task performance [13]. In dermatology, Bozorgpour et al.'s DermoSegDiff model leveraged boundary-aware diffusion to achieve high-precision segmentation, surpassing GAN-based methods [14]. Stable Diffusion is increasingly adopted; Yu et al.'s MedDiff-FM exploited diffusion models' versatility across various medical imaging tasks [15].

In contrast to prior diffusion-based research, our study introduces a novel approach by generating both images and segmentation masks from a single prompt, streamlining the process. Additionally, we integrate large language models to enrich clinical descriptions, thereby improving data diversity and relevance for medical applications.

2.2. Data Augmentation

Data augmentation enhances the robustness and generalization of medical image analysis models. Traditional techniques, such as geometric transformations (e.g., rotation, flipping, scaling) and color adjustments (e.g., brightness, contrast), are simple to implement but provide limited diversity. Chlap et al. reviewed their use in medical imaging, noting their insufficiency for capturing complex lesion morphologies [16]. More advanced methods like mixup and CutMix create new samples by blending images, yet in medical contexts, they risk distorting critical clinical features, such as lesion boundaries.

Generative models offer a solution by producing diverse, clinically relevant samples. Abdelhalim et al.'s self-attention progressive GAN (PGAN) augmented skin lesion datasets, boosting classification performance [17]. Shin et al. illustrated GANs' potential in generating synthetic MRIs for tumor segmentation [18]. Diffusion models exhibit even greater potential; Montoya et al.'s MAM-E model generated mammograms for breast cancer classification, outperforming traditional augmentation techniques [19].

Our method capitalizes on Stable Diffusion's conditional generation capabilities to target minority classes, addressing class imbalance effectively. Through domain-specific fine-tuning and conditional prompts, we ensure that synthetic images closely align with real images in clinically significant features, such as texture and boundaries, enhancing their utility in medical image analysis.

3. Method

Our method aims to generate high-quality synthetic skin lesion images and their segmentation masks using Stable Diffusion to augment the training dataset for classification and segmentation tasks.And the detailed pipeline is shown in Fig. 1.

3.1. Stable Diffusion

Diffusion models represent a class of powerful generative models capable of producing high-quality images through an iterative denoising process. The Denoising Diffusion Probabilistic Model (DDPM), proposed by Ho et al. [12], established the theoretical foundation for diffusion models. The forward diffusion process incrementally introduces Gaussian noise to the data across multiple timesteps, mathematically expressed as:

$$q(x_t \mid x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{1}$$

where β_t denotes the noise schedule at timestep t, and x_t represents the noisy data. The reverse process learns to recover the original data by modeling the denoising distribution:

$$p_\theta(x_{t-1} \mid x_t) = N\left(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t) I\right) \tag{2}$$

Here, μ_θ and σ_θ are parameters predicted by a neural network, typically a UNet architecture.

Stable Diffusion, introduced by Rombach et al. [8], enhances computational efficiency by conducting the diffusion process within the latent space of a Variational Autoencoder (VAE). The workflow involves encoding an image x into a latent representation z, performing diffusion and denoising in

this latent space, and decoding the resulting latent back into the image domain. This generation process can be formulated as:

$$z_T \sim N(0,I), z_0 = f_\theta(z_T, c), x = Decoder(z_0) \tag{3}$$

where z_T is the initial noise, c is the conditioning input (e.g., a text prompt), and f_θ denotes the denoising network (UNet).

Owing to their ability to generate high-quality images, diffusion models have gained prominence in medical imaging. Khader et al. [20] and Wang et al. [21] developed 3D medical image generation methods, respectively, which combine VQ-GAN and ControlNet to improve the generation quality. In the field of dermatology, Sagers et al. [22] and Akrout et al. [23] leveraged diffusion models to produce synthetic data, significantly enhancing classifier performance, particularly across diverse populations and datasets.

In this study, we adopt Stable Diffusion-2.0 as our baseline model due to its optimal balance of performance and compatibility with consumer-grade hardware. We utilize text prompts to conditionally generate skin lesion images and masks, thereby augmenting the training data for downstream classification and segmentation tasks, as detailed in subsequent sections.

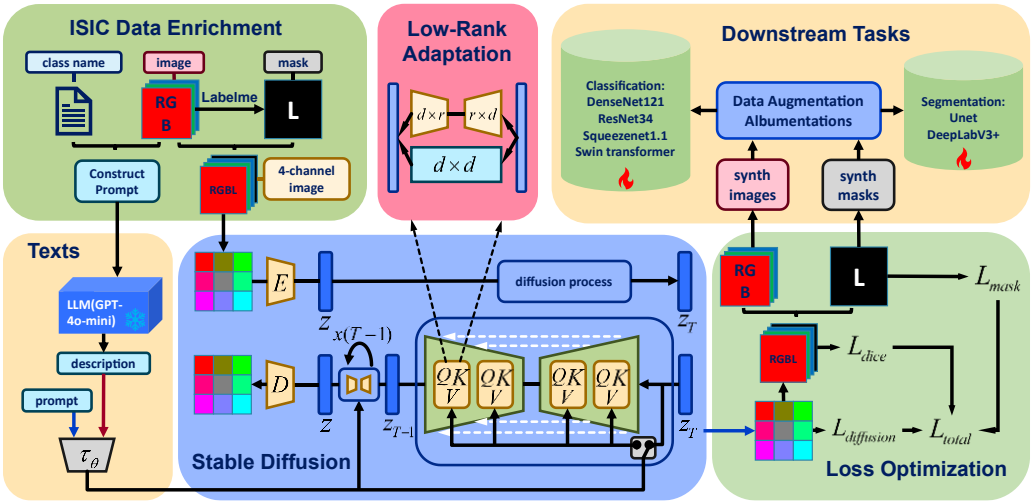

Figure 1. The Pipeline of our Proposed SkinDualGen Approach

3.2. Low-Rank Adaptation

To efficiently tailor Stable Diffusion to the medical imaging domain, we employ Low-Rank Adaptation (LoRA) [7]. LoRA approximates weight updates using low-rank matrices, substantially reducing computational overhead while maintaining model efficacy. For a pre-trained weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA expresses the weight update as:

$$W + \Delta W, \Delta W = AB \tag{4}$$

where $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times k}$, and $r \ll \min(d,k)$ is the rank, Only A and B are trainable, significantly decreasing the number of parameters requiring optimization.

We integrate LoRA adapters into the attention layers of Stable Diffusion's UNet, which are critical for incorporating text-based conditions to align generated images with clinical descriptions. This approach reduces the number of trainable parameters by over 90%. Furthermore, LoRA is extensible and can be combined with other techniques such as ControlNet for additional conditional control [24], or DreamBooth for personalized generation [25].

3.3. One-Prompt-Dual-Generation

To simultaneously generate images and masks from a single prompt, we construct four-channel images, with the first three channels representing the RGB image and the fourth channel encoding the segmentation mask. Specifically, the input convolutional layer of the encoder is adapted from 3 to 4 channels, keeping the same kernel size of 3 × 3 and step size of 1 as in the original model. Similarly, the output convolutional layer of the decoder is adapted to generate a 4-channel output, with the same kernel size and step size. Additionally, the UNet's first convolutional layer is updated to process four-channel latent inputs. In a nutshell, this unified four-channel representation enables the VAE to jointly encode and decode images and masks, streamlining the generation process while maintaining consistency.

The training process leverages multi-task learning, optimizing a combination of loss functions. The primary diffusion loss is defined as the mean squared error (MSE) between the predicted and actual noise in the latent space:

$$L_{diffusion} = E_{t,z_t,\varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta\left(z_t, t, c\right)\right\|_2^2\right] \tag{5}$$

We also incorporate a binary cross-entropy (BCE) loss to assess the discrepancy between the predicted mask logits and the ground truth:

$$L_{mask} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right] \tag{6}$$

Additionally, we employ the Dice loss to enhance the overlap between predicted and actual mask regions:

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \varepsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \varepsilon} \tag{7}$$

where ε is a smoothing term to ensure numerical stability.

The total loss is a weighted combination of these components:

$$L_{total} = \lambda_1 L_{diffusion} + \lambda_2 L_{mask} + \lambda_3 L_{dice} \tag{8}$$

During training, the four-channel images are encoded into the VAE's latent space, subjected to noise at random timesteps, and denoised by the UNet. In inference, a single prompt produces a four-channel output. The increase in the number of parameters and computational cost due to these architectural modifications is negligible and limited to the additional weights of the first convolutional layer of the encoder and the last convolutional layer of the decoder. This increase is negligible compared to the parameter size of the entire model. In addition, by generating the image and mask from a single cue, our approach utilizes the shared computation of the diffusion process, potentially reducing the overall computational overhead compared to generating the image and mask separately.

3.4. Datasets Enrichment based on Large Language Model

In the field of dermatology, publicly available datasets are limited, with the ISIC [26] series being among the most prominent, encompassing over 13,000 dermoscopic images. These datasets, curated

by the International Skin Imaging Collaboration (ISIC). Most images are accompanied by clinical metadata, which have been reviewed and annotated by recognized experts, including detailed dermoscopic features. However, the sheer volume of these datasets presents challenges for individual researchers conducting large-scale analyses.

For this study, we selected the ISIC 2020 dataset due to its diverse range of skin lesion categories. From this dataset, we curated a refined subset comprising 1,990 images across seven categories. As the original ISIC 2020 dataset lacks segmentation masks, we employed the Labelme tool to manually annotate each image, generating corresponding segmentation masks to support downstream segmentation tasks.

In recent years, Large Language Models (LLMs) have emerged as powerful tools in medical imaging, owing to their exceptional capabilities in text comprehension, analysis, and generation. Our approach involved retrieving each category along with its corresponding RGB skin lesion images and using a meticulously designed prompt to guide GPT-4o-mini in generating detailed descriptions. These descriptions aim to encapsulate fine-grained features using professional medical terminology, adhering to the prompt structure: "Analyze this {category} dermatology image. Describe in medical terms and give a sentence. Use ICD-11 terminology and begin with 'a dermoscopic lesion photo of {category} for skin cancer diagnosis,...'". Unlike Medghalchi et al.'s approach of merely employing diverse adjectives to expand monotonous fixed statements [27], our LLM-based semantic expansion ensures greater diversity and specificity. Following this process, we generated detailed descriptions for each image, reviewed by dermatologists. The resulting dataset is well-suited for small-scale multimodal research prioritizing richness.

3.5. Synthesize Images and Masks for Downstream Tasks

The high-quality synthetic images and masks generated in this study are utilized to augment training datasets for classification and segmentation tasks, with the goal of enhancing model performance and mitigating class imbalance issues.

In classification tasks, class imbalance poses a significant challenge, with certain lesion types (e.g., melanoma) being underrepresented. To address this, we could generate additional synthetic samples for minority classes using text prompts such as "a dermoscopic lesion photo of melanoma for skin cancer diagnosis" to produce melanoma image-mask pairs. These synthetic images were combined with real images, significantly improved model generalization by increasing data diversity. For segmentation tasks, synthetic masks provide additional training pairs, enabling the model to learn more lesion boundaries. The consistency between synthetic masks and their corresponding images ensures that the model can effectively learn from diverse samples. During training, both real and synthetic image-mask pairs were employed, expanding the dataset and reducing the risk of overfitting.

In addition to synthetic data, we incorporated traditional data augmentation techniques to further enrich the dataset. For classification tasks, we applied transformations from the Torchvision library, while for segmentation tasks, we utilized the Albumentations library, ensuring consistency between images and masks, markedly enhancing the performance of downstream tasks.

4. Experiments

4.1. Experimental Setting

4.1.1. Datasets

4.1.1.1 Skin Cancer ISIC and GPT-based Descriptions

As outlined in Section III-D, this novel dataset comprises 1,990 skin lesion images spanning seven categories and consists of three components: RGB images, L masks, and JSON descriptions. It is publicly accessible on Kaggle. The training set was employed to fine-tune Stable Diffusion-2.0 and constitutes the real data portion of the hybrid training set for downstream classification and

segmentation models. The test set was used to evaluate the performance of the trained downstream models.

4.1.1.2 HAM10000

The HAM10000 (Human Against Machine with 10,000 training images) dataset is a comprehensive collection of 10,015 color images designed for the study and training of skin lesion classification models. The dataset exhibits significant class imbalance, presenting challenges for model training. In this study, it serves as a test set to evaluate the generalizability of classification and segmentation models trained on the hybrid dataset to unseen data.

4.1.2. Models

4.1.2.1 Image-Mask Generation:

For the generation of images and masks, we selected Stable Diffusion-2.0 and fine-tuned it using Low-Rank Adaptation (LoRA) [7]. Among the various open-source models in the Stable Diffusion series, SD1.4 and SD1.5 employ a general CLIP model, which demonstrates suboptimal text-image alignment. SD2.1 utilizes OpenCLIP, offering improved performance; however, it requires input and output resolutions of 768 pixels and occasionally generates fish-scale artifacts. SDXL and SD3.5, due to their extensive parameter counts [28], are incompatible with consumer-grade GPUs. Thus, we opted for SD2.0 as the base model for modification and fine-tuning, balancing performance and hardware compatibility.

4.1.2.2 Classification:

For classification tasks, we utilized four models: DenseNet121, ResNet34, SqueezeNet1.1, and Swin Transformer, each fine-tuned using open-source pre-trained weights with adjustments to the final layer.

4.1.2.3 Segmentation:

For segmentation tasks, we adopted U-Net and DeepLabV3+ models, fine-tuning their final layers using open-source pre-trained weights.

4.1.3. Metrics

4.1.3.1 Image-Mask Generation:

To evaluate the similarity between generated and real images, we adopt three widely recognized metrics: Fréchet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), and Multi-Scale Structural Similarity (MS-SSIM).

4.1.3.2 Classification:

For classification tasks, we employ four standard metrics: Accuracy, Sensitivity, Precision, and F1-score .

4.1.3.3 Segmentation:

For segmentation tasks, we utilize four established metrics: Dice Coefficient, Intersection over Union (IoU), Average Surface Distance (ASD), and Hausdorff Distance (HD).

4.2. Fine-tuning and Making Expanded Datasets

This section outlines the experimental setup leveraging the pretrained Stable Diffusion-2.0 checkpoint, a text-to-image generative model sourced from the Hugging Face, chosen for its robust generative capabilities.

For fine-tuning, we employed the novel ISIC-GPT dataset introduced in the previous section. The Stable Diffusion-2.0 model was adapted to process four-channel inputs and outputs, with LoRA applied to efficiently fine-tune the attention layers of the UNet component. After evaluating ranks of 2, 4, and 8, we selected a rank of 4 for LoRA fine-tuning of the attention layers, as it achieved an

optimal balance between image fidelity and structural consistency. The base UNet and VAE parameters remained frozen, with only the LoRA parameters trained, substantially reducing the number of trainable parameters, computational cost, and training duration. RGB images were resized to 256×256 using LANCZOS interpolation, while L-format segmentation masks were resized using nearest-neighbor interpolation. Data augmentation was conducted using the Albumentations, applying pixel-level transformations and spatial transformations, synchronized across RGB images and masks. During data loading, a batch size of 4 was used with 12 worker processes to accelerate retrieval. Fine-tuning spanned 100 epochs with a learning rate of 1e-4, a batch size of 4, and the AdamW optimizer was employed for efficiency. Training was performed on an NVIDIA GeForce RTX 4090 GPU, completing in approximately 5 hours.

In the inference phase, four-channel outputs were generated and split into RGB and L channels. RGB channels were denormalized to their original range, while the L channel was processed via a Sigmoid function with a 0.5 threshold to yield a binary mask, saved in PNG format. The prompt structure guiding the diffusion model was "a dermoscopic lesion photo of {class_name} for skin cancer diagnosis," where {class_name} represents one of seven skin lesion categories. Random seeds were used to enhance output diversity, with a resolution of 512×512 pixels and the DDIM scheduler. We used the Optuna optimizer for Bayesian hyperparameter tuning, we identified num_inference_steps=45 and guidance_scale=1.22 as the optimal settings. Image quality assessment for the synthetic dataset, presented in TABLE I, involved computing three metrics between the real training and test sets, and between the real training and synthetic training sets. The synthetic training set's FID (68.601) is higher than the real test set's (42.058), indicating some distributional divergence. However, this value remains acceptable for downstream tasks, as evidenced by performance gains in classification and segmentation. Clinically, the synthetic images retain sufficient realism and diversity, supporting model training effectively. Compared to GAN-based methods, our approach may exhibit slightly higher FID but offers greater sample diversity and stability, avoiding mode collapse—a common GAN limitation. The LPIPS and MS-SSIM metrics, though slightly irregular due to the modest dataset size, show comparable values across groups, suggesting sufficient realism in the generated data. Fig. 2 offers a visual comparison of synthetic images and masks against real ones, revealing that synthetic skin lesion images are nearly indistinguishable from real images to the naked eye, underscoring their high fidelity and potential as real-data substitutes.

Table 1. Image Quality Assessment for Generated Images

Data for Calculations	FID	LPIPS	MS-SSIM
Real Train vs Real Test	42.058	0.545	0.190
Real Train vs Syth Train	68.601	0.524	0.210

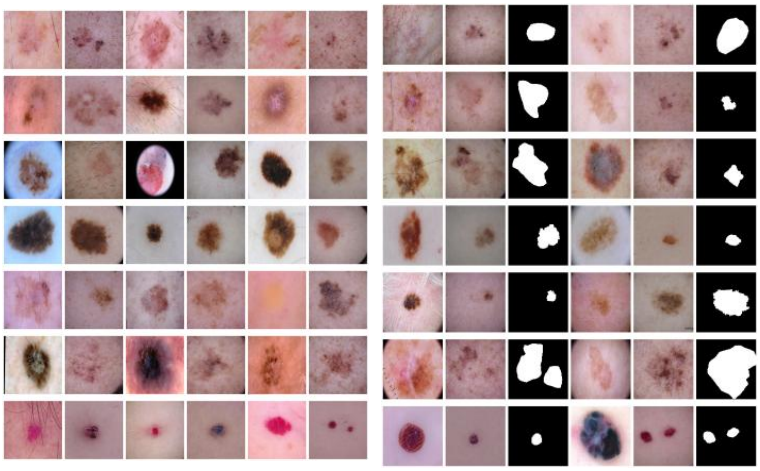

Figure 2. Comparison of Image-Mask Pairs (One row corresponds to one category): Original & Generated Image(left), Original, Generated Image & Mask(right)

Consequently, we established three datasets—real, synthetic, and hybrid—all of identical scale, comprising skin lesion images and corresponding masks, primed for subsequent comparative experiments.

4.3. Quantitative and Qualitative Analysis for Tasks

4.3.1. Training and Evaluation for Classification Tasks

In this study, the training data was organized into three configurations: real data only, synthetic data only, and a hybrid dataset comprising 50% real and 50% synthetic data. Input images were resized to 224×224 pixels and augmented, followed by normalization. The test set consisted exclusively of real data, resized and normalized without further augmentation. Training was conducted with a batch size of 32, spanning 50 epochs for DenseNet121 and ResNet34, and 40 epochs for SqueezeNet1.1 and Swin Transformer. The learning rate was set to 0.0001. A 5-fold Stratified K-Fold cross-validation approach ensured balanced class representation, with all metrics calculated via macro-averaging. The quantitative results are shown in TABLE II. Overall, the hybrid dataset, integrating the clinical relevance of real data with the diversity of synthetic data, mitigated the distributional constraints of real-only data and the domain gaps of synthetic-only data. This approach enhanced classification performance across all models, with accuracy and F1-score improvements ranging from 8% to 15%. Increased data diversity effectively reduced overfitting and improved generalization. The suboptimal results with synthetic-only data likely stem from distributional shifts or label noise relative to the real test set, impeding generalization. Employing diffusion models for data augmentation sharpened inter-class distinctions and reduced intra-class variability, thereby boosting classification accuracy.

Table 2. Comparison of classification and segmentation performance on real\synth\hybrid data

Models	Train Dataset	Accuracy	Sensitivity	Precision	F1score
DenseNet121	Real only	71.257	71.257	73.775	71.289
	Synth only	50.057	50.057	58.547	46.875
	50%Real+50%Synth	**80.286**	**80.286**	**82.160**	**80.446**
ResNet34	Real only	71.600	71.600	74.504	71.609
	Synth only	45.943	45.943	60.142	43.055
	50%Real+50%Synth	**77.257**	**77.257**	**79.340**	**77.454**
Squeezenet1.1	Real only	63.943	63.943	67.800	63.925
	Synth only	47.771	47.771	53.213	45.012
	50%Real+50%Synth	**71.029**	**71.029**	**73.747**	**70.862**
Swin Transformer	Real only	74.571	74.571	77.642	74.021
	Synth only	57.143	57.143	63.506	56.027
	50%Real+50%Synth	**80.971**	**80.971**	**82.690**	**80.901**
Models	Train Dataset	Dice	IoU	ASD	HD
U-Net	Real only	79.397	65.836	**58.351**	**95.775**
	Synth only	64.459	47.575	64.303	114.456
	50%Real+50%Synth	**80.122**	**66.840**	59.101	97.882
DeepLabV3+	Real only	81.373	68.595	57.063	92.113
	Synth only	67.968	51.515	66.556	112.226
	50%Real+50%Synth	**81.951**	**69.425**	**56.023**	**89.778**

4.3.2. Training and Evaluation for Segmentation Tasks

For segmentation tasks, we employed U-Net and DeepLabV3+ models, built on ResNet34 and ResNet50 backbones, respectively. The data configurations mirrored those of the classification tasks, with images resized to 512×512 pixels. Training utilized a batch size of 8 over 20 epochs, with a learning rate of 0.001 and Dice Loss. The quantitative results are shown in TABLE II. Overall, the hybrid dataset, by incorporating synthetic mask diversity, refined the models' capacity to predict

edges and fine structures, addressing the shortcomings of real-only data in complex lesion scenarios. This resulted in Dice and IoU improvements of 0.5%–1% and substantial reductions in ASD and HD (5–10 units). The poorer performance of synthetic-only data likely arises from geometric distortions in mask generation or domain mismatches with the real test set, leading to imprecise boundary predictions. Dice Loss optimized overlap regions effectively, and mixed-precision training accelerated convergence. DeepLabV3+ consistently outperformed U-Net across all metrics, particularly with the hybrid dataset, achieving a Dice coefficient of 81.951% and HD of 89.778, highlighting its efficacy in complex medical image segmentation.

4.3.3. Evaluating the Robustness on HAM10000

TABLE III presents the robustness evaluation of our models on the HAM10000 dataset, encompassing both classification and segmentation tasks. The hybrid dataset generally enhanced sensitivity and F1-scores in classification tasks, improving positive class detection and overall model performance. However, due to the limited sample size of the training set, accuracy improvements were modest, with some models (e.g., ResNet34) even experiencing slight declines, highlighting the complexity of the HAM10000 dataset. In segmentation tasks, the Dice coefficient and IoU showed limited gains (approximately 1%), with overall performance remaining low (Dice < 31%). This may be attributed to constraints in image resolution or annotation quality. Resolution mismatch (600×450 in HAM10000 vs. 512×512 in training) may cause domain shift. Future work could adopt resolution normalization. Among the models, Swin Transformer and DeepLabV3+ demonstrated relatively superior performance on the hybrid dataset, suggesting potential for further robustness improvements through optimized data preprocessing.

4.3.4. Visualizing for Classification Tasks

Fig. 3 illustrates heatmaps generated using three explainable AI (XAI) techniques—GradCAM, Saliency, and Occlusion—to interpret the decision-making processes of our classification models on the test set. The hybrid dataset enhances the feature space through increased diversity, mitigating overfitting and directing model attention toward pathologically relevant regions. The complementary insights from GradCAM (class-discriminative localization), Saliency (pixel sensitivity), and Occlusion (impact of occlusion) collectively strengthen model interpretability and trustworthiness.

Table 3. Results Evaluating The Robustness on HAM10000

Models	Dataset	Accuracy	Sensitivity	Precision	F1score
DenseNet121	Real	34.930	59.227	38.660	37.633
	Hybrid	**35.990**	**66.391**	**41.249**	**40.765**
ResNet34	Real	**35.968**	59.645	39.001	37.898
	Hybrid	34.333	**65.071**	**39.547**	**38.655**
Squeezenet1.1	Real	25.368	51.408	29.833	27.495
	Hybrid	**27.359**	**57.198**	**35.381**	**33.793**
Swin Transformer	Real	34.011	**66.194**	39.908	38.197
	Hybrid	**34.201**	60.199	**40.169**	**39.208**
Models	Dataset	Dice		IoU	
U-Net	Real	29.878		17.568	
	Hybrid	**30.953**		**18.321**	
DeepLabV3+	Real	29.399		17.235	
	Hybrid	**30.110**		**17.730**	

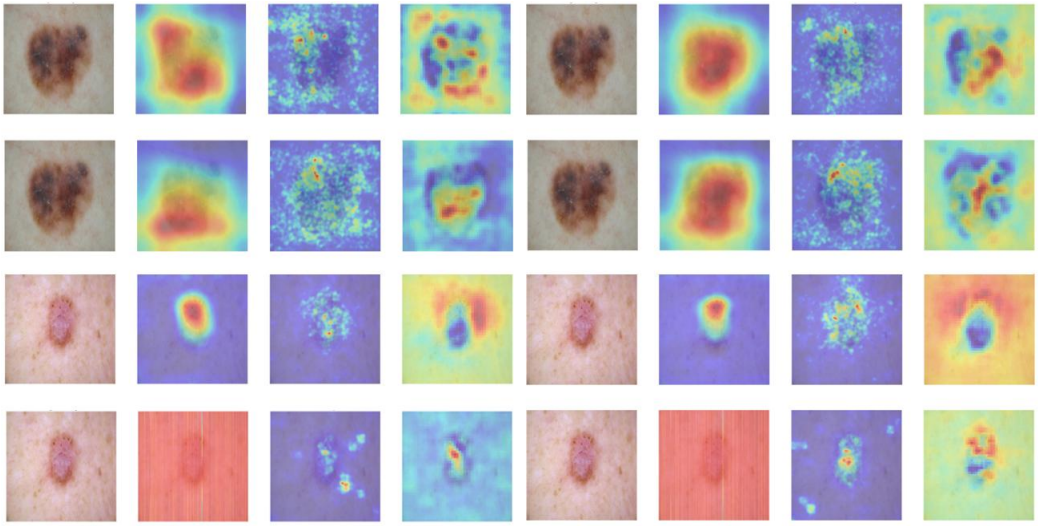

Figure 3. Heatmaps of XAI methods (Gradcam, Saliency, Occlusion) for the testset: One row corresponds to one model, the first four images show models trained on the real dataset only, and the next four show models trained on the hybrid dataset. Models trained solely on synthetic data were excluded due to poor performance. Visualized samples are ISIC_0010034 and ISIC_0030261, with each set of four images including the original image followed by attention maps from the three XAI methods.

5. Conclusion

This paper's experiments demonstrate that SkinDualGen significantly enhances the performance of classification and segmentation models by generating high-quality synthetic skin lesion images and masks, effectively addressing the challenges of data scarcity and class imbalance. But limitations exist, including potential biases in synthetic data, which may lead to uneven model performance across populations. Second, occasional geometric distortions in masks require refined prompt engineering, increasing application complexity. Future research could extend SkinDualGen to other medical imaging modalities like CT and MRI, strengthen privacy protection and fairness and 3D medical image generation, ensuring wider clinical applicability.

References

[1] Shorten, Connor, Khoshgoftaar, Taghi M. A survey on image data augmentation for deep learning. Journal of big data, 6(1): 1-48, 2019.

[2] Koetzier, Lennart R, Wu, Jie, Mastrodicasa, Domenico, et al. Generating synthetic data for medical imaging. Radiology, 312(3): e232471, 2024.

[3] Ibrahim, Mahmoud, Al Khalil, Yasmina, Amirrajab, Sina, et al. Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. Computers in biology and medicine, 189: 109834, 2025.

[4] Ktena, Ira, Wiles, Olivia, Albuquerque, Isabela, et al. Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine, 30(4): 1166-1173, 2024.

[5] Mutepfe, Freedom, Kalejahi, Behnam Kiani, Meshgini, Saeed, et al. Generative adversarial network image synthesis method for skin lesion generation and classification. Journal of Medical Signals \& Sensors, 11(4): 237-252, 2021.

[6] Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, et al. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022.

[7] Edward J Hu, yelong shen, Phillip Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations, 2022.

[1] Kingma, Diederik P, Welling, Max, et al. Auto-encoding variational bayes. , 2013.

[8] Baur, Christoph, Denner, Stefan, Wiestler, Benedikt, et al. Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. Medical image analysis, 69: 101952, 2021.

[9] Goodfellow, Ian J, Pouget-Abadie, Jean, Mirza, Mehdi, et al. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[10] Yi, Xin, Walia, Ekta, Babyn, Paul. Generative adversarial network in medical imaging: A review. Medical image analysis, 58: 101552, 2019.

[11] Ho, Jonathan, Jain, Ajay, Abbeel, Pieter. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33: 6840-6851, 2020.

[12] Kazerouni, Amirhossein, Aghdam, Ehsan Khodapanah, Heidari, Moein, et al. Diffusion models in medical imaging: A comprehensive survey. Medical image analysis, 88: 102846, 2023.

[13] Bozorgpour, Afshin, Sadegheih, Yousef, Kazerouni, Amirhossein, et al. Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation. International workshop on predictive intelligence in medicine, 2023.

[14] Yu, Yongrui, Gu, Yannian, Zhang, Shaoting, et al. MedDiff-FM: A Diffusion-based Foundation Model for Versatile Medical Image Applications. arXiv preprint arXiv:2410.15432, 2024.

[15] Chlap, Phillip, Min, Hang, Vandenberg, Nym, et al. A review of medical image data augmentation techniques for deep learning applications. Journal of medical imaging and radiation oncology, 65(5): 545-563, 2021.

[16] Abdelhalim, Ibrahim Saad Aly, Mohamed, Mamdouh Farouk, Mahdy, Yousef Bassyouni. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. Expert Systems with Applications, 165: 113922, 2021.

[17] Shin, Hoo-Chang, Tenenholtz, Neil A, Rogers, Jameson K, et al. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, 2018.

[18] Montoya-del-Angel, Ricardo, Sam-Millan, Karla, others. MAM-E: Mammographic synthetic image generation with diffusion models. Sensors, 24(7): 2076, 2024.

[19] Khader, Firas, Mu. Denoising diffusion probabilistic models for 3D medical image generation. Scientific Reports, 13(1): 7303, 2023.

[20] Wang, Haoshen, Liu, Zhentao, Sun, Kaicong, et al. 3D MedDiffusion: A 3D Medical Diffusion Model for Controllable and High-quality Medical Image Generation. arXiv preprint arXiv:2412.13059, 2024.

[21] Sagers, Luke W, Diao, James A, Groh, Matthew, et al. Improving dermatology classifiers across populations using images generated by large diffusion models. arXiv preprint arXiv:2211.13352, 2022.

[22] Akrout, Mohamed, Gyepesi, Balint, Hollo, Peter, et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2023.

[23] Zhang, Lvmin, Rao, Anyi, Agrawala, Maneesh. Adding conditional control to text-to-image diffusion models. Proceedings of the IEEE/CVF international conference on computer vision, 2023.

[24] Ruiz, Nataniel, Li, Yuanzhen, Jampani, Varun, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023.

[25] ISIC Archive, "ISIC 2020 Challenge Dataset," [Online]. Available: https://challenge.isic-archive.com/data/#2020, 2020.

[26] Medghalchi, Yasamin, Zakariaei, Niloufar, Rahmim, Arman, et al. MEDDAP: Medical Dataset Enhancement via Diversified Augmentation Pipeline. arXiv preprint arXiv:2403.16335, 2024.

[27] Dustin Podell, Zion English, Kyle Lacey, et al. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. The Twelfth International Conference on Learning Representations, 2024.