Diffusion Model-Based Data Augmentation for Lung Ultrasound Classification with Limited Data

Xiaohui Zhang[*†] XIAOHUI8@ILLINOIS.EDU
University of Illinois Urbana-Champaign, USA

Ahana Gangopadhyay[*] AHANA.GANGOPADHYAY@GE.COM
GE HealthCare, USA

Hsi-Ming Chang HSI-MING.CHANG@GE.COM
GE HealthCare, USA

Ravi Soni RAVI.SONI@GE.COM
GE HealthCare, USA

Abstract

Deep learning models typically require large quantities of data for good generalization. However, acquiring labeled medical imaging data is expensive, particularly for rare pathologies. While standard data augmentation is routinely performed to improve data variety, it may not be sufficient to improve the performance of downstream tasks with a clinical diagnostic purpose. Here we investigate the applicability of SinDDM (Kulikov et al., 2023), a single-image denoising diffusion model, for medical image data augmentation with lung ultrasound (LUS) images. Qualitative and quantitative evaluation of perceptual quality of the generated images were conducted. A multi-class classification task to detect various pathologies from LUS images was also employed to demonstrate the effectiveness of synthetic data augmentation using SinDDM. We further evaluated the image generation performance of FewDDM, an extended version of SinDDM trained on a limited number of images instead of a single image. Our results show that both SinDDM and FewDDM are able to generate images superior in quality compared to single-image generative adversarial networks (GANs), and are also highly effective in augmenting medical imaging data with limited number of samples to improve downstream task performance.

Keywords: Single Image Denoising Diffusion Model, Synthetic Image Generation, Data Augmentation, Lung Ultrasound Classification, Limited Data, Class Imbalance

1. Introduction

Recent advances in deep learning have accelerated its usage in various medical imaging applications such as classification, segmentation, anomaly detection, denoising, reconstruction, etc., across anatomies and across imaging modalities including magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), X-Ray, ultrasound, etc. (Kim et al., 2019; Suganyadevi et al., 2022; Fan et al., 2023; Ma et al., 2023; Anandasabapathy et al., 2021; Dutta et al., 2022; Fan et al., 2022). However, deep learning models are data-hungry, typically requiring hundreds or thousands of images to learn relevant patterns from high-dimensional data. Acquiring labeled medical imaging data is expensive and time-consuming due to the nature of the acquisition process, the medical expertise needed for annotation as well as privacy and security concerns. This problem becomes even more acute for rare pathologies or abnormalities, making the training data unbalanced and causing deep learning models to focus more on the majority class(es). Basic image manipulation techniques such as geometric transformations, cropping, color space augmentations, noise injections, etc. to improve data variety may not always be sufficient for medical imaging applications and may not be label-preserving in nature for all modalities or tasks (Shorten and Khoshgoftaar, 2019; Zhao et al., 2019).

[*] These authors contributed equally
[†] This work was done while the author was an intern at GE HealthCare.

A popular alternative approach is to augment training data by incorporating synthetic images generated by deep learning-based techniques like neural style transfer or generative models such as generative adversarial networks (GANs) (Ma et al., 2020), variational autoencoders (Diamantis et al., 2022) and diffusion models (Akrout et al., 2023). In particular, denoising diffusion models (DDMs) are becoming increasingly popular in data augmentation for medical imaging because of their ability to generate diverse, high-quality samples (Pinaya et al., 2022; Moghadam et al., 2023; Chambon et al., 2022). However, DDMs typically require large datasets for training in order to learn the underlying data distribution, making them unsuitable for usage in the medical imaging domain, particularly for rare abnormalities. A recent line of research focuses on adapting such models to work with a limited number of images, or even a single image. This typically requires an adjustment of the receptive field such that the model can learn internal patch statistics within a single image (Nikankin et al., 2022; Wang et al., 2022; Kulikov et al., 2023). A recent development in this context is SinDDM, a heirarchical DDM trained on multiple scales of a single image, which is able to generate diverse random samples of arbitrary dimensions from the image (Kulikov et al., 2023).

In this paper, we investigate the applicability of SinDDM in medical image data augmentation using lung ultrasound (LUS) images as a case study. Lung ultrasound is an inexpensive, reliable and non-invasive way to detect pulmonary diseases (Chavez et al., 2014; Abdalla et al., 2016), and has also been widely used to complement COVID-19 diagnosis in the recent pandemic (Smith et al., 2020; Sultan and Sehgal, 2020). Here we consider a lung ultrasound (LUS) image classification task with COVID-19, bacterial pneumonia and healthy controls using a publicly available dataset of LUS images as well as videos curated and labeled by medical experts (Born et al., 2021). The number of pneumonia images in the dataset, already less than the other classes, was further reduced to create an artificial class imbalance. This enabled us to investigate the usefulness of synthetic images generated by SinDDM in improving the downstream task performance in the scenario of severe class imbalance. We further extended SinDDM to FewDDM, training on a limited number of images instead of a single image when more than one sample is available. Extensive experiments show that both SinDDM and FewDDM are able to generate synthetic images that are able to considerably boost downstream classification performance. Furthermore, qualitative and quantitative evaluations of the generated images show that the DDM-based approaches are able to generate more realistic synthetic images in comparison to GAN-based approaches when trained on a single image, while preserving pathological markers better.

2. Related work

In this section, we review the existing literature on synthetic data generation approaches for medical imaging applications, particularly when data availability is severely limited.

2.1. Synthetic image generation in medical imaging

Synthetic image generation using deep learning techniques is becoming increasingly popular in medical imaging applications due to data availability and privacy issues inherent to the domain (Kebaili et al., 2023). Besides data augmentation, medical image generation also finds its usage in domain translation applications (Lyu and Wang, 2022; Li et al., 2023). Neural style transfer (Gatys et al., 2016) techniques have been used to generate clinical images of skin lesions with underrepresented skin colors to improve cancer detection (Rezk et al., 2022) or to generate kidney histology images to improve histological classification performance (Cicalese et al., 2020). Another popular approach is deep generative modeling, which learns the underlying data distribution and generates new data by sampling from the learned distribution. Generative adversarial networks (GANs) and their variants have been widely used to generate synthetic data in the medical imaging domain (Shin et al., 2018a,b; Sandfort et al., 2019; Tang et al., 2019; Ma et al., 2020). More recently, diffusion models have emerged as the de-facto standard in image generation owing to their ability to produce high-quality and highly diverse images (Yang et al., 2022; Croitoru et al., 2023; Dhariwal and Nichol, 2021). Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) have been used for generating synthetic high-resolution 3D brain MRIS (Pinaya et al., 2022; Dorjsembe et al., 2022), 4D temporal volume cardiac MRIs (Kim and Ye, 2022), histopathology images of brain cancer (Moghadam et al., 2023), chest X-Ray images with different abnormalities (Chambon et al.,

2022; Packhäuser et al., 2022), etc. However, these models typically need to be trained on hundreds or thousands of domain-specific images in order to learn the underlying data distribution. This is at odds with the data availability problem pertinent to the medical imaging domain, particularly for rare abnormalities.

2.2. Synthetic image generation from limited data

A number of recent works explore the use of generative models to generate synthetic images from a limited number of samples, or even one sample. Unconditional single-image GANs have been proposed previously for texture generation (Li and Wand, 2016; Bergmann et al., 2017), but do not perform well on natural images with varying textures and non-repetitive structures (Shaham et al., 2019). More recently, Shocher et al. (2019) proposed InGAN, a single-image conditional GAN for image retargeting in the natural image domain. SinGAN (Shaham et al., 2019) is an unconditional heirarchical GAN model operating on different scales of a single natural image to learn internal patch distributions within the image and generate new samples of arbitrary sizes and considerable variability. A number of approaches have also been proposed in the context of DDPMs in order to leverage the high quality and mode coverage of such models as well as utilize their conditioning capabilities. SinDDM (Kulikov et al., 2023) adopted a similar multi-scale approach as SinGAN. SinDiffusion (Wang et al., 2022) trained a diffusion model at a single scale, instead redesigning the network architecture to develop a patch-wise receptive field in order to learn internal patch statistics within a single image. SinFusion (Nikankin et al., 2022) also utilized network redesigns to adjust the receptive field of the model and trained on large random crops of the input image to generate new samples, while extending its applicability to video generation, extrapolation and upsampling.

Generative models trained on limited data have also been explored in the medical imaging domain. Tub-GAN (Zhao et al., 2018) used style transfer to generate diverse retinal and neuronal images from a small number of training examples. MinimalGAN (Zhang et al., 2023) is another style-based GAN that modeled content and style separately to produce diverse outputs from limited data, and can be trained with a single image or multiple images. Thambawita et al. (2022) proposed SinGAN-Seg, a modified version of the SinGAN architecture along with a style-transfer based fine-tuning step, in order to generate synthetic medical images and their corresponding segmentation masks by training on a single image, and evaluated their approach for a polyp segmentation task. However, there has been little exploration of the effectiveness of single-image diffusion model-based approaches in the medical imaging domain, particularly for data augmentation purposes in clinical diagnostic tasks.

3. Methods

SinDDM, a hierarchical denoising diffusion model (DDM), can be trained on a single image and can generate novel high-quality variants of the training image. In this study, we investigate the feasibility of leveraging SinDDM for data augmentation in the context of a real-world medical imaging application. Specifically, in addition to assessing the perceptual image quality of the generated samples, we evaluate the feasibility of leveraging trained SinDDMs for data augmentation to improve the performance of a clinical task. Our study also investigates the use of SinDDMs trained with a limited number of images instead of one single image.

3.1. SinDDM

SinDDM is an unconditional denoising diffusion model that employs a multi-scale diffusion process to learn the internal statistics of structures within a single training image (Kulikov et al., 2023). The forward multi-scale diffusion process combines both blur and noise over training image x at scale s and timestep t, $(s,t) \in \{0,...,N-1\} \times \{0,...,T\}$, as

$$x_t^s = \sqrt{\bar{\alpha}_t}(\gamma_t^s \tilde{x}^s + (1-\gamma_t^s)x^s) + \sqrt{1-\bar{\alpha}_t}\epsilon, \qquad (1)$$

where $\epsilon \in \mathcal{N}(0, \boldsymbol{I})$. $\gamma_t^s \in [0,1]$ is a non-increasing monotonic function of t and $\bar{\alpha}_t$ follows a cosine schedule (Nichol et al., 2021). In the blurry version of the training image in different scales $(\tilde{x}^{N-1},...,\tilde{x}^0)$, $\tilde{x}^0 = x^0$ and $\tilde{x}^s = (x^{s-1}) \uparrow^r$ for every $s \geq 1$, where $\uparrow$ represents upsampling operations using bicubic interpolation. As t increases, x_t^s becomes both noisier and blurrier.

In order to sample an image, SinDDM first follows the standard DDM approach at image scale $s = 0$ by starting with random noise at timestep $t = T$ and gradually removing noise until a clean sample is generated at $t = 0$. An upsampling operation is

then applied to obtain image sample at scale $s = 1$ and noise is added. A reverse diffusion process is computed to form a sample at this scale. This process is repeated until the finest scale $s = N - 1$ is reached.

To drive the reverse diffusion process, a single fully convolutional denoiser comprising 4 convolutional blocks is trained on various scales of the image to predict x_0^s based on x_t^s. A small receptive field of 35×35 is achieved to capture the statistics of the fine details within each scale. In addition, both the scale s and the timestep t are injected into the model by the use of joint embedding, which has been shown to improve generation quality and training time.

3.2. SinGAN

In this study, we also compare the image generation performance of SinDDM with that of the existing GAN-based single-image generative model SinGAN (Shaham et al., 2019). SinGAN is an unconditional generative model trained to learn the internal distribution of patches within the image. The model consists of a pyramid of patch-GANs, where both training and inference are done in a coarse-to-fine manner. At each image scale, the generator learns to generate image samples in which all the overlapping patches cannot be distinguished by the discriminator from the patches in the down-sampled training image. During the inference, the generation of an image starts at the coarsest scale and sequentially passes through each of the generator up to the finest scale, with noise added at each image scale.

3.3. POCOVID-Net for lung ultrasound classification

For task-based image quality evaluation of the synthetic LUS images generated by SinDDM, we adopted POCOVID-Net (Born et al., 2021), a deep learning-based model for the downstream task of classifying LUS images into COVID-19, bacterial pneuomonia and healthy patients. POCOVID-Net comprises the backbone neural architecture of VGG-16 pre-trained on ImageNet (Deng et al., 2009). The backbone network was followed by one dense layer of 64 neurons with ReLU activation, dropout of 0.5, batch normalization and an output layer having 3 nodes with a softmax activation function.

4. Experiments

4.1. Dataset

For this work, we used the LUS data made publicly available by Born et al. (2021) after removing all images and videos with non-commercial licenses. Figure 1 illustrates a representative example from each of the COVID-19, bacterial pneumonia and healthy classes. The selected dataset includes 106 videos and 32 images recorded with convex probe on patients with COVID-19, bacterial pneumonia and healthy controls respectively, as shown in Table 1. Since the data were originally collected from various sources, the videos have different lengths and frame rates. Here, the videos and images were pre-processed as described in previous work (Born et al., 2021). The videos were split into image frames at a frame rate of 3 Hz and the images were cropped to a standard aspect ratio. Our final dataset consists of 1994 images in total, with 823, 267 and 904 images from COVID-19, bacterial pneumonia and healthy classes respectively.

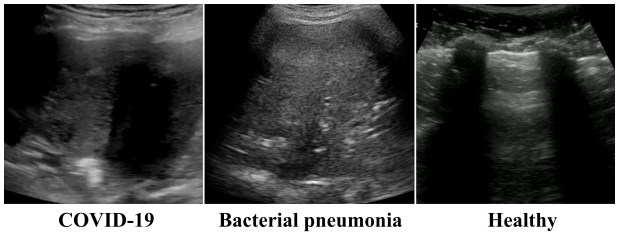

Figure 1: LUS image examples of COVID-19, bacterial pneumonia and healthy controls.

Similar to previous studies (Diaz-Escobar et al., 2021; Barros et al., 2021; Born et al., 2021), 5-fold cross-validation was used to verify the robustness of the classification results. The dataset was randomly partitioned into five folds at the patient-level. Four out of the five folds were merged to form the training set while the remaining fold was used as the validation set, making each fold the validation set in turn.

4.2. Experimental settings

4.2.1. SinDDM training with limited pneumonia data

To employ SinDDM to generate synthetic high-quality LUS images, we first train SinDDM models

Table 1: Number of videos and images in the selected dataset per class, and the total number of corresponding extracted image frames.

Class	Videos	Images	Frames
COVID-19	56	19	823
Pneumonia	14	8	267
Healthy	36	5	904
Total	106	32	1994

for each of the 8 single LUS images of pneumonia patients. Note that in this study, we only augment data based on LUS images rather than on image frames extracted from videos. The SinDDM models were trained using the Adam optimizer with its default Torch parameters. Using V100 GPUs, we train our model for 120×10^3 steps with an initial learning rate of 0.001, which is reduced by half on [20, 40, 70, 80, 90, 110] $\times 10^3$ steps. The batch size was set to be 32. Apart from the vanilla SinDDM which was trained on a single LUS image, we also investigated the training of SinDDM on a few (2-3) LUS images. For simplicity, this modified version of SinDDM will be referred as *FewDDM* in the rest of the paper.

4.2.2. SinGAN training

For the training of SinGAN, we employed the default configuration in the original paper (Shaham et al., 2019). The SinGAN model was trained sequentially from the coarsest to the finest scale. Once each GAN is trained, it is fixed. The training loss for each individual GAN is comprised of an adversarial loss and a reconstruction term. Models in each scale were trained for 2000 epochs.

4.2.3. Task-based evaluation using POCOVID-Net

Starting with the pre-trained POCOVID-Net (Born et al., 2021), we fine-tuned only the weights of the last three layers, resulting in a total of ∼2.4 M training parameters and ∼12.4 M non-trainable parameters. The model was trained to minimize cross-entropy loss by using Adam optimizer (Kingma and Ba, 2014). Each model was trained for 40 epochs with a batch size of 8 and early stopping strategy was enabled. In the default training scheme (Born et al., 2021), data augmentation transformations such as rotations (up to 10°), translations (up to 10%), and horizontal and vertical flips were used.

5. Results

5.1. Quantitative evaluation of synthetic images from SinDDM

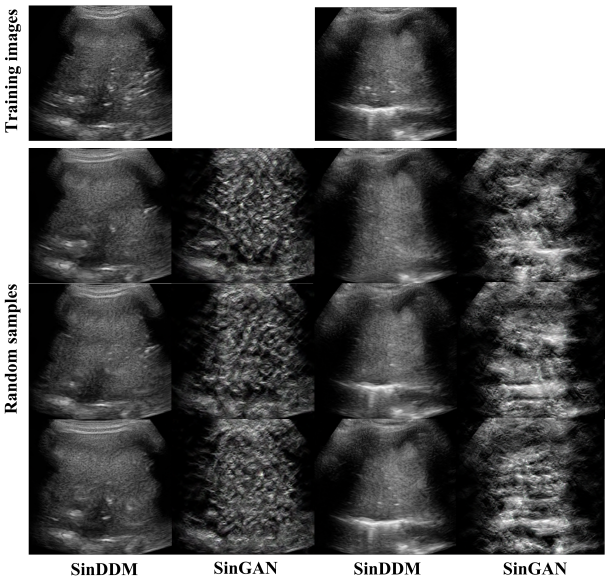

Figure 2: Novel variants of two different LUS pneumonia images generated by SinDDM and SinGAN.

Examples of novel syntehtic LUS images of pneumonia generated by SinDDM and SinGAN are shown in Figure 2. Close inspection reveals that images generated by SinGAN are dominated by artifacts, while SinDDM successfully captures the pathological features related to the diagnosis of pneumonia such as consolidation and air bronchograms. The global structure of the original LUS images is well-preserved and reasonable variations in the local structures can be observed as well.

Frechet Inception Distance (Heusel et al., 2017), a popular metric used to evaluate synthetic images produced by generative models, measures the difference between the distributions of activation vectors after the last pooling layer in the ImageNet-trained Inception v3 network for the generated and real images. For evaluating the images generated by SinDDM and

for comparing SinDDM against SinGAN, we use the Single Image Frechet Inception Distance (SIFID) proposed by Shaham et al. (2019), which uses the internal distribution of deep features generated by a real and fake image pair at the convolutional layer before the second pooling layer. Table 2 reports the average SIFID for images generated by SinGAN and SinDDM from the same set of real images. It can be seen that SinDDM achieves a considerably lower SIFID score on lung ultrasound images. We further computed the average structural similarity (SSIM) between pairs of real and synthetic images to provide a comparison of luminance, contrast and structure in the images generated by different models relative to the real images. As seen from Table 2, SinDDM variants have a much higher structural similarity to the real images in general. We also include results from a qualitative evaluation of the synthetic images by an expert sonographer in the Discussion section.

Table 2: Average SIFID and SSIM for images generated by SinGAN and SinDDM

Model	SIFID	SSIM
SinGAN	1.16	0.23
SinDDM	**0.38**	**0.43**

5.2. Task-based evaluation of synthetic images from SinDDM

In order to investigate the effectiveness of SinDDM-based data augmentation, we consider a three-class classification task of detecting lung ultrasound pathologies with artificially introduced class imbalance in the training data. Specifically, we compare the classification performance of the POCOVID-Nets that are trained on datasets consisting of only 25% of the pneumonia data available in each training fold. Before augmenting the pneumonia data using SinDDM, we first investigate the effectiveness of regular data augmentation techniques (rotation, translation etc.) which were used in a previous study (Born et al., 2021). It can be observed from Table 3 that while mild improvement was shown in both the accuracy and balanced accuracy with regular data augmentation, there is no improvement in term of the evaluation metrics reported for the minority class of pneumonia.

A total of 8 SinDDMs and SinGANs were trained corresponding to the 8 single pneumonia LUS images referred to in Table 1, respectively. After SinDDMs were trained on individual pneumonia images, 32 novel image samples were generated from each of the models respectively. These synthetic pneumonia images were added to the corresponding training folds. As reported in Table 3, by synthetically increasing the total number of pneumonia images, the accuracy and balanced accuracy metrics greatly improved in comparison to the ones without any data augmentation. It is worth noting that metrics such as precision, recall and F1-score of the pneumonia class improved significantly as well. These results demonstrate the effectiveness of data augmentation by using SinDDM in the LUS pathology classification task. On the other hand, while SinGAN does not outperform SinDDM in term of visual quality as shown in Table 2, increasing the number of available pneumonia images for training by adding the synthetic images from SinGAN still provides significant improvements in the classification performance.

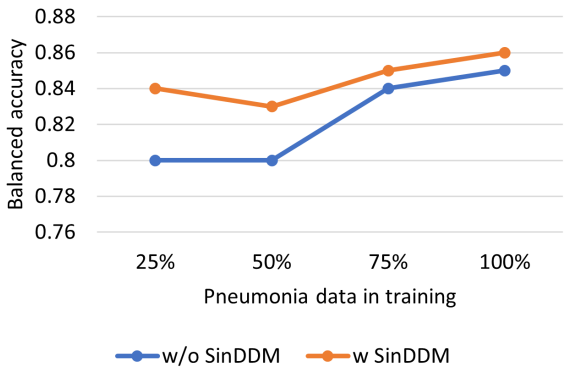

Figure 3: Balanced accuracies without and with SinDDM data augmentation for varying percentages of pneumonia data in training.

In addition, we compare the balanced accuracy values without and with SinDDM data augmentation when different percentages of pneumonia data (split at patient-level) are used in the training. As before, for each training fold, each trained SinDDM generated 32 synthetic images to augment the dataset. POCOVID-Nets were retrained and evaluated for each dataset variation. Here, the task performance of four comparisons using 25%, 50%, 75% and 100% of the available training data are presented in Fig. 3. The results confirm that the task performance

can be improved with SinDDM data augmentation on the minority class. It can be seen that the improvement of balanced accuracy is most significant when the number of real pneumonia images is severely limited, and there is some improvement even when more patient data becomes available for training.

5.3. Task-based evaluation of synthetic images from FewDDM

In addition to using one single image for training SinDDM, it is also of our interest to investigate the image generation capability of SinDDM when trained on a limited number of samples. In each training iteration, we sample a batch of noisy variants of a randomly chosen input image among a few training images instead of sampling noisy variants of a single image. Therefore, for 3 out of the 5 training folds where more than one single pneumonia images exist, three FewDDMs were trained on all such pneumonia images within the training fold. Figure 4 shows novel pneumonia LUS image samples generated by FewDDM trained on all 3 of the images in the top row. It can be seen that the predictive characteristics of pneumonia in LUS such as consolidation and fluid bronchogram are learned by the FewDDM and preserved well in the generated image variants.

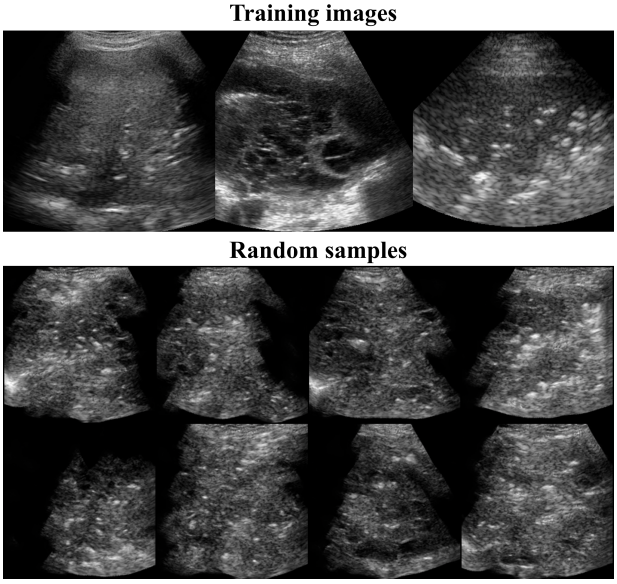

Figure 4: Novel variants of the pneumonia LUS images generated by FewDDM.

After the FewDDMs were trained, the same number of synthetic pneumonia images were added to each training fold for consistent comparison of the classification task with Section 5.2. As reported in Table 3, both the overall accuracy and balanced accuracy metrics improve compared to those of either SinDDM or SinGAN. Specifically, the F1-score of the pneumonia class is considerably increased when regular data augmentation is added to the training data along with synthetic images from FewDDM. This could be interpreted as more data variety captured by FewDDM when trained on more than one image, enabling the POCOVID-Net to be more robust in differentiating between different pathologies.

6. Discussion

Class imbalance is a commonly observed issue for rare pathologies or abnormalities in medical imaging data. Additionally, the small scale of dataset acquired in medical imaging due to the complexity, high cost and privacy concerns of patient data collection poses significant challenges when researchers design and deploy deep learning-based models for downstream clinical tasks such as disease detection and lesion segmentation (Willemink et al., 2020). In the cirsumstances where both extremely severe class imbalance and limited number of samples exist, we showed that traditional data augmentation techniques such as translation and rotation may not be effective to improve the downstream task performance.

In this study, we investigated the applicability of the recently proposed single-image denoising diffusion model, SinDDM, to synthesize lung ultrasound images for data augmentation. Built on a multi-scale diffusion process, each SinDDM model was trained on a single pneumonia lung ultrasound image to generate diverse novel samples. When compared with SinGAN, SinDDM demonstrated high sample quality in terms of producing sufficient varieties in the pulmonary details of LUS images while preserving pathological features related to the diagnosis of pneumonia as well as the global structure of the original image.

Qualitatively, 32 randomly sampled synthetic images produced by each of SinGAN, SinDDM and FewDDM were assessed by a trained sonographer tasked with labeling them as real or synthetic. The qualitative evaluation results are given in Table 4, where a lower fake detection rate indicates that the model can generate more realistic-looking synthetic

Table 3: Comparison of the tested classification models on 5-fold cross validation for each class. In the training dataset, only 25% of the pneumonia data is used. Precision abbreviated to Prec., recall to Rec., accuracy to Acc., balanced accuracy to Bal. Acc. Results are compared among models trained with or without regular data augmentation, and with or without data augmented by using SinGAN, SinDDM and FewDDM.

	Class	**Prec.**	**Rec.**	**F1-score**	**Acc.**	**Bal. Acc.**
w/o data aug.	COVID-19	0.76 ± 0.19	0.91 ± 0.14	0.80 ± 0.13	0.82 ± 0.14	0.77 ± 0.20
	Pneumonia	0.79 ± 0.40	0.67 ± 0.35	0.72 ± 0.37		
	Healthy	0.94 ± 0.08	0.73 ± 0.33	0.76 ± 0.25		
w/ regular aug.	COVID-19	0.81 ± 0.14	0.93 ± 0.07	0.86 ± 0.09	0.86 ± 0.13	0.80 ± 0.19
	Pneumonia	0.73 ± 0.37	0.66 ± 0.34	0.69 ± 0.35		
	Healthy	0.96 ± 0.04	0.79 ± 0.28	0.83 ± 0.21		
w/ SinGAN aug.	COVID-19	0.76 ± 0.19	0.91 ± 0.14	0.81 ± 0.12	0.82 ± 0.15	0.78 ± 0.82
	Pneumonia	0.98 ± 0.03	0.69 ± 0.32	0.75 ± 0.32		
	Healthy	0.94 ± 0.07	0.72 ± 0.34	0.75 ± 0.28		
w/ SinGAN aug. and regular aug.	COVID-19	0.90 ± 0.10	0.83 ± 0.16	0.85 ± 0.10	0.86 ± 0.12	0.86 ± 0.09
	Pneumonia	0.84 ± 0.29	0.86 ± 0.15	0.79 ± 0.21		
	Healthy	0.90 ± 0.07	0.90 ± 0.13	0.90 ± 0.09		
w/ SinDDM aug.	COVID-19	0.81 ± 0.16	0.86 ± 0.12	0.82 ± 0.11	0.85 ± 0.09	0.80 ± 0.14
	Pneumonia	0.98 ± 0.03	0.67 ± 0.34	0.72 ± 0.34		
	Healthy	0.87 ± 0.07	0.87 ± 0.18	0.86 ± 0.11		
w/ SinDDM aug. and regular aug.	COVID-19	0.87 ± 0.08	0.87 ± 0.11	0.87 ± 0.08	0.88 ± 0.08	0.84 ± 0.13
	Pneumonia	0.99 ± 0.02	0.68 ± 0.32	0.75 ± 0.29		
	Healthy	0.88 ± 0.12	0.96 ± 0.06	0.91 ± 0.06		
w/ FewDDM aug.	COVID-19	0.83 ± 0.16	0.85 ± 0.12	0.83 ± 0.12	0.87 ± 0.09	0.85 ± 0.08
	Pneumonia	0.95 ± 0.06	0.82 ± 0.11	0.87 ± 0.08		
	Healthy	0.89 ± 0.09	0.89 ± 0.18	0.87 ± 0.11		
w/ FewDDM aug. and regular aug.	COVID-19	0.90 ± 0.08	0.88 ± 0.07	0.89 ± 0.05	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.90 \pm 0.05}$
	Pneumonia	0.98 ± 0.03	0.85 ± 0.12	0.91 ± 0.07		
	Healthy	0.91 ± 0.08	0.96 ± 0.06	0.93 ± 0.03		

samples. None of the synthetic images produced by SinGAN were able to evade detection, mostly due to over-saturated white regions and non-uniform gains according to the sonographer's notes. SinDDM had the lowest fake detection rate out of the 3 models, which could be attributed to realistic-looking rib shadows and the random nature of tissue patterns as noted by the sonographer. FewDDM fared somewhat worse in comparison to SinDDM due to the absence of pleural lines or chest walls in some images, but were noted to have a random nature of tissue pattern as seen in real images.

Quantitatively as well, it was shown through metrics such as SIFID and SSIM that DDM-based approaches generate more realistic images compared to the GAN-based approaches. However, it remains an open question in the medical imaging community to appropriately select and define a universal image quality metric to evaluate the performance of generative models such as GANs and diffusion models. One reason is that quality evaluation of synthetic images is subject to imaging modalities, the objects to be imaged, and the downstream clinical tasks. In practice, the medical image features looked at by radiologists might not be the same as what the downstream

Table 4: Qualitative evaluation results for SinGAN, SinDDM and FewDDM

Model	Fake detection rate
SinGAN	100%
SinDDM	**9.37%**
FewDDM	34.37%

network focuses on. Moreover, different radiologists might focus on different pathological features in the medical image as well. More recent works also highlight the need for assessing the ability of generative models on learning medical image statistics such as texture and morphology features (Kelkar et al., 2023; Deshpande et al., 2023).

One important purpose of synthetic data augmentation in medical imaging is to improve the performance of downstream diagnostic tasks. Here, augmenting the LUS image dataset with synthetic pneumonia images generated by SinDDM achieved significant improvement on the pathology classification performance, particularly on the minority pneumonia class. While SinGAN failed to produce LUS images with high visual quality compared to SinDDM, the task performance still benefits from the synthetic data augmentation. This is worth noting since traditional image quality metrics may not be necessarily correlated with the objective task-based image quality measures (Badal et al., 2019). Therefore, evaluating image quality in medical imaging in both aspects is of great importance for developing deep learning-based medical imaging applications (Zhang et al., 2021; Kelkar et al., 2021; Li et al., 2021).

A practical scenario when working on a small-scale medical image dataset is that there exist only a few images of a particular minority class. Instead of training different SinDDMs for each individual image which could be tedious and computationally burdensome, we investigated the effectiveness of data augmentation with a modified FewDDM that can be trained with a limited number of samples. While the visual quality of global structures of the generated images is not as satisfying as those from SinDDM, the downstream task performance with FewDDM outperformed both SinDDM and SinGAN. By learning varieties in local structures from multiple samples instead of just one, details in diverse images generated with FewDDM could potentially allow for a more robust and generalized downstream model. It is also worth noting that for all three models considered, a combination of regular and synthetic data augmentation techniques yielded the best performance for the downstream task.

In SinDDM, the receptive field of the model is designed to be small to avoid memorization of the single training image (Kulikov et al., 2023). Receptive fields play an important role based on the task at hand. While larger receptive fields can capture global features, smaller receptive fields help learn local details. The relationship between the design of receptive fields in SinDDM and image quality of the synthetic images remains to be investigated in a future study.

References

W Abdalla, M Elgendy, AA Abdelaziz, and MA Ammar. Lung ultrasound versus chest radiography for the diagnosis of pneumothorax in critically ill patients: A prospective, single-blind study. *Saudi journal of anaesthesia*, 10(3):265, 2016.

Mohamed Akrout, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincső, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. *arXiv preprint arXiv:2301.04802*, 2023.

S Anandasabapathy, X Zhang, M Anastasio, R Richards-Kortum, and E Petrova. An optical, endoscopic brush for high-yield diagnostics in esophageal cancer. In *Endoscopic Microscopy XVI*, volume 11620, page 116200B. SPIE, 2021.

Andreu Badal, Kenny H Cha, Sarah E Divel, Christian G Graff, Rongping Zeng, and Aldo Badano. Virtual clinical trial for task-based evaluation of a deep learning synthetic mammography algorithm. In *Medical Imaging 2019: Physics of Medical Imaging*, volume 10948, pages 164–173. SPIE, 2019.

Bruno Barros, Paulo Lacerda, Celio Albuquerque, and Aura Conci. Pulmonary covid-19: learning spatiotemporal features combining cnn and lstm networks for lung ultrasound video classification. *Sensors*, 21(16):5486, 2021.

Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. Learning texture manifolds with the periodic spatial gan. *arXiv preprint arXiv:1705.06566*, 2017.

Jannis Born, Nina Wiedemann, Manuel Cossio, Charlotte Buhre, Gabriel Brändle, Konstantin Leidermann, Julie Goulet, Avinash Aujayeb, Michael Moor, Bastian Rieck, et al. Accelerating detection of lung pathologies with explainable ultrasound image analysis. *Applied Sciences*, 11(2):672, 2021.

Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting

pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.

Miguel A Chavez, Navid Shams, Laura E Ellington, Neha Naithani, Robert H Gilman, Mark C Steinhoff, Mathuram Santosham, Robert E Black, Carrie Price, Margaret Gross, et al. Lung ultrasound for the diagnosis of pneumonia in adults: a systematic review and meta-analysis. *Respiratory research*, 15(1):1–9, 2014.

Pietro Antonio Cicalese, Aryan Mobiny, Pengyu Yuan, Jan Becker, Chandra Mohan, and Hien Van Nguyen. Stypath: Style-transfer data augmentation for robust histology image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 351–361. Springer, 2020.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Rucha Deshpande, Muzaffer Özbey, Hua Li, Mark A Anastasio, and Frank J Brooks. Assessing the capacity of a denoising diffusion probabilistic model to reproduce spatial context. *arXiv preprint arXiv:2309.10817*, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Dimitrios E Diamantis, Panagiota Gatoula, and Dimitris K Iakovidis. Endovae: Generating endoscopic images with a variational autoencoder. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022.

Julia Diaz-Escobar, Nelson E Ordonez-Guillen, Salvador Villarreal-Reyes, Alejandro Galaviz-Mosqueda, Vitaly Kober, Raúl Rivera-Rodriguez, and Jose E Lozano Rizk. Deep-learning based detection of covid-19 using lung ultrasound imagery. *Plos one*, 16(8):e0255886, 2021.

Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022.

Kaushik Dutta, Ziping Liu, Richard Laforest, Abhinav Jha, and Kooresh Isaac Shoghi. Deep learning framework to synthesize high-count preclinical pet images from low-count preclinical pet images. In *Medical Imaging 2022: Physics of Medical Imaging*, volume 12031, pages 351–360. SPIE, 2022.

Zong Fan, Xiaohui Zhang, Jacob A Gasienica, Jennifer Potts, Su Ruan, Wade Thorstad, Hiram Gay, Pengfei Song, Xiaowei Wang, and Hua Li. A novel adversarial learning strategy for medical image classification. *arXiv preprint arXiv:2206.11501*, 2022.

Zong Fan, Ping Gong, Shanshan Tang, Christine U Lee, Xiaohui Zhang, Zhimin Wang, Pengfei Song, Shigao Chen, and Hua Li. An auxiliary attention-based network for joint classification and localization of breast tumor on ultrasound images. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 34–40. SPIE, 2023.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, 9(4):81, 2023.

Varun A Kelkar, Xiaohui Zhang, Jason Granstedt, Hua Li, and Mark A Anastasio. Task-based evaluation of deep image super-resolution in medical imaging. In *Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*, volume 11599, pages 207–213. SPIE, 2021.

Varun A Kelkar, Dimitrios S Gotsis, Frank J Brooks, KC Prabhat, Kyle J Myers, Rongping Zeng, and Mark A Anastasio. Assessing the ability of generative adversarial networks to learn canonical medical image statistics. *IEEE transactions on medical imaging*, 2023.

Boah Kim and Jong Chul Ye. Diffusion deformable model for 4d temporal medical image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 539–548. Springer, 2022.

Mingyu Kim, Jihye Yun, Yongwon Cho, Keewon Shin, Ryoungwoo Jang, Hyun-jin Bae, and Namkug Kim. Deep learning in medical imaging. *Neurospine*, 16(4):657, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. Sinddm: A single image denoising diffusion model. In *International Conference on Machine Learning*, pages 17920–17930. PMLR, 2023.

Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 702–716. Springer, 2016.

Kaiyan Li, Weimin Zhou, Hua Li, and Mark A Anastasio. Assessing the impact of deep neural network-based image denoising on binary signal detection tasks. *IEEE transactions on medical imaging*, 40 (9):2295–2305, 2021.

Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *arXiv preprint arXiv:2304.02742*, 2023.

Qing Lyu and Ge Wang. Conversion between ct and mri images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022.

Baoqiang Ma, Yan Zhao, Yujing Yang, Xiaohui Zhang, Xiaoxi Dong, Debin Zeng, Siyu Ma, and Shuyu Li. Mri image synthesis with dual discriminator adversarial learning and difficulty-aware attention mechanism for hippocampal subfields segmentation. *Computerized Medical Imaging and Graphics*, 86:101800, 2020.

Baoqiang Ma, Jiapan Guo, Tian-Tian Zhai, Arjen van der Schaaf, Roel JHM Steenbakkers, Lisanne V van Dijk, Stefan Both, Johannes A Langendijk, Weichuan Zhang, Bingjiang Qiu, et al. Ct-based deep multi-label learning prediction model for outcome in patients with oropharyngeal squamous cell carcinoma. *Medical Physics*, 2023.

Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2023.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Yaniv Nikankin, Niv Haim, and Michal Irani. Sinfusion: Training diffusion models on a single image or video. *arXiv preprint arXiv:2211.11743*, 2022.

Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. *arXiv preprint arXiv:2211.01323*, 2022.

Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

Eman Rezk, Mohamed Eltorki, Wael El-Dakhakhni, et al. Improving skin color diversity in cancer detection: deep learning approach. *JMIR Dermatology*, 5(3):e39143, 2022.

Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):16884, 2019.

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4570–4580, 2019.

Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 1–11. Springer, 2018a.

Younghak Shin, Hemin Ali Qadir, and Ilangko Balasingham. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access*, 6: 56007–56017, 2018b.

Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and retargeting the "dna" of a natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4492–4501, 2019.

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

MJ Smith, SA Hayward, SM Innes, and ASC Miller. Point-of-care lung ultrasound in patients with covid-19–a narrative review. *Anaesthesia*, 75(8): 1096–1104, 2020.

S Suganyadevi, V Seethalakshmi, and K Balasamy. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1):19–38, 2022.

Laith R Sultan and Chandra M Sehgal. A review of early experience in lung ultrasound in the diagnosis and management of covid-19. *Ultrasound in Medicine & Biology*, 46(9):2530–2545, 2020.

You-Bao Tang, Sooyoun Oh, Yu-Xing Tang, Jing Xiao, and Ronald M Summers. Ct-realistic data augmentation using generative adversarial network for robust lymph node segmentation. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, pages 976–981. SPIE, 2019.

Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A Hicks, Hugo L Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A Riegler. Singan-seg: Synthetic training data generation for medical image segmentation. *PloS one*, 17(5):e0267976, 2022.

Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Sindiffusion: Learning a diffusion model from a single natural image. *arXiv preprint arXiv:2211.12445*, 2022.

Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295 (1):4–15, 2020.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.

Xiaohui Zhang, Varun A Kelkar, Jason Granstedt, Hua Li, and Mark A Anastasio. Impact of deep learning-based image super-resolution on binary signal detection. *Journal of Medical Imaging*, 8(6): 065501–065501, 2021.

Yipeng Zhang, Quan Wang, and Bingliang Hu. Minimalgan: diverse medical image synthesis for data augmentation using minimal training data. *Applied Intelligence*, 53(4):3899–3916, 2023.

Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 8543–8553, 2019.

He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical image analysis*, 49:14–26, 2018.