Enhanced Diffusion Model with Multi-level Embeddings for Medical Image Data Augmentation in Skin Disease

Mujung Kim[1], Jisang Yoo[1], Soon Chul Kwon[2], Byung Jun Kim[3],
Changsik John Pak[4], Chong Hyun Won[4], Suk-Ho Moon[5], Woo Jin Song[6],
Han Gyu Cha[6], Kyung Hee Park[7]

[1]*Dept. of Electronic Engineering, Kwangwoon Univ., Seoul, Republic of Korea*
[2]*Graduate School of Smart Conv., Kwangwoon Univ., Seoul, Republic of Korea*
[3]*Dept. of Reconstr. & Plastic Surg., Seoul Natl. Univ. Hospital, Seoul, Republic of Korea*
[4]*Dept. of Plastic & Reconstr. Surg., Asan Medical Ctr., Seoul, Republic of Korea*
[5]*Dept. of Plastic & Reconstr. Surg., Seoul St. Mary's Hospital, Seoul, Republic of Korea*
[6]*Dept. of Plastic & Reconstr. Surg., Soonchunhyang Univ. Hospital, Republic of Korea*
[7]*Dept. of Nursing Science, The Univ. of Suwon, Hwaseong, Republic of Korea*

Emails: [1]{kmj1026, jsyoo}@kw.ac.kr, [2]ksc0226@kw.ac.kr,
[3]bjkim79@snu.ac.kr, [4]{iloveps.jcp, drwon}@amc.seoul.kr, [5]nasuko@catholic.ac.kr,
[6]{pswjsong, prs.cha}@schmc.ac.kr, [7]khpark@suwon.ac.kr

***Abstract*—Deep learning has enabled applications in medical diagnosis, education, and research. However, obtaining large-scale, high-quality data remains challenging due to privacy regulations and the scarcity of rare disease data.**

Recent approaches focus on deep learning-based image generation models to create synthetic data, increasing its diversity and quality for medical applications. This study proposes an improved diffusion-based model for high-quality image generation across diverse domains. Inspired by the 8-channel VAE from Mefusion, we modified the VAE structure in Stable Diffusion to reduce artifacts. To address the loss of detailed representations in the Latent Diffusion model's compression process, we introduced multi-level embeddings and adapter layers. These additions improved synthetic data quality in the dermatology domain.

Using the HAM10000 dataset, we generated synthetic data for seven skin disease conditions and conducted classification experiments to evaluate its utility. The classification accuracy using synthetic data alone was comparable to using original data. Training with both synthetic and original data improved accuracy from 87% to 90%.

Our results confirm that synthetic data from our diffusion model is effective for dermatological training. Visual and quantitative evaluations further highlight its potential for medical applications.

***Index Terms*—Diffusion, Data augmentation, Skin disease classification, Medical imaging**

I. Introduction

With the rapid advancement of deep learning technology in recent years, there has been an active movement to adopt it in the medical field [1]–[3]. Deep learning is being applied in various medical areas such as diagnosis [4]–[7], lesion detection [8], [9], and image segmentation [10], [11], and high-quality data that accurately reflects disease characteristics is essential for effective training of these models.

However, obtaining data in the medical field is extremely challenging due to various constraints, including privacy protection issues, lack of data for rare diseases, and strict regulations in clinical data collection processes. To overcome these limitations, recent studies [12] focusing on generating synthetic data using deep learning-based generative models to supplement insufficient medical data has been gaining attention.

In recent years, within the field of deep learning image generation, Diffusion-based models [13] have gained significant attention by demonstrating higher training stability, diversity, and High quality image synthesis capabilities compared to previous generative methods such as Variational Autoencoder (VAE) [14] and Generative Adversarial Networks (GAN) [15]. Diffusion models not only offer easier training and high-quality image generation but have also shown exceptional performance in various tasks including image editing [16], super-resolution [17], and style transfer [18]. In the medical field, research utilizing these Diffusion-based generative models to create high-quality synthetic data has been actively progressing, with studies effectively generating synthetic data for various diseases using models such as Imagen [19] and Stable Diffusion [20]. Notably, the study in [21] improved upon Stable Diffusion's VAE architecture to generate synthetic images of chest X-rays, iris lesions, and histopathological images, demonstrating superior generation quality when compared to GAN-based models.

Diffusion-based models are capable of generating realistic images by effectively capturing overall image information. however, they still face challenges in retaining fine-grained details. To address this issue, methods for maintaining detailed

representations in image generation have been actively studied.

ELITE [22] preserves fine details during personalized image generation in the general image domain by employing attention operations between text prompts and local features extracted through a multi-layer network. Similarly, Instantbooth [23] improves fine-grained representations during personalized image generation by adding an adapter layer to the diffusion denoising network to incorporate rich patch information from the input image.

In the medical field, subtle differences in fine details can lead to critical consequences, making the preservation of detailed representations even more crucial. Therefore, it is essential that synthetic medical images reflect patterns, textures, and shapes of the lesion area accurately to ensure that disease-specific details are not lost. In medical applications, not only is the fast generation of realistic images important, but generating images with the precision necessary for clinical use is also critical. Otherwise, the generated images may not be suitable for real-world applications, limiting their clinical contribution.

In this paper, we propose an improved diffusion-based method for the skin disease domain by modifying the VAE structure of the Latent Diffusion Model and incorporating lesion masks and multi-level embeddings. We utilized the HAM10000 [24] dataset, which contains data for seven types of skin tumors, to enhance the ability of the Stable Diffusion model to preserve detailed representations.

First, inspired by the artifact improvement approach suggested by Medfusion [21], we adopted the VAE channel expansion technique of Stable Diffusion. In addition, we introduced multi-level embeddings extracted from lesion images to the diffusion training process to effectively learn detailed representations from the original data. These learned representations were incorporated into the adaptive layer during the diffusion denoising process to be utilized in synthetic data generation. We also applied lesion masks to extract only the lesion area from the input image, enabling focused learning on the lesion region.

To verify the effectiveness of the generated data, we visually inspected the results and used the synthetic data obtained with the proposed method to train five representative classification models, including VGG [25] and ResNet [26]. We compared their classification accuracy with models trained using only the original data. When using the same amount of data, the models trained with synthetic data showed performance comparable to those trained with original data. Furthermore, when augmenting the original dataset with a larger amount of synthetic data, accuracy increased from 87% to 90%, achieving the highest accuracy. This demonstrates that synthetic data can play an important role when data is limited, such as in cases of rare diseases or when privacy issues make data collection challenging.

The proposed method is similar to ELITE in terms of preserving detailed representations using multi-level embeddings but differs in several aspects. While ELITE focuses on personalized image generation, our study emphasizes data augmentation, aiming to generate diverse sample outputs rather than fixed targets. Additionally, whereas ELITE uses a local mapping network to focus on relationships within the text space, we added an adapter layer to the diffusion U-Net structure to focus on the utilization of visual tokens. Finally, while ELITE was applied to general image domains, we focused on the medical domain, specifically generating synthetic data for skin disease images and conducting downstream tasks with this data.

We make the following contributions:

- We introduce a novel enhancement to diffusion networks through a dedicated Multi Level embeddings, specifically designed for preserving fine-grained characteristics of skin lesions. This architecture, combined with lesion-specific masks, significantly improves the preservation of critical disease features including structural patterns, textures, and subtle pathological indicators. The integration of these components enables more accurate representation of disease-specific details that are crucial for medical diagnosis.
- We demonstrate the practical effectiveness of our approach through comprehensive evaluation using four widely-adopted classification models including VGG and ResNet architectures. Our experimental results demonstrate that when we secured a larger amount of data through synthetic data compared to using original data alone, the classification accuracy significantly improved from 87% to 90%. This improvement validates both the quality of our synthetic data and its utility in addressing data scarcity issues in medical imaging applications.

II. Related Work

A. Medical Data Augmentation

Data augmentation is a technique that secures additional data needed to improve model performance by utilizing original data to address issues such as overfitting and data imbalance that can occur when training data is insufficient. Traditional data augmentation techniques include rotation, resizing, translation, flipping, and affine transformations.

However, medical data has specialized and complex characteristics, making it difficult to generate new essential features through simple geometric transformations alone. Moreover, medical data collection is more challenging compared to other fields due to privacy protection issues and data availability constraints.

To overcome these limitations, recent advances in generative deep learning technology have led to the development of data augmentation methods through deep learning-based synthetic data generation. After GANs [12] gained attention for their high-quality synthesis capabilities, various GAN-based synthetic data augmentation methods [27]–[29] have been studied in the medical data field. In [30], skin cancer data synthesis was conducted using STGAN, which combines universal knowledge with class-specific knowledge.

Recently, Diffusion [13] has gained attention for producing high-quality realistic data through stable training that avoids the unstable learning issues of GANs, and research utilizing this for medical synthetic data augmentation [31], [32] is also increasing. In [21], the VAE structure of the Latent Diffusion Model(LDM) [20] was expanded to 8 channels to improve artifacts in generated medical images, and demonstrated better performance than GAN-based models in quantitative metrics such as FID across various medical imaging fields including retinal images, colorectal cancer histology images, and chest X-rays.

B. Diffusion Models

Diffusion Probability Models (DPM) [33] introduced the principle of learning data distribution using Markov chains, specifically transforming a simple known probability distribution into a target distribution. Denoising Diffusion Probabilistic Models (DDPM) [13] expanded this concept to specialize in image generation. DDPM has been receiving prominent attention in computer vision, particularly in the field of image generation. Unlike GAN-based models that suffered from learning stability issues such as mode collapse, DDPM demonstrated advantages by showing excellent realism and diversity while maintaining stable training. DDPM enabled high-quality image generation by optimizing the denoising process. To address DDPM's limitation of requiring long progressive sampling time, Denoising Diffusion Implicit Models (DDIM) [34] enabled high-quality image generation with fewer diffusion steps through deterministic sampling of noisy latent variables. In [35], the diffusion model was reinterpreted from a Stochastic Differential Equation (SDE) perspective, extending DDPM's discrete-time diffusion process to a continuous domain, allowing for more flexible sampling and various noise schedules. The study in [36] surpassed GANs in terms of image diversity and quality, introducing Classifier Guidance to enable conditional generation. Latent Diffusion Models (LDM) [20] increased computational efficiency by performing the diffusion process in compressed latent space rather than pixel space, bringing innovation to high-resolution image generation. It also enabled flexible conditional image generation through text prompts. Diffusion-based models are being utilized in various studies for medical data augmentation, as shown in [21], [31], leveraging these advantages. In this paper, inspired by previous research that demonstrated excellent potential in diffusion-based medical data synthesis, we conducted research using the LDM model Stable-Diffusion.

III. Method & Material

A. Overview

The goal of our model is to reduce the loss of detail representations and improve artifacts in Diffusion-based image synthesis to generate high-quality medical synthetic data. Our model's workflow is shown in Fig.1. The process consists of the following steps: First, to improve artifacts in generated images, we expand and pre-train stable diffusion's existing 4-channel VAE to 8 channels. Next, for higher focus on the disease itself, we use mask images corresponding to the lesion area as a preprocessing step to remove unnecessary background and extract lesion images that contain only the skin lesion region. To obtain multi-level detailed expressions from the extracted lesion images, we extract multi-level embeddings through the CLIP [37] image encoder composed of various levels. To incorporate these detailed expressions into Diffusion training, we add separate adapter layers to the Transformer blocks within the Diffusion U-net, through which the detailed features of the lesions are transmitted to the Diffusion Model. Each component is explained in the following sections.

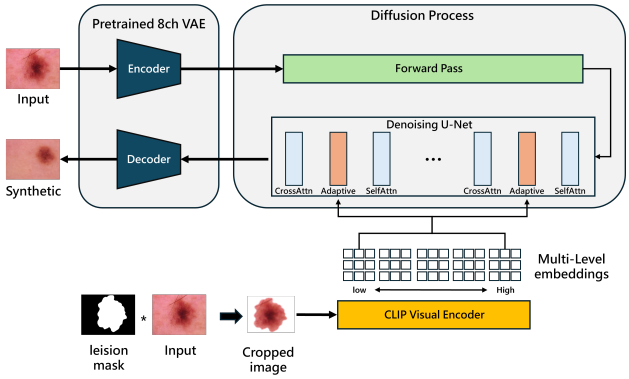

Fig. 1. **Workflow for proposed method.** A pre-trained 8-channel VAE enhances image quality, while lesion masks isolate the lesion area to produce cropped images. These cropped images are transformed into multi-level embeddings using a CLIP Visual Encoder, which are then integrated into the diffusion process through adapter layers in the Denoising U-Net. This enables the model to effectively learn and preserve detailed features such as lesion patterns, textures, and structures.

B. Pre-train 8ch-VAE

The Variational Autoencoder (VAE) [14] is a crucial component in the Stable Diffusion model, playing a vital role in image input and output. Since the VAE directly affects the model's performance during the compression and reconstruction of image representations, improving it is essential for enhancing image quality. In medical imaging, improving artifacts and inaccuracies is particularly important, as these can interfere with accurate diagnosis. Inspired by Medfusion's demonstration that expanding the VAE embedding channels in the Stable Diffusion model from 4 to 8 channels significantly improved reconstruction quality despite a slight decrease in compression ratio, we adopted an 8-channel VAE for the skin disease domain. Expanding the VAE embedding channels from 4 to 8 enables the model to capture more information, allowing for richer feature representation. Specifically, through a higher-dimensional latent space, the model can preserve detailed image characteristics and effectively reduce distortions such as artifacts.

To pre-train the 8-channel VAE, we used a combination of KL, L1, L2, and SSIM loss functions, each serving a specific purpose in optimizing the VAE's performance. The KL divergence ensures that the latent space follows a standard Gaussian distribution, which enhances the stability and efficiency of the

sampling process. The L1 loss reduces the absolute differences between the reconstructed and original images, thereby preserving the overall structural integrity. The L2 loss minimizes squared differences, making the model sensitive to small errors and improving reconstruction accuracy. Finally, the SSIM loss preserves the structural similarity between the reconstructed and original images, which is critical for maintaining fine-grained medical details, such as lesion textures and patterns.

This combination of loss functions was chosen to maintain the original Stable Diffusion model's configuration, ensuring a balance between latent space regularization (via KL divergence) and high-fidelity reconstruction quality (via L1, L2, and SSIM losses). As this study focuses on the direct comparison between the original 4-channel VAE and the expanded 8-channel VAE, we retained the same loss configuration to isolate the impact of channel expansion.

In this research, we pre-trained the 8-channel VAE using the HAM10000 [24], which contains 7 classes, and incorporated it as the first stage of our model pipeline. This approach enhances the model's ability to represent skin disease images more accurately and mitigates the quality degradation that can occur during the image generation process through the expanded channel configuration. The pre-trained 8-channel VAE plays a critical role in generating more precise and detailed skin disease images.

Figure 2 compares the reconstruction results of the original 4-channel VAE and the 8-channel VAE on the HAM10000 dataset. By using 8 embedding channels, we observed improvements in reconstruction quality, particularly in reducing artifacts and preserving finer details in the lesion areas, compared to the 4-channel configuration.

Although the expansion to 8 channels resulted in a slight decrease in compression ratio, it also led to an approximate 1.5× increase in training and inference time. However, this trade-off is considered acceptable in the medical domain, where accurate and detailed image generation is of paramount importance. The improved image quality, including reduced artifacts and better preservation of lesion details, justifies the additional computational cost. Future work could focus on optimizing the computational efficiency of the expanded model without compromising the quality of the generated images.

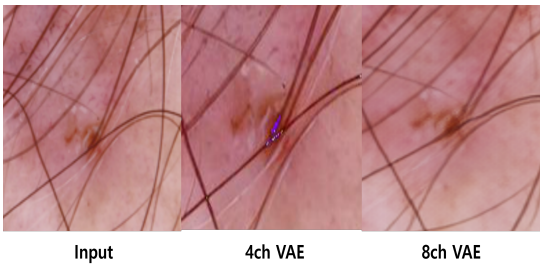

Fig. 2. Comparison of reconstructed images: input image (left), reconstruction through 4-channel VAE (middle), and 8-channel VAE (right). The 8-channel VAE demonstrates clearer and more accurate reconstruction of lesion patterns and textures.

C. Multi-Level Embeddings for Detailed Feature Representation

Existing Diffusion generative models sometimes struggle to maintain detailed features. In this study, we introduced a method to incorporate additional information about detailed features into the Diffusion learning process to address this limitation. First, to focus on disease characteristics by isolating the lesion area from the background, we extract lesion images using mask images. This ensures a clear focus on regions requiring analysis.

$$x_l := x_i \cdot m_i, \tag{1}$$

where x_i is the input image, m_i is the lesion mask image of the corresponding image, and x_l is the extracted lesion image. To extract detailed feature information from these lesion images, we generate multi-level embeddings consisting of 5 levels using a pre-trained CLIP image encoder [37]. Each layer captures different levels of features: lower layers extract fine details such as texture and color patterns, middle layers capture local structures, and upper layers capture overall shapes and structures of lesions. The transformed embeddings capture diverse characteristics of lesions, from detailed features to overall structure, playing a crucial role in preserving important details during the subsequent image generation process. Inspired by GLIGEN [38], to integrate the extracted detailed feature information into the Diffusion model, we added learnable adaptive layers to the transformer blocks of the Diffusion U-net. These adaptive layers are placed between self-attention and cross-attention, allowing detailed feature information to be effectively utilized during the learning process. The adapter layer is defined as follows:

$$v = v + \beta \cdot tanh(\gamma) \cdot SelfAttn([v, e_m]), \tag{2}$$

where v represents visual tokens, β is a constant controlling the importance of the adapter layer, γ is a learnable scalar value (initialized to 0), and e_m is the average of multi-level embeddings. Through this, the Diffusion model can learn more accurate and detailed lesion representations.

D. Datasets

The HAM10000 dataset [24] is a comprehensive dermoscopic dataset created to advance automatic classification of skin tumors, featuring data for seven types of skin tumors. This dataset includes 10,015 dermoscopic images from seven skin disease classes, including both benign and malignant tumors. The classes are:

- Melanocytic nevi(NV)
- Melanoma(MEL)
- Benign keratosis-like lesions(BKL)
- Basal cell carcinoma(BCC)
- Actinic keratoses and intraepidermal carcinoma(AKIEC)
- Vascular lesions(VASC)
- Dermatofibroma(DF)

This dataset is publicly available and widely used for training and evaluating machine learning models for skin disease classification tasks.

In this study, we utilized the HAM10000 dataset to train and evaluate our diffusion-based data augmentation model. Following standard procedures, we split the dataset into training, validation, and test sets to ensure unbiased evaluation. Additionally, to address the class imbalance issue in the dataset, we applied data augmentation techniques to generate synthetic samples, enhancing the diversity of the training data.

All images were preprocessed to standardize their size and quality. Specifically, each image was resized to 256 × 256 pixels, and normalization was applied to ensure consistent pixel intensity across the dataset. Furthermore, we utilized the lesion masks provided within the dataset for multi-level embedding, enabling the model to focus on relevant regions and preserving details crucial for accurate classification.

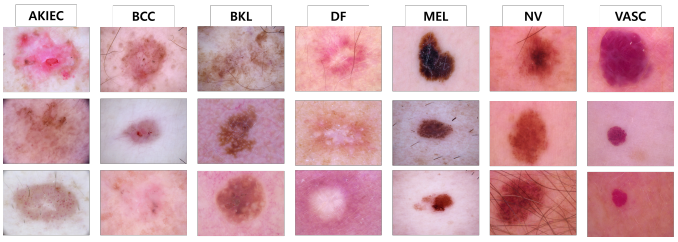

Fig. 3. The HAM10000 dataset: consisting of 10,015 images of seven types of skin tumors.

E. Implementation Details

While using Stable Diffusion as our base model, we modified its components to suit our objectives. The VAE was pretrained using the Adam optimizer to encode 224×224 images into a 32×32 latent space, combining KL, L1, L2, and SSIM losses with a batch size of 16. The diffusion process consisted of forward (encoding to Gaussian noise over 1000 steps) and backward (denoising) stages, trained with AdamW optimizer and L1 loss to supervise the difference between actual and estimated noise distributions. During the sampling stage, we used 150 steps with a batch size of 8, setting $\beta = 1$ and initial $\gamma = 0$ for adapter training. Implementation used Python 3.9.19 with PyTorch 1.12 and CUDA 11.4 on two NVIDIA RTX 3090 GPUs with a learning rate of 1e-5.

IV. Experiment & Result

A. Visual Result

As shown in Figure 4, we generated images of 7 classes of skin diseases from the HAM10000 dataset and compared them visually with the original images. With the introduction of multi-level embedding, it is evident that detailed lesion characteristics are generally well preserved. However, in the case of 'DF' (Dermatofibroma), the method failed to perfectly capture subtle textures such as small wounds around the lesion. This suggests a limitation in capturing fibroma-like indented or raised textures in detail, compared to characteristics like color or patterns. Nevertheless, it was confirmed that important features such as the lesion's color and shape were well maintained through multi-level embedding.

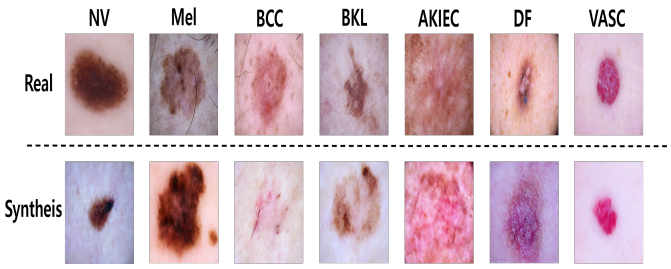

Fig. 4. Comparison of real images (upper row) and synthetic images generated by our model (lower row) for 7 skin diseases. Detailed features such as colors and patterns are well preserved across 6 classes, except for DF.

B. Downstream task

To validate the effectiveness of synthetic data, we created four types of classification datasets with 250 original images, 500 original images, 500 synthetic images, and 1000 hybrid images combining synthetic and original data. We trained five common classification models, including VGG [25]and ResNet [26], and compared the classification performance. The dataset configurations were as follows:

- Original250: 250 original images
- Original500: 500 original images
- Synthetic500: 500 synthetic images
- Hybrid1000: original500 + synthetic 500

As shown in Table I, using synthetic data alone showed similar or slightly decreased classification accuracy compared to using only original data. This is due to the fact that synthetic data cannot perfectly reproduce real-world characteristics. However, the main purpose of this study was to validate the utility of synthetic data to address data scarcity. To this end, we trained models using a hybrid dataset (Hybrid1000) that combined 500 original images with 500 synthetic images. The hybrid data achieved the highest classification accuracy. We quantitatively confirmed the effectiveness of synthetic data as it demonstrated that the average classification accuracy increased from about 87% to 90% when using sufficient training data supplemented with synthetic data compared to using original data alone. These results suggest that synthetic data can help mitigate data scarcity in the medical domain.

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY USING FIVE COMMON CLASSIFICATION MODELS ACROSS FOUR DATASETS COMPOSED OF REAL AND SYNTHETIC DATA.

Dataset	VGG13	VGG16	VGG19	ResNet18	ResNet34	Avg
Origin 250	83.24	83.79	86.55	83.51	83.95	84.08
Origin 500	84.63	85.23	87.55	88.43	89.32	86.91
Synthetic 500	82.75	83.98	86.21	88.21	89.14	87.01
Hybrid 1000	**89.64**	**90.03**	**90.31**	**89.85**	**90.54**	**90.07**

Conclusion

In this study, we proposed an enhanced diffusion-based model to address medical image data scarcity, particularly for

skin diseases. By expanding Stable Diffusion's VAE structure to 8 channels and introducing multi-level embeddings, our model effectively preserves detailed lesion characteristics including patterns, textures, and structures. Using the HAM10000 dataset, we generated synthetic images for 7 skin diseases and compared them with real medical data, confirming successful preservation of key features like color and shape. Experiments with various classification models demonstrated that synthetic data achieved comparable accuracy to real data, and when combined, improved classification accuracy from 87% to 90%. This shows our model's effectiveness in addressing medical data scarcity through high-quality synthetic data generation. Future research will focus on expanding to diverse medical imaging domains and improving the reproduction of fine details. This study demonstrates the potential for supporting AI model training in situations where medical data acquisition is challenging.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2023-00258639) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

[1] X. Chen, et al., "Recent advances and clinical applications of deep learning in medical image analysis," Medical Image Analysis, vol.79, pp.102444, 2022.

[2] S.K. Zhou, H. Greenspan and D. Shen, "Deep learning for medical image analysis," Academic Press, 2023.

[3] R. Aggarwal, et al., "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis," NPJ Digital Medicine, vol.4, no.1, pp.65, 2021.

[4] M. Li, Y. Jiang, Y. Zhang and H. Zhu, "Medical image analysis using deep learning algorithms," Frontiers in Public Health, vol.11, pp.1273253, 2023.

[5] S. Aslani and J. Jacob, "Utilisation of deep learning for COVID-19 diagnosis," Clinical Radiology, vol.78, no.2, pp.150-157, 2023.

[6] L. Huang, et al., "Rapid, label-free histopathological diagnosis of liver cancer based on Raman spectroscopy and deep learning," Nature Communications, vol.14, no.1, pp.48, 2023.

[7] Y. Kim, et al., "Adaptive Collaboration Strategy for LLMs in Medical Decision Making," arXiv preprint arXiv:2404.15155, 2024.

[8] L. Gaur, U. Bhatia, N.Z. Jhanjhi, G. Muhammad and M. Masud, "Medical image-based detection of COVID-19 using deep convolution neural networks," Multimedia Systems, vol.29, no.3, pp.1729-1738, 2023.

[9] J. Wolleb, F. Bieder, R. Sandkühler and P.C. Cattin, "Diffusion models for medical anomaly detection," International Conference on Medical Image Computing and Computer-Assisted Intervention, pp.35-45, 2022.

[10] Y. Ye, et al., "MedUniSeg: 2D and 3D Medical Image Segmentation via a Prompt-driven Universal Model," arXiv preprint arXiv:2410.05905, 2024.

[11] T. Koleilat, et al., "MedCLIP-SAMv2: Towards Universal Text-Driven Medical Image Segmentation," arXiv preprint arXiv:2409.19483, 2024.

[12] Y. Chen, et al., "Generative adversarial networks in medical image augmentation: a review," Computers in Biology and Medicine, vol.144, pp.105382, 2022.

[13] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol.33, pp.6840-6851, 2020.

[14] D.P. Kingma, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[15] I. Goodfellow, et al., "Generative adversarial nets," Advances in Neural Information Processing Systems, vol.27, 2014.

[16] G. Couairon, J. Verbeek, H. Schwenk and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," arXiv preprint arXiv:2210.11427, 2022.

[17] H. Chung, E.S. Lee and J.C. Ye, "MR image denoising and super-resolution using regularized reverse diffusion," IEEE Transactions on Medical Imaging, vol.42, no.4, pp.922-934, 2022.

[18] Z. Wang, L. Zhao and W. Xing, "Stylediffusion: Controllable disentangled style transfer via diffusion models," Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[19] C. Saharia, et al., "Photorealistic text-to-image diffusion models with deep language understanding," Advances in Neural Information Processing Systems, vol.35, pp.36479-36494, 2022.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-resolution image synthesis with latent diffusion models," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10684-10695, 2022.

[21] G. Müller-Franzes, et al., "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," Scientific Reports, vol.13, no.1, pp.12098, 2023.

[22] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang and W. Zuo, "ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation," Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[23] J. Shi, W. Xiong, Z. Lin and H.J. Jung, "Instantbooth: Personalized text-to-image generation without test-time finetuning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.8543-8552, 2024.

[24] P. Tschandl, C. Rosendahl and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol.5, no.1, pp.1-9, 2018.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014

[26] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.770-778, 2016.

[27] M. Hammami, D. Friboulet and R. Kechichian, "Cycle GAN-based data augmentation for multi-organ detection in CT images via YOLO," 2020 IEEE International Conference on Image Processing (ICIP), 2020.

[28] P. Chaudhari, H. Agrawal and K. Kotecha, "Data augmentation using MG-GAN for improved cancer classification on gene expression data," Soft Computing, vol.24, pp.11381-11391, 2020.

[29] T. Pang, J.H.D. Wong, W.L. Ng and C.S. Chan, "Semi-supervised GAN-based radiomics model for data augmentation in breast ultrasound mass classification," Computer Methods and Programs in Biomedicine, vol.203, pp.106018, 2021.

[30] Q. Su, H.N.A. Hamed, M.A. Isa, X. Hao and X. Dai, "A GAN-based data augmentation method for imbalanced multi-class skin lesion classification," IEEE Access, 2024.

[31] M. Akrout, 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, pp.81-90, 2023.

[32] A. Kebaili, J. Lapuyade-Lahorgue and S. Ruan, "Deep learning approaches for data augmentation in medical imaging: a review," Journal of Imaging, vol.9, no.4, pp.81, 2023.

[33] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," International Conference on Machine Learning, pp.2256-2265, 2015.

[34] J. Song, C. Meng and S. Ermon, "Denoising diffusion implicit models," arXiv preprint arXiv:2010.02502, 2020.

[35] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon and B. Poole, "Score-based generative modeling through stochastic differential equations," arXiv preprint arXiv:2011.13456, 2020.

[36] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," Advances in Neural Information Processing Systems, vol.34, pp.8780-8794, 2021.

[37] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, vol.1, no.2, pp.3, 2022.

[38] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li and Y.J. Lee, "GLIGEN: Open-Set Grounded Text-to-Image Generation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.22511-22521, 2023.