ASME Journal of Medical Devices
Online journal at:
https://asmedigitalcollection.asme.org/medicaldevices

Toward Smart Ultrasound Image Augmentation to Advance Tumor Treatment Monitoring: Exploring the Potential of Diffusion Generative Model

Emmanuel Yangue
School of Industrial Engineering & Management,
Oklahoma State University,
Stillwater, OK 74078

Ashish Ranjan
Department of Radiation Oncology,
UT Southwestern Medical Center,
Dallas, TX 75390

Yu Feng
School of Chemical Engineering,
Oklahoma State University,
Stillwater, OK 74078

Chenang Liu[1]
School of Industrial Engineering & Management,
Oklahoma State University,
Stillwater, OK 74078
e-mail: chenang.Liu@okstate.edu

Medical imaging is a crucial tool in clinics to monitor tumor treatment progress. In practice, many imaging tools (such as magnetic resonance imaging (MRI) and computed tomography (CT) scans) are in general costly and may also expose patients to radiation, leading to potential side effects. Recent studies have demonstrated that ultrasound imaging, which is safe, low-cost, and easy to access, can monitor the drug delivery progress in solid tumors. However, the noisy nature of ultrasound images and the high-level uncertainty of cancer disease progression are still challenging in ultrasound-based tumor treatment monitoring. To overcome these barriers, this work presents a comparative study to explore the potential advantages of the emerging diffusion generative models against the commonly applied state-of-the-art generative models. Namely, the denoising diffusion models (DDMs), against the generative adversarial networks (GAN), and variational auto-encoders (VAE), are used for analyzing the ultrasound images through image augmentation. These models are evaluated based on their capacity to augment ultrasound images for exploring the potential variations of tumor treatment monitoring. The results across different cases indicate that the denoising diffusion implicit models (DDIM)/kernel inception distance (KID)-inception score (IS) model leveraged in this work outperforms the other models in the study in terms of similarity, diversity, and predictive accuracy. Therefore, further investigation of such diffusion generative models could be considered as they can potentially serve as a great predictive tool for ultrasound image-enabled tumor treatment monitoring in the future.
[DOI: 10.1115/1.4065905]

Keywords: diffusion generative models, image augmentation, tumor treatment monitoring, ultrasound imaging

1 Introduction

1.1 Background and Motivation. Medical imaging techniques play a crucial role in monitoring and validating study hypotheses of tumor treatment progress [1], as many of them have become more capable and precise. The commonly applied techniques in clinics include X-rays, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), optical coherence tomography (OCT), and ultrasound imaging [1]. With the images available, gaining a deeper understanding of how tumors react to pharmacological interventions stands out as one of the key research areas in image-guided cancer research, e.g., image-guided drug delivery (IGDD) [2,3]. Furthermore, although some of the medical imaging techniques (e.g., X-rays, CT, MRI) used for IGDD could be very accurate, their high cost and potential exposure to radiation lead to significant concerns in broadening their applications in tumor treatment monitoring, e.g., tracking the progress of drug delivery. These concerns also motivate researchers and practitioners to consider ultrasound imaging, as it is safe to use, easy to access, and relatively inexpensive to use compared to other imaging techniques [1].

Image-guided drug delivery is utilized to visualize and represent tumor response (tissue changes) to cancer treatment, and a major challenge remains the development of accurate methods to advance drug delivery monitoring and control [4]. In IGDD using ultrasound imaging, the presence of speckles (noise) affects the resolution quality of ultrasound images, leading to challenges in interpreting the images. Therefore, several prior studies have focused on enhancing ultrasound images using various techniques. For example, some traditional techniques employed in the field include histogram equalization, spatial and frequency domain techniques, and filtering techniques, among others [5–7]. Nevertheless, several of the traditional image-enhancing techniques may not be robust enough to address the complexities of ultrasound images in IGDD,

[1]Corresponding author.
Manuscript received February 23, 2024; final manuscript received July 1, 2024; published online August 6, 2024. Assoc. Editor: Zhenpeng Qin.

particularly when considering tumor progression over time. Therefore, several studies are instead using machine learning (ML) algorithms to address these complexities [8,9].

1.2 Deep Generative Models and Ultrasound Images. In tumor treatment monitoring studies, the tumor has the potential to grow or diminish, following different trajectories, which require thorough investigation. As the collected real data can only reflect the progress that has already occurred, it is also highly valuable to enable the generation of potential variations, which would be beneficial for better consideration of therapy design. Predictive methods for monitoring variations can assist researchers in developing new treatment designs for tumors. This approach can help save resources and enhance therapy planning in advance compared to real-time monitoring, which may require more corrective actions and resources. According to the literature, ultrasound images have been extensively utilized for treatment effectiveness and diagnosis [3,4], and therefore, they could potentially be used for studies that explore occurring variations in tumor treatment monitoring.

Therefore, there is a need to develop models that can comprehend the underlying distribution of the ultrasound images used for tumor treatment monitoring. Recent advances in generative models have demonstrated their great potential to understand these distributions. Deep generative models (DGMs) are neural network models capable of understanding the underlying distribution of a dataset. DGMs are mainly used for data augmentation though they could also be leveraged for other analytics tasks. In practice, they can usually provide more diversity and generalization to the augmented datasets, compared to the traditional transformation techniques (e.g., cropping, stretching, flipping) [10] and oversampling techniques (e.g., SMOTE) [11]. As a result, they can potentially be utilized for ultrasound image augmentation to advance tumor treatment monitoring [12]. Examples of these DGMs include generative adversarial networks (GAN) and diffusion models both of which have been used in numerous applications for medical imaging analysis [13,14]. For example, Qin et al. [15] used GANs to generate synthetic chest X-rays for limited and imbalanced datasets. Nevertheless, it is worth noting that employing DGMs in augmenting ultrasound images also comes with some critical challenges.

1.3 Challenges in Applying Deep Generative Models to Ultrasound Images. The first challenge concerns the complex textures and features of ultrasound images [4]. Ultrasound images depict weak tissue contrast, which makes it impractical for online drug discharge and chemotherapy distribution [4]. Moreover, ultrasound images often exhibit noisy interference (speckles) with unclear patterns making feature extraction even more challenging [4]. For instance, according to the literature, there is a significant technical barrier to portraying drug particles with short-time circulation in ultrasound images [16]. Such high data complexities also bring challenges for the application of DGMs in understanding the underlying distribution of ultrasound images, for example, leading to issues such as mode collapse, vanishing gradients, poor convergence, data copying, etc. [17,18]. Another challenge in applying DGMs to ultrasound imaging in IGDD is their capacity to predict realistic variations (tissue changes) while reducing the level of unrealistic tumor responses (model hallucination). More specifically, in medical studies, due to the required ethics in healthcare, DGMs must include an appropriate level of fairness in representing sensitive attributes [19].

1.4 Objectives. To address these limitations, this study evaluates the capability and potential of a newly improved diffusion generative model compared to other DGMs for predicting high-quality and realistic tumor responses (variations) through image augmentation. Specifically, the study assesses several popular denoising diffusion models (DDMs), generative adversarial networks (GANs), as well as the widely applied variational auto-encoders (VAEs) in this comparative analysis. One diffusion model introduced in this study, denoising diffusion implicit models (DDIM)/kernel inception distance (KID)-inception score (IS) [20], has been modified to better address the challenges of ultrasound images mentioned earlier (more details in Sec. 2.1.2). The timestep capability of diffusion models to reconstruct samples provides strong motivation for its utilization over other DGMs, as cancer treatment monitoring requires an understanding of the spatial and temporal aspects of tumor progression over time. Additionally, to support this study, a set of ultrasound images is collected from experiments on mice with colon cancer, encompassing both before and after treatment phases.

The main contribution of this study is to establish the predictive potential of DGM, particularly, our recently developed diffusion model, termed DDIM/KID-IS, which is leveraged and tailored for this work, to monitor tumor treatment responses via image augmentation. This represents the preliminary stage in enabling advanced tumor monitoring research using DGM. The generated samples demonstrate the effectiveness of the models through the comparison and validation of three aspects:

(1) *Image fidelity*: The deep generative models' ability to generate ultrasound images resembling real ultrasound images, indicating an understanding of the underlying distribution of the ultrasound images crucial for tumor treatment monitoring. Five evaluation metrics are employed to validate this step, namely kernel inception distance (KID), learned perceptual image patch similarity (LPIPS), structural similarity index measure (SSIM), muti-scale SSIM (MS-SSIM), and peak signal-to-noise ratio (PSNR).
(2) *Diversity generation*: The ability of the deep generative models to produce samples with a notable level of diversity which would suggest that the generative models can emulate potential tumor response variations (i.e., growth or reduction of tumors with different trajectories). Entropy (H) and kernel density estimation (KDE) are used to estimate the level of diversity present in the generated ultrasound images.
(3) *Predictive capability*: The predictive capability of the DGMs is evaluated against a test set using K-means clustering, the root mean square error (RMSE), and the mean absolute error (MAE).

This study substantially presents the potential of DDIM/KID-IS and its capability to take the lead in terms of ultrasound image augmentation for tumor treatment monitoring. The remainder of this paper is organized in the following way. Section 2 presents the different DGM used in this study, including DDIM/KID-IS, the evaluation metrics, and the dataset and experimental setup used to validate the effectiveness of the models. Section 3 exhibits the results and discussion. Lastly, Sec. 4 summarizes the conclusion and discusses future research directions. In addition, Table 1 provides the list of all abbreviations and their meanings used in this paper.

2 Methods: Comparative Analysis of Generative Models in Ultrasound Image Synthesis

Three major categories of DGM are investigated and compared in this study, i.e., GANs, VAEs, and DDMs, (Fig. 1). This section is subdivided into four main subsections. In Sec. 2.1, multiple DDMs used in this study are explored, including one augmented DDM recently developed by the authors. Section 2.2 presents the other two popular DGM (GAN and VAE). Section 2.3 presents the different evaluation metrics used to evaluate and validate the best models. Finally, Sec. 2.4 elaborates on the dataset and experimental setup.

2.1 Denoising Diffusion Model. As discussed in Sec. 1, the diffusion model is an emerging generative model with impressive generative capacities similar to those of GANs [21,22]. The DDMs (Fig. 1(*c*)) progressively (stepwise) transform the data into random noise and then generate new data by removing noise based on a

Table 1 List of abbreviations used in the paper

Set	List of abbreviations and meaning
A–O	CT: computed tomography; CNN: convolutional neural network; DCGAN; deep convolution generative adversarial network; DDM: denoising diffusion model; DDIM: denoising diffusion implicit models; DDPM: denoising diffusion probabilistic model; DGM: deep generative models; GAN: generative adversarial network; GAN-ADA: generative adversarial network with adaptative discriminator augmentation; H: Entropy; IGDD: image-guided drug delivery; IS: inception score; KDE: kernel density estimation; KID: kernel inception distance; LPIPS: learned perceptual image patch similarity; MAE: mean absolute error; ML: machine learning; MRI: magnetic resonance imaging; MS-SSIM: multiscale structural similarity index measure; NF: normalizing flow; OCT: optical coherence tomography; OOD: out of distribution.
P–Z	PET: positron emission tomography; PSNR: peak signal-to-noise ratio; RMSE: root mean square error; ROI: region of interest; SSIM: structural similarity index measure; VAE: variational auto-encoder; WGAN: Wasserstein generative adversarial network

denoising process. The timestep capability of the diffusion model motivates using DDM for monitoring tumor treatment using ultrasound images for IGDD. Tumor treatment using IGDD requires an understanding of both the spatial and temporal evolution of the tumor. The iterative timestep diffusion destruction and reconstruction of diffusion models should enable learning intricate patterns of ultrasound images to effectively monitor tumor treatment responses. In this study, three DDMs are leveraged: (1) the denoising diffusion probabilistic model (DDPM) [21], (2) the denoising diffusion implicit model (DDIM) [22], and (3) the DDIM/KID-IS, which was originally developed in our recent study [20] and tailored in this work.

2.2 Denoising Diffusion Probabilistic Models and Denoising Diffusion Implicit Models. Denoising diffusion probabilistic models (DDPM) have demonstrated its capability of generating high-quality images. However, one of its practical disadvantages is its slow sampling process [22]. Training a DDPM involves many steps and iterations before achieving high-quality generation, as it uses a Markovian diffusion process (Fig. 2). As reported by Akbar et al. [23], their DDPM took a day and a half to generate 100,000 images compared to a few minutes for styleGAN. In DDPM, a standard Gaussian distribution is used (with variance $\beta_1, \ldots, \beta_T$) to transform the data into noise, while the denoising step is performed through a neural network. The forward and reverse diffusion of DDPM are represented by Eqs. (1) and (2), respectively. For this study, DDPM is investigated with three models assessed at various timesteps: 100 timesteps (DDPM_100), 300 timesteps (DDPM_300), and 1000 timesteps (DDPM_1000).

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right) \tag{1}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{2}$$

where $\mathcal{N}(\cdot)$ represents the normal distribution, t is the random timestep, and β_t is the variance scale coefficient. $\mathbf{x}_t$ is the diffused data at time-step t. $\mu_\theta(\cdot)$ and $\Sigma_\theta(\cdot)$, respectively, denote the mean and variance coefficient of the reverse step with the learnable parameter θ.

Denoising diffusion implicit models (DDIM), which is an extension of the DDPM, utilizes a forward and reverse non-Markovian diffusion process to perturb the data and generate new samples [22]. Specifically, as illustrated in Fig. 2, the non-Markovian diffusion process leveraged by DDIM uses the noisy input $\mathbf{x}_t$ and a prediction $\mathbf{x}_0$, along with a reverse conditional distribution denoted by $i_\varphi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, to determine a sample $\mathbf{x}_{t-1}$. This mechanism is further described by Eq. (3). Through this approach, DDIM provides faster sampling generation compared to DDPM [22]

$$\mathbf{x}_{t-1} = \sqrt{\omega_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\omega_t}\tau_\theta^{(t)}(\mathbf{x}_t)}{\sqrt{\omega_t}}\right) + \sqrt{1-\omega_{t-1}-\varphi_t^2}\cdot\tau_\theta^{(t)}(\mathbf{x}_t) + \varphi_t\tau_t \tag{3}$$

where ω_t is the noise rate at step t, τ is the noise variable, θ represents the learnable parameter capturing the probabilistic relationship between the noisy and the clean images, and φ controls the stochasticity of the forward process.

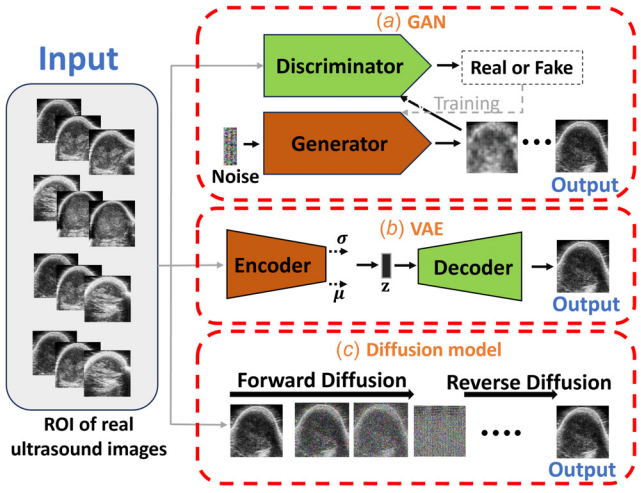

Fig. 1 Overview of DGM used in this comparative study to generate synthetic ultrasound images: (*a*) GAN generates samples with a generator playing an adversarial game against a discriminator; (*b*) VAE utilizes a variational encoder and decoder for sample generation, and (*c*) diffusion generative model generates synthetic sample by reconstructing new samples from noisy samples

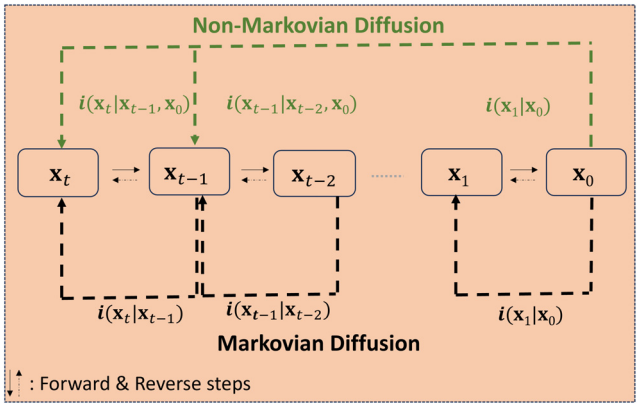

Fig. 2 An illustration of Markovian (bold black) versus non-Markovian (green) diffusion processes used in diffusion models. In Markovian diffusion, $i(\mathbf{x}_t|\mathbf{x}_{t-1})$, the probability of a future state is dependent on the current state, whereas non-Markovian diffusion $i(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)$, is dependent on the history of the system.

2.2.1 Denoising Diffusion Implicit Model/Kernel Inception Distance-Inception Score. Denoising diffusion implicit model/KID-IS is an augmented DDIM via improved similarity metrics to provide a balance between similarities and diversity, which was developed in our previous studies as a monitoring tool for additive manufacturing [20]. DDIM/KID-IS can reduce the number of unrealistic new samples while increasing the potential of exploring diversity. As depicted in Fig. 3(*a*), DDIM/KID-IS involves three training steps: (1) noise injection (or forward diffusion), (2) denoising, and (3) calculation of losses (image and noise) and KID-IS score. Due to the challenges of ultrasound image textures, DDIM/KID-IS in this study is modified and adapted to capture the underlying distribution of the ultrasound images.

(1) The DDIM/KID-IS has been modified to include two feature extractions in the KID-IS distance metric calculation. The two features' extractions are Inception and Xception CNN-based encoders which can capture information from the ultrasound images at various scales. Notice that after each training epoch, the KID-IS score is measured between the real and the generated ultrasound images. Afterward, the KID-IS score along lines with the image loss and noise loss is used to update the denoising step of the following training epochs.
(2) The other change concerns the U-Net architecture of the diffusion model used for denoising. Ultrasound images have noisy interference (speckle), and therefore while diffusion models inject noise into the dataset, it might lead to more texture complexity. To bypass this challenge the U-net of the DDIM/KID-IS has been made to have a wider block architecture to provide a better denoising step. Figure 3(*b*) provides more details about the KID-IS metric included in our diffusion model, indicated by the arrow connecting the KID-IS score in Figs. 3(*a*) and 3(*b*). Since KID-IS is calculated using features from both the real inputs and the generated outputs, two arrows connect these components to Fig. 3(*b*). With these modifications, DDIM/KID-IS should provide better image generation but could also lead to extended training time.

The distance metric in DDIM/KID-IS represented by Eq. (4) is a combination of inception score (IS) [24] and kernel inception distance (KID) [25]. To validate the necessity of using the updated KID-IS in this study, the original version of the model, termed KID-IS/O, is also evaluated.

$$\text{KID-IS} = \frac{2(\text{m-KID} \times \text{IS})}{\text{m-KID} + \text{IS}} \tag{4}$$

Specifically, the m-KID and IS can be expressed respectively as Eqs. (5) and (6).

$$\text{m-KID} = E_{\mathbf{x},\mathbf{x}\sim R}\left[k_c(\mathbf{x},\mathbf{x})\right] - 2E_{\mathbf{x}\sim\mathbf{R},\mathbf{y}\sim\mathbf{G}}\left[k_c(\mathbf{x},\mathbf{y})\right] + E_{\mathbf{y},\mathbf{y}\sim\mathbf{G}}\left[k_c(\mathbf{y},\mathbf{y})\right] \tag{5}$$

where $E_{\mathbf{y},\mathbf{y}\sim\mathbf{G}}\left[k_c(\mathbf{y},\mathbf{y})\right]$ represents the expected mean kernel extracted from the generated images ($\mathbf{G}$), and $E_{\mathbf{x}\sim\mathbf{R},\mathbf{y}\sim\mathbf{G}}\left[k_c(\mathbf{x},\mathbf{y})\right]$ represents the expected cross-mean kernel between the real images and the generated images. $\mathbf{x}$ and $\mathbf{y}$ are vector data points or samples from the data distribution of $\mathbf{R}$ and $\mathbf{G}$, respectively. k_c is a combination of a polynomial kernel and a Gaussian kernel.

$$\text{IS} = \exp\left(E_{\mathbf{y}\sim\mathbf{G}}\left[D_{KL}\left(p(v|\mathbf{y})||p(v)\right)\right]\right) \tag{6}$$

where $E_{\mathbf{y}\sim\mathbf{G}}$ represents the expectation of the generated ultrasound images $\mathbf{G}$ and $\mathbf{y}$ is an image sampled from $\mathbf{G}$. The $D_{KL}(\cdot)$ is the Kullback–Leiber divergence between the conditional class distribution $p(v|\mathbf{y})$ of the generated ultrasound images and the marginal class distribution $p(v)$ of all generated samples. v is the output class label.

2.3 Two Other Popular Deep Generative Models: Generative Adversarial Network and Variational Auto-Encoders

2.3.1 Generative Adversarial Network. Generative adversarial network is a deep generative model using a generator to capture the distribution of the training data while simultaneously attempting to deceive the discriminator in an adversarial game, as illustrated in Fig. 1(*a*) [26]. GAN could also be formulated by a minimax game, as shown in the following equation:

$$\min_G \max_D V(D,G) = E_{\mathbf{x}\sim P_{\text{data}(\mathbf{x})}}[\log\ (D(\mathbf{x}))] + E_{z\sim P_{\text{data}(\mathbf{z})}}[\log(1 - D(G(\mathbf{z})))] \tag{7}$$

where G is the generator network mapping z (the latent space) to $\mathbf{x}$ (the input space). D is the discriminator mapping $\mathbf{x}$ to a classification of the generated ultrasound images as real or fake.

In practice, GAN might suffer from training issues such as poor convergence and generalization [27], vanishing gradient and mode collapse [17], optimization, and memorization issues [28]. It is important to note that some of these issues can also affect other DGM such as VAE. Thus, GAN has also been extended to improve upon the standard model based on its limitations and specific applications.

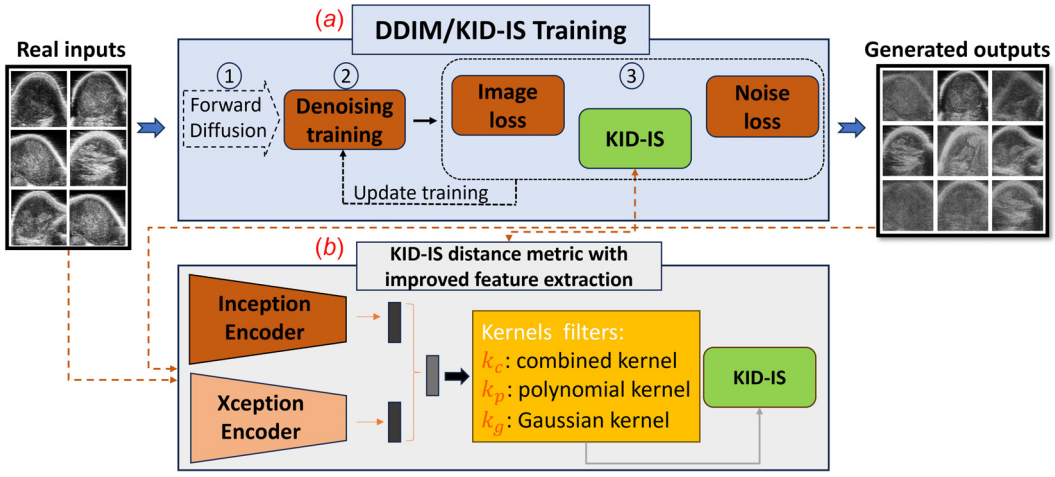

Fig. 3 An overview of DDIM/KID-IS framework. Real ultrasound images are injected with noise and reconstructed. The KID-IS function, added to the diffusion model, extracts features of the real and reconstructed samples using the Inception and Xception encoders to calculate a score that strikes a balance between similarity and diversity.

For this study, four GAN models are utilized to generate artificial ultrasound images. The first model is the Wasserstein GAN (WGAN). WGAN is a modified version of the original GAN which uses a critic and the Wasserstein distance [29]. The second GAN model considered in this study is the deep convolutional GAN (DCGAN), which is a specialized GAN using CNN as the architecture of the generator and discriminator [30]. The third GAN model is the least square GAN (LSGAN) which employs the least squares loss function instead of the binary loss used in DCGAN [31]. Finally, the last GAN model used in this study is the GAN adaptative discriminator augmentation (GAN-ADA). GAN-ADA is used to reduce the risk of overfitting created by a very confident discriminator on a small dataset [32].

2.3.2 Variational Auto-encoder. The VAE is an extension of the autoencoder (AE) that incorporates a latent distribution, enabling the generation of samples to be reconstructed with potentially unseen samples as shown in Fig. 1(*b*) [33]. The Kullback-Leibler loss is used for the latent space with a normal distribution. The VAE decoder, which represents the generative component of the VAE, can be represented by the following equation:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_{\theta\mathbf{x}}(\mathbf{x}|\mathbf{z})p_{\theta\mathbf{z}}(\mathbf{z}) \tag{8}$$

where $p_{\theta\mathbf{z}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}_\mathrm{W}, \mathbf{I}_\mathrm{W})$, $\mathbf{x}$ is the input from the data distribution, and $\mathbf{z}$ is the latent code. $p_{\theta\mathbf{x}}(\mathbf{x}|\mathbf{z})$ is a parametric conditional, and $p_{\theta\mathbf{z}}(\mathbf{z})$ is a generic notation for a parametric probability density function (pdf) of the latent code $\mathbf{z}$. $\mathcal{N}(\cdot)$ represents the normal distribution. $\mathbf{0}_\mathrm{W}$ is the zero-vector of size W, and $\mathbf{I}_\mathrm{W}$ is the identity matrix of size W.

For this study, three VAE models are investigated including: (1) the standard VAE; (2) vector quantized VAE (VQ-VAE); and (3) VAE-GAN. Specifically, VQ-VAE uses a discrete latent space for simpler optimization instead of a continuous latent space which is used in standard VAE [34]. In practice, optimizing the gradient descent for continuous latent space tends to be quite difficult. Moreover, VAE-GAN is another extended version of the VAE that utilizes adversarial training to improve the generated samples from the standard VAE [35]. In this adversarial training, the generator of the GAN is replaced by the VAE, which is attempting to deceive the discriminator.

2.4 Evaluation Metrics

2.4.1 Similarity Evaluation. To evaluate the efficiency of the generative models in monitoring tumor treatment through image augmentation, their generated images are evaluated for similarity (Image fidelity). The similarity test assesses the quality of the generated images in terms of visual and perceptual similarity, noise presence, luminance, contrast, and structural similarities. For this purpose, five evaluation metrics are used for the similarity test in this study. The first metric is the Kernel Inception Distance (KID). KID is a quality distance metric measuring the feature representation between the generated and real images using kernel filters to create a score of similarity [25]. KID tends to align with human visual judgment, with a lower score implying better image quality. KID score is calculated using the following equation:

$$\mathrm{KID} = E_{\mathbf{x},\mathbf{x}\sim\mathbf{R}}\left[k(\mathbf{x},\mathbf{x})\right] - 2E_{\mathbf{x}\sim\mathbf{R},\mathbf{y}\sim\mathbf{G}}\left[k(\mathbf{x},\mathbf{y})\right] + E_{\mathbf{y},\mathbf{y}\sim\mathbf{G}}\left[k(\mathbf{y},\mathbf{y})\right] \tag{9}$$

where $E_{\mathbf{x},\mathbf{x}\sim\mathbf{R}}\left[k(\mathbf{x},\mathbf{x})\right]$ denotes the expected mean kernel features extracted from the real images ($\mathbf{R}$). $E_{\mathbf{y},\mathbf{y}\sim\mathbf{G}}\left[k(\mathbf{y},\mathbf{y})\right]$ represents the expected mean kernel extracted from the generated images ($\mathbf{G}$). $E_{\mathbf{x}\sim\mathbf{R},\mathbf{y}\sim\mathbf{G}}\left[k(\mathbf{x},\mathbf{y})\right]$ represents the expected cross-mean kernel between the real images and the generated images. $\mathbf{x}$ and $\mathbf{y}$ are vector data points or samples from the data distribution of $\mathbf{R}$ and $\mathbf{G}$, respectively.

The second metric is the Learned Perceptual Image Patch Similarity (LPIPS), which is an objective metric that utilizes CNN feature representation ability to map a distance that correlates with human judgment [36]. The Structural Similarity Index Measure (SSIM), as shown in Eq. (10), is the third metric used to measure image quality in terms of luminance, contrast, and structural similarities [37]

$$\mathrm{SSIM} = \frac{\left(2\mu_R\mu_G + \tau^2 l^2\right)\left(2\sigma_{RG} + \tau'^2 l^2\right)}{\left(\mu_R^2 + \mu_G^2 + \tau^2 l^2\right)\left(\sigma_R^2 + \sigma_G^2 + \tau'^2 l^2\right)} \tag{10}$$

where μ_R, μ_G represent the pixel mean of the real images ($\mathbf{R}$) and generated images ($\mathbf{G}$). σ_R^2, σ_G^2 represent the variance of the real and generated images, and σ_{RG} represents the covariance between $\mathbf{R}$ and $\mathbf{G}$. l is the dynamic range, and τ *and* τ' constant values set at 0.01 and 0.03, respectively. When SSIM is extended to measure structural similarity index using different image scales, termed multiscale SSIM (MS-SSIM) [38], as shown in the following equation:

$$\text{MS-SSIM} = \frac{1}{s}\sum_{1}^{s}\mathrm{SSIM} \tag{11}$$

where s is the number of scales. It will be the fourth metric in this study.

The last metric is the peak signal-to-noise ratio (PSNR) as shown in Eq. (12). PSNR evaluates the extent of noise relative to the peak signal within the two sets of images [37]. A high value of PSNR implies a high similarity between the real and generated ultrasound images.

$$\mathrm{PSNR} = 10\log_{10}\frac{R^2}{\mathrm{MSE}} \tag{12}$$

where R is the maximum pixel fluctuation in the real images, and MSE is the mean square error difference between the real images ($\mathbf{R}$) and generated images ($\mathbf{G}$). It is important to note that PSNR, SSIM, and MS-SSIM are not necessarily used to match human visual judgment but can serve as indicators of quality, and therefore they are complementary to the other two metrics in this study.

2.4.2 Diversity Evaluation. The generated samples must exhibit a certain degree of realistic diversity not observed in the training dataset. Achieving this level of realistic diversity suggests that the generative model has attained an understanding of the data distribution, enabling it to predict potential outcomes of tumor treatment. To assess the level of diversity of the generated ultrasound images, the entropy (H) and kernel density estimation (KDE) of the generated images are calculated. The entropy (H), as shown in Eq. (13), quantifies the level of randomness of pixel intensities [39]. The entropy difference between the real and generated images is then calculated. A high entropy difference score would indicate more details and diversity between the two sets of images, whereas a lower score would indicate more uniformity or simplicity between the two sets of images. H and KDE must be considered along with the similarity score measurements of Sec. 2.3.1, as it is expected that unrealistic or imaginary generated samples would also lead to an extremely high H and KDE.

$$\mathrm{H} = -\sum_{i=1}^{\mathrm{I}} p_i \log_2 p_i \tag{13}$$

where I is the number of pixel intensities in the image and p_i probability mass function of the pixel intensity. Moreover, KDE is also used to get some visual insights into the pixel intensity distribution. A diverse set of generated ultrasound images should produce a KDE that is inherently different from the KDE of real ultrasound images.

2.4.3 Prediction Evaluation. Although generative models can generate ultrasound images that can be diverse in features and similar to those seen within the training dataset, the predictive ability of the models must also be tested. The generative models are trained

on a set of ultrasound images that include a small minority class, with a separate test set reserved for evaluation. First, the predicted ultrasound images are assessed for their similarity and diversity in terms of the extracted feature space compared to the test samples. The predicted samples and test samples are clustered using the K-means clustering technique. Afterward, the silhouette score (SC) is used to determine the optimal number of clusters. SC measures the degree to which clusters are distinct from each other [40]. SC ranges between −1 (i.e., wrong cluster) and 1 (i.e., correct cluster) and it can be calculated using Eq. (14). If the predicted samples and the test samples are clustered together, then it implies they have a high level of closeness in terms of extracted feature space. It is important to note that this clustering test does not measure the level of accuracy in prediction, quality, or similarity, but rather the proximity in the extracted feature space. The dimensions of the images are reduced by using principal component analysis before K-means clustering is applied.

$$\mathrm{SC} = \frac{u - v}{\max(u, v)} \tag{14}$$

where u is the mean distance for points in the same cluster and v is the mean distance for different clusters. Finally, the root-mean-square error (RMSE) and the MAE between the predicted samples and the real samples are calculated. A good prediction should result in low RMSE and MAE. RMSE and MAE are illustrated, respectively, in the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{g} - r)^2} \tag{15}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|(\hat{g} - r)\right| \tag{16}$$

where $\hat{g}$ is the predicted generated ultrasound image, r is the real ultrasound image, and n is the total number of samples.

2.5 Dataset Description and Experimental Setup. In this study, ultrasound images are collected from an experimental dataset of mice subjects bearing colon cancer. The tumor was initiated in athymic nude mice with colon carcinoma cells. Each mouse in the study was anesthetized with less than 5% isoflurane and afterward kept in a 37 °C room. Several treatments were used to treat the mice and the formulations have already been reported in previous studies [3,4,41]. The ultrasound images were captured using a Visual Sonics Vevo 2100 ultrasound MS550D transducer (22–55 MHz) when the tumors grew to a magnitude between 300 and 400 cm^3. Note that just the body part of the mouse bearing the tumor was captured. The ultrasound images are assessed per time with several frames at various points of time. Each mouse is treated under the rules, procedures, and regulations of the Oklahoma State University's Institutional Animal Care and Use Committee (OSU IACUC).

Two cases are used to validate the generative models (See Table 2). In case 1, 900 ultrasound images (or frames) of a mouse from ten image classes (videos) are used to validate the similarity and diversity of the generated ultrasound against the real ultrasound images. The treatment group of this mouse is done with high-intensity focused ultrasound (HIFU) and echogenic low-temperature sensitive liposomes (E-LTSL). The mouse is injected with 10 mg of E-LTSL and 50 *ul* of saline at 42 °C and 37 °C. The low temperature is due to the heat sensitivity of E-LTSL. After the ultrasounds are collected, the ROI is manually segmented based on domain knowledge (See Fig. 4) and then fed to the DGMs discussed in Secs 2.2 and 2.3.

Table 2 Cases 1 and 2 setups

Parameter	Case 1	Case 2
Test	Similarity and diversity	Reality and prediction
Number of videos	10	4
Sample size (Frames)	900	610
Treatment	(HIFU) + (E-LTSL)	Doxorubicin
Mouse and cage number	Cage 10 mouse 2	Cage 10 mouse 4
Temperature	42 °C	37 °C

In case 2, after validating the efficiency of each deep generative model (by similarity and diversity), the best models are used and evaluated on a new mouse with a different treatment called doxorubicin. This is to determine the capability of the DGM to generate data not seen during training or with a few occurrences in the training set. This test aims to determine the efficacy of the models to predict potential tumor evolution following the administration of a new treatment. For case 2, 610 ultrasound images (or frames) from four classes (videos) of ultrasound images are used to train the model. Three out of the four classes of images have 200 ultrasound images each per class. The fourth class, which represents the minority class from which ultrasound images must be predicted, has only ten images present in the entire dataset. The generative models are used to predict/generate samples from the imbalance/minority class. Note that for each case, the ultrasound images are frames captured from videos after the treatment is injected into the ROI. For case 1, the 900 frames are captured from ten videos, and for case 2, the 610 frames are captured from four videos.

All ultrasound images are trained at a size of 128 × 128, with a batch size of 10, over 300 epochs. Training is mostly implemented using an Nvidia RTX A2000 12 GB GPU on a Windows 10 operating system with the TensorFlow 2.10 library and Python 3. Due to the computational resource demands of certain models, models such as DDPM, which require a more powerful GPU, were trained using a Tesla T4 GPU on Google Colab or Kaggle. Table 3 presents the architecture and hyperparameters of all models used in this study. Specifically, the reported training time (t) in Table 4 is measured using a Tesla T4 GPU on Google Colab for all models trained in the study to ensure consistency in the reported metric.

3 Results and Discussion

3.1 Image Fidelity: Similarity Evaluation. The generated ultrasound images by the thirteen generative models are evaluated on their likeness to the real images using five evaluation metrics. The results are reported in Table 4 and can be supported by examples of generated ultrasound images to validate the findings through human visual inspection (see Fig. 5). Visually five out of the thirteen models could capture and represent the noisy complexity of tumor ultrasound images, namely, DDPM_1000, DDIM/KID-IS, DCGAN, WGAN, and VQ-VAE. For GANs, both GAN-ADA and LSGAN could not fully comprehend the feature distribution of the training set leading to training instability with no convergence [17]. As depicted in Fig. 6, the generators for GAN-ADA and LSGAN collapsed before being able to generate realistic samples, resulting in unrealistic generated images. Therefore, the shown images in Fig. 5 for LSGAN and GAN-ADA are collected earlier in training.

Variational auto-encoders and DDMs tend to have smoother training compared to the GAN models since GAN training involves an adversarial game. However, some VAEs and DDMs also result in some poorly generated samples. For instance, VAE and VAE-GAN suffered from mode collapse by generating only one particular ultrasound image from the entire training dataset, in addition to a loss of color (resulting in almost pure grey), as shown in Fig. 5. DDPM_100 also generated some low-quality samples, since it requires a lot more sampling steps and epochs to generate high-quality ultrasound images, according to Ref. [22]. While DDIM improves the quality of the generated samples dramatically (the KID score is almost seven times better than DDPM_100, see Table 4), it tends to generate more unrealistic ultrasound images than its GAN counterpart.

Furthermore, as reflected in Table 4, increasing the timesteps from 100 to 1000 does not impact the training time for DDPM, but it does impact the inference time. It is important to note that although

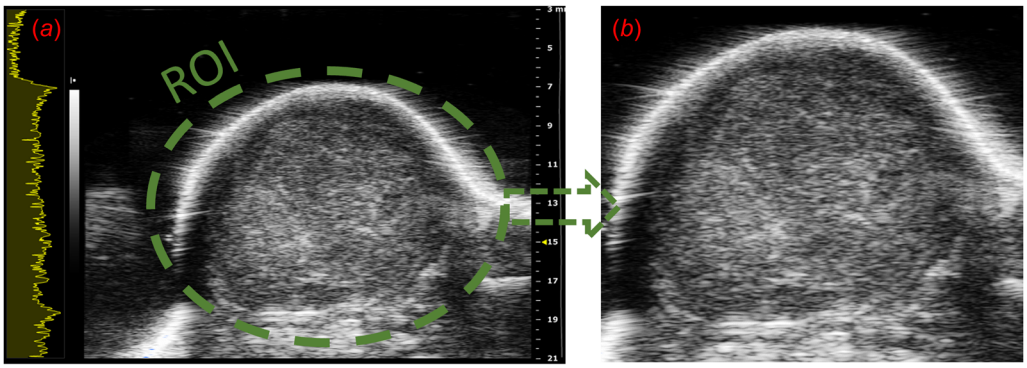

Fig. 4 A sample of an ultrasound image of a mouse injected with drug treatment: (*a*) raw ultrasound image including ROI, frequency plot, and the remaining part; and (*b*) extracted ROI from the raw ultrasound image

Table 3 Hyperparameter and architecture of the DDMs, GAN models, and VAE models used in this study

Diffusion model architecture					
Model	Weight decay	Learning rate	Min/max signal rates	timesteps	U-net block depth/width
DDIM based	0.0002	0.001	0.02/0.98	20	2 / (32, 64, 96, 128) except DDIM/KID-IS: (32, 64, 96, 128, 160, 192, 224, 256)
DDPM based	N/A	0.0002	N/A	100, 300, 1000	

GAN architecture					
	beta	Learning rate	Latent dimension	Generator nodes	Discriminator
LSGAN	N/A	0.0002	128	128, 256, 512	128, 128, 128
GAN-ADA	0.5	0.0002	128	128, 128, 128, 128	128, 128, 128, 128
DCGAN	N/A	0.0001	128	128, 256, 512	128, 128, 128
WGAN	0.5/0.9	0.0002	128	512, 256, 128	128, 256, 512

VAE architecture				
	Latent dimension	Encoder	Decoder	Discriminator/VQ
VAE	128	128, 128, 128	128, 256, 512	N/A
VAE-GAN	128	128, 128, 128	128, 256, 512	128, 64, 32, 128
VQ-VAE	128	32, 64	64, 32	64

the training times for DDPM_100 and DDPM_300 are slightly lower than DDIM/KID-IS, DDIM/KIDS still achieves a higher sampling quality with faster inference time. DDPM_100 and DDPM_300 did not generate any high-quality samples because the number of timesteps (100 and 300) is considered too low for DDPM to achieve high-performance generation. DDPM trained with 1000 steps was capable of generating samples with comparable or superior quality to that of DDIM/KID-IS, which is trained with only 20 steps. It is

Table 4 Five similarity evaluation metrics results for thirteen models with their training time

Models		KID ↓	MS-SSIM ↑	SSIM ↑	PSNR ↑	LPIPS ↓	Training (t) (Seconds)
DDM	DDIM	5.5923	0.9929	0.0997	11.9628	0.3110	12
	KID-IS/O	4.7939	0.9951	0.1570	13.4488	0.2976	38
	DDIM/KID-IS	3.5747	0.9969	0.2819	**16.2234**	0.1570	43
	DDPM_100	38.2904	0.9915	0.0724	11.2846	0.5124	39
	DDPM_300	14.4327	0.9918	0.1180	11.6431	0.3155	39
	DDPM_1000	**3.2235**	0.9964	0.3299	15.0944	**0.1308**	39
GAN	DCGAN	8.83	0.9927	0.1298	11.904	0.3085	23
	WGAN	5.5784	0.9956	0.2375	13.8279	0.2229	45
	LSGAN	66.3273	0.9868	0.0811	9.4452	0.5372	33
	GAN-ADA	15.4744	0.9876	0.0957	10.1203	0.6750	5
VAE	VAE-GAN	52.7710	0.9880	0.1607	10.5255	0.4359	45
	VAE	25.9777	0.9888	0.1500	10.7611	0.4593	17
	VQ-VAE	6.9951	**0.9990**	**0.5858**	**20.7641**	0.194	4

For KID and LPIPS, "↓" means that the lowest score implies the highest quality. For MS-SSIM, SSIM, and PSNR, "↑" means that the highest score implies the highest quality. The reported training (t) is the average training time per epoch in seconds, measured using a Tesla T4 GPU on Google Colab. For each similarity evaluation metric, the best model is highlighted in black bold, and the second best is underlined.

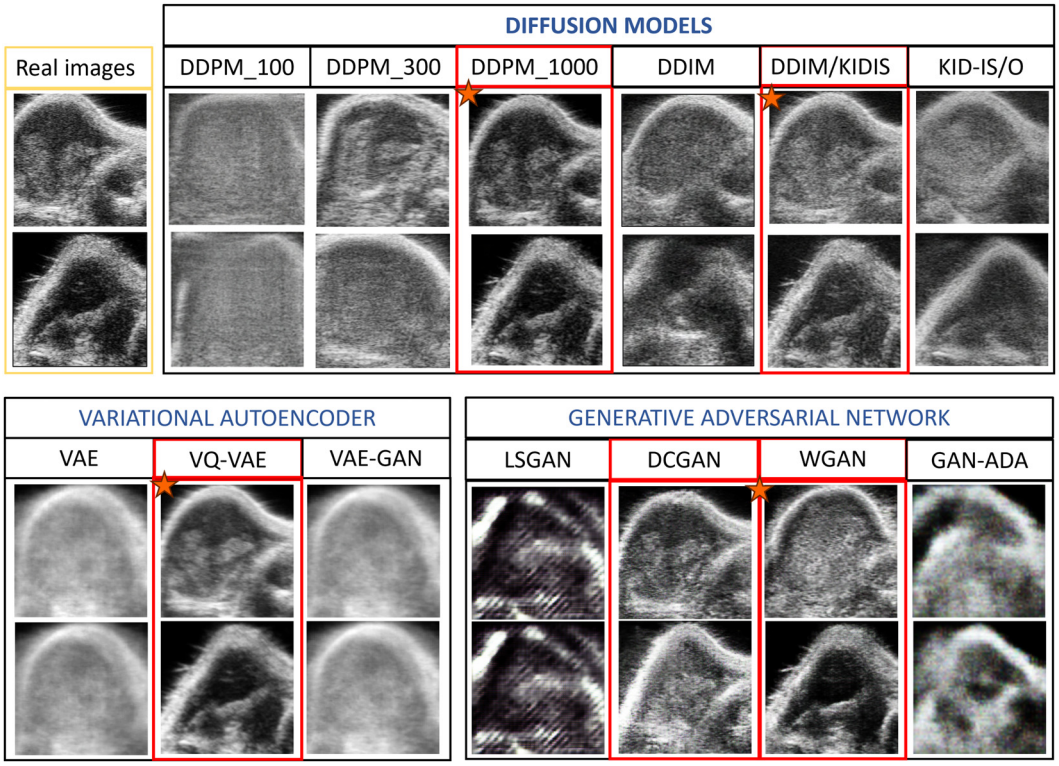

Fig. 5 Generated images for the thirteen generative models compared against real ultrasound images. From left to right: Top row: Diffusion models - DDPM_100, DDPM_300, DDPM_1000, DDIM, DDIM/KID-IS, KID-IS/O. Bottom row (right side): GAN models—LSGAN, DCGAN, WGAN, GAN-ADA. Bottom row (left side): VAE models—VAE, VQ-VAE, VAE-GAN.

noteworthy that the inference time for DDPM_1000 to generate 100 samples is about 700 s (more than 11 min), whereas it takes less than 10 s for the other DGM models, including our DDIM/KID-IS model.

When analyzing the five best models selected visually as depicted in Fig. 5, DDIM/KID-IS and DDPM_1000 achieved the highest quality overall based on the five similarity metrics, though the high inferential time of DDDPM_1000 makes it impractical. Moreover, VQ-VAE had the highest value for SSIM, MS-SSIM, and PSNR as seen in Table 4. Metrics such as SSIM, which measure features like contrast and luminance, imply that VQ-VAE generates brighter and higher contrast ultrasound images compared to DDIM/KID-IS. WGAN and DCGAN also seem to generate high-quality ultrasound images visually similar to real mages, with low evaluation metrics. In Sec. 3.2, the level of diversity of the generated ultrasound images is analyzed.

3.2 Diversity Generation Evaluation. In this section, the level of diversity among the five best-generated models (see Fig. 5) is analyzed. Figure 7 depicts the training loss of four out of the five generative models with the highest quality from the similarity evaluation in Sec. 3.1. It can be observed that DDIM/KID-IS and VQ-VAE have smoother and better training convergence compared to their GAN counterparts. Although DCGAN seems to generate ultrasound images with a high likeness to the real images, the model is affected by mode collapse. A large proportion of samples originate from a specific class, indicating that the generator is not generalizing well enough to capture more features from other ultrasound images presented in the training set.

To determine the level of diversity for each of the best models, the entropy difference between the generated images and the real images is reported in Table 5. Also, the KDE plot of each of the four best models is displayed in Fig. 8.

The entropy difference of the VQ-VAE is very small and close to zero, implying the generated samples from that model are almost similar or uniform to the real samples. This could be a sign of data copying and the incapability of the model to generate unobserved data. This is validated by the KDE plot of the VQ-VAE (Fig. 8(*b*)), which closely resembles the distribution of the real images. Models

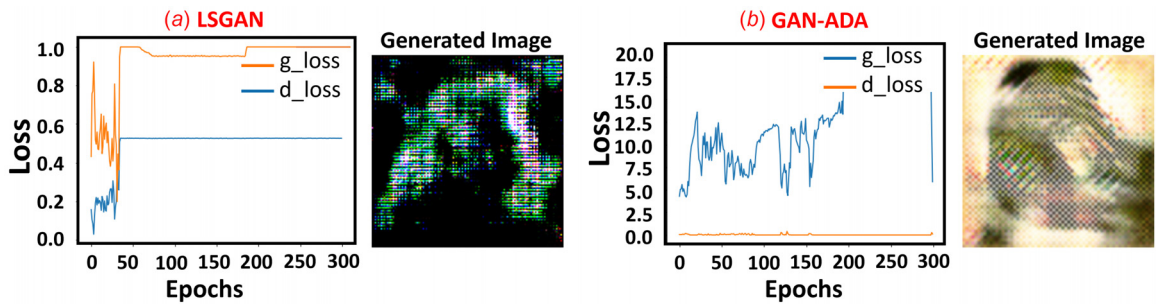

Fig. 6 Training loss (left) and a randomly generated sample (right) of two unsuccessful GAN generative models: (*a*) LS-GAN and (*b*) GAN-ADA. Both generators collapsed before generating realistic samples.

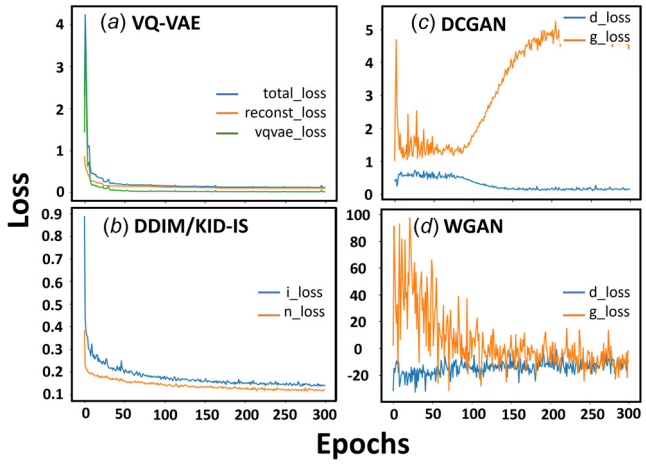

Fig. 7 Training loss versus epoch for some of the best models out of the thirteen: both (*a*) VQ-VAE; (*b*) DDIM/KID-IS lead to smoother training convergence; (*c*) DC-GAN experiences mode collapse; and (*d*) WGAN shows converging losses, but training is not smooth

Table 5 Entropy differences for the best models

Models	Entropy difference
DDIM/KID-IS	**0.0660**
DDPM_1000	0.0197
DCGAN	0.0339
VQ-VAE	0.0099
WGAN	0.0206

The highest score implies a good level of diversity. The best model is highlighted in black bold.

such as DCGAN and VQ-VAE are not suitable for tumor treatment monitoring with ultrasound images because the generated ultrasound images lack diversity to explore potential responses to tumor treatment. The behavior of the VQ-VAE is similar to standard AE which does not have a normal distribution and cannot generate new unobserved samples. DDIM/KID-IS has the highest entropy difference score and the most diversified KDE compared to the other models. This indicates that the diffusion model can generate ultrasound images similar to real ultrasound images with a degree of diversity, which will be potentially useful for monitoring the progress of tumor treatment.

3.3 Prediction Capability Evaluation. The predicted samples and the real test samples were first evaluated using a clustering technique to determine the relationship of the extracted feature space between the real ultrasound images and the generated ultrasound images. Both predicted and real samples were grouped in a common folder without labeling the class of each sample.

Fig. 9 depicts the results of the optimal number clusters using K-means and the silhouette score for different models. Since the feature space of each sample image is extracted using principal component analysis, K-means clustering does not play the role of determining if the predicted samples match the real test samples. Instead, this test evaluates how close or diversified the feature spaces of the predicted samples are from the real images. The predicted samples could be very similar to the test samples but different in pixel measurements. The K-means clustering result shows that the predicted samples are well diversified among themselves for all the generated samples. For instance, while comparing the real samples to DDIM/KID-IS generated samples, although there are two distinct groups, K-means shows that the best number of clusters should be three (i.e., the highest silhouette score). In other words, the extracted feature spaces from the generated samples are diversified enough to be separated into subclusters. Similar findings are observed for all the generated models. However, it could also suggest that the predictive capability of the generative models could be further improved to generate synthetic samples with extracted features close to the real test feature space.

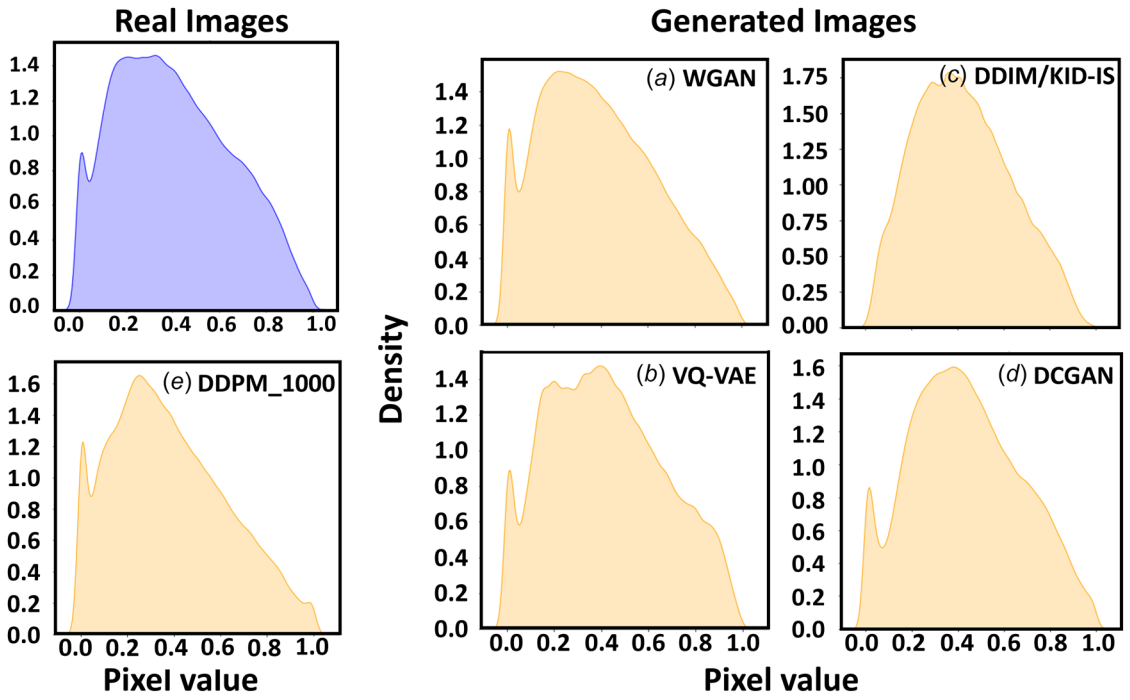

Fig. 8 KDE plot comparing real images to the generated images based on pixel value: (*a*) WGAN, (*b*) VQ-VAE, (*c*) DDIM/KID-IS, (*d*) DCGAN, and (*e*) DDPM_1000. The similarity of VQ-VAE's KDE to the real images suggests potential data copying. DDIM/KID-IS achieves the most diversified KDE.

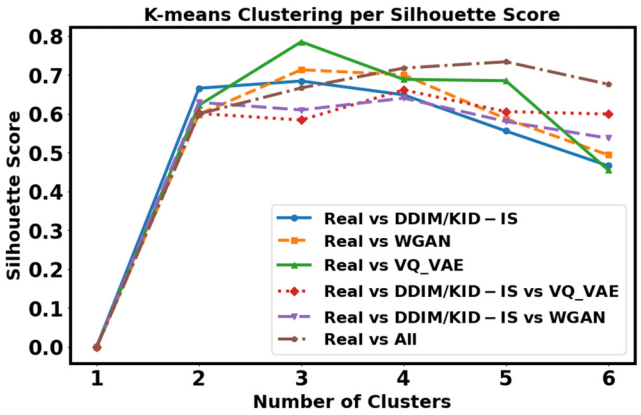

Fig. 9 The optimal number of clusters for different models. All models achieve the highest silhouette score with three clusters, although there are only two distinct groups, implying the level of diversity within the generated samples.

Table 6 Predictive performance of the three best models per category

Models	RMSE ↓	MAE ↓	LPIPS ↓
DDIM/KID-IS	**0.1284**	**0.1010**	**0.1606**
WGAN	0.2139	0.1480	0.2037
VQ-VAE	0.1812	0.1211	0.1889

For RMSE and MAE, the "↓" indicates that lower error signifies better prediction. The best model is highlighted in bold.

Furthermore, Table 6 reports the performance of the predictive ability of the best generative models per category. Although VQ-VAE has a lower RMSE and MAE than WGAN, the model suffers from data copying. Many of the generated samples by VQ-VAE for the minority class are duplicates from previous epoch training, making those predicted images unreliable. Furthermore, DDIM/KID-IS demonstrates its superiority over the other two models by having the lowest RMSE and MAE for the predicted samples.

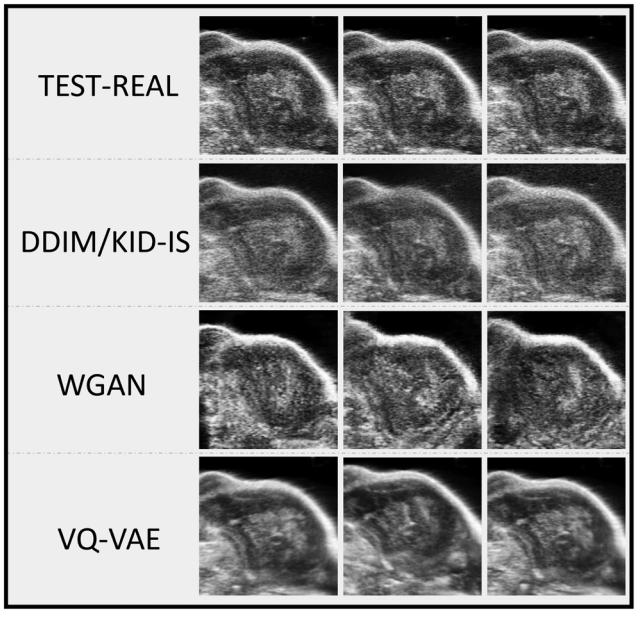

Fig. 10 Predicted synthetic ultrasound images using various generative models. From top to bottom: real samples; DDIM/KID-IS; WGAN; VQ-VAE.

Moreover, DDIM/KID-IS has the lowest LPIPS, as seen in Table 6 implying that the predicted minority samples have some perceptual similarity to the test samples. It is worth noting that due to the low sample size of the minority class, fewer cases of generated samples from the minority class are expected. In future work, the models can be modified for conditional image augmentation. Figure 10 displays samples of the predicted samples generated by the three generative models alongside the real test samples. Although DDIM/KID-IS can predict certain aspects of the test set and demonstrate its potential to monitor tumor treatment, it remains challenging to discern and quantify tumor treatment effects on the features of the weakened tissue in ultrasound images. Establishing a feature control-based model might offer a potential solution to address this issue in future work and also demonstrate clinical benefits.

4 Conclusions and Future Work

Image augmentation is the first step in monitoring tumor treatment response using generative models. When these generative models generate synthetic ultrasound images of high quality that are realistic and diverse, they can help predict the potential responses to tumor treatment. Studying such augmentation would be beneficial for improving therapy design.

This paper presents a comparative study of various generative models and their capability to monitor tumor treatment of colon cancer in mice via image augmentation. Specifically, DDMs (DDIM/KID-IS, KID-IS/O, DDIM, DDPM) are compared to widely used models GAN (DCGAN, WGAN, LSGAN, GAN-ADA) and VAEs (VAE, VQ-VAE, VAE-GAN) across two cases study. After testing each model in terms of similarity and diversity, the models DDPM_1000, DDIM/KID-IS, VQ-VAE, WGAN, and DCGAN seem to outperform their counterparts with DDIM/KID-IS getting the best performance in terms of perceptual quality metrics and diversity (entropy difference and KDE plot). Models such as VAE and VQ-VAE tend to suffer from significant issues such as mode collapse or data copying. In terms of prediction, DDIM/KID-IS had the lowest RMSE, MAE, and LIPSI, demonstrating the potential of diffusion models to excel in tumor treatment monitoring using ultrasound images. Despite the performance of DDIM/KID-IS, the model did generate a few instances of unrealistic samples showing that the model still needs further improvement.

Several studies still need to be undertaken for better understanding in the future work, which could be summarized by three main directions. First, the diffusion model needs to be further fine-tuned to accommodate the noisy and complex features of ultrasound and reduce the number of unrealistic generated synthetic ultrasound images. Second, the predictive model needs to be tailored to capture clear potential changes in the tumor ROI due to the provided treatment. These generated potential responses need to be investigated to determine the reasons that caused them to enhance therapy design. This implies establishing a quantified relationship between tumor responses and parameters impacting the responses. Additionally, future analysis must include more fairness metrics in the study to capture an appropriate level of sensitive attributes and implement relative measurements demonstrating medical significance with clinical benefits and interpretation. Finally, more diffusion models need to be investigated across additional cases and various conditions to establish the effectiveness of DDIM/KID-IS in monitoring tumor treatment in ultrasound images and possibly other image modalities.

Acknowledgment

This research report was supported by the Oklahoma Center for the Advancement of Science and Technology (OCAST) Health Research Program under Award Number HR23-049 and partially supported by the National Science Foundation (NSF) under Award No. TI-2234619. The authors would also like to thank the Oklahoma State University CEAT Engineering Research and Seed Funding Program for continued support.

Funding Data

- National Science Foundation (Award No. 2234619; Funder ID: 10.13039/100000001).

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets and codes generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

References

[1] NIH News in Health, 2019, "Medical Scans Explained," NIH News in Health, Bethesda, MD, accessed Dec. 7, 2023, https://newsinhealth.nih.gov/2019/11/medical-scans-explained
[2] Priester, M. I., and ten Hagen, T. L. M., 2023, "Image-Guided Drug Delivery in Nanosystem-Based Cancer Therapies," Adv. Drug Delivery Rev., **192**, p. 114621.
[3] Mirniaharikandehei, S., VanOsdol, J., Heidari, M., Danala, G., Sethuraman, S. N., Ranjan, A., and Zheng, B., 2019, "Developing a Quantitative Ultrasound Image Feature Analysis Scheme to Assess Tumor Treatment Efficacy Using a Mouse Model," Sci. Rep., **9**(1), p. 7293.
[4] Ektate, K., Kapoor, A., Maples, D., Tuysuzoglu, A., VanOsdol, J., Ramasami, S., and Ranjan, A., 2016, "Motion Compensated Ultrasound Imaging Allows Thermometry and Image Guided Drug Delivery Monitoring From Echogenic Liposomes," Theranostics, **6**(11), pp. 1963–1974.
[5] Bharti, P., and Mittal, D., 2020, "An Ultrasound Image Enhancement Method Using Neutrosophic Similarity Score," Ultrason Imaging, **42**(6), pp. 271–283.
[6] Liu, C., Kapoor, A., VanOsdol, J., Ektate, K., Kong, Z., and Ranjan, A., 2018, "A Spectral Fiedler Field-Based Contrast Platform for Imaging of Nanoparticles in Colon Tumor," Sci. Rep., **8**(1), p. 11390.
[7] Li, Y., VanOsdol, J., Ranjan, A., and Liu, C., 2022, "A Multilayer Network-Enabled Ultrasonic Image Series Analysis Approach for Online Cancer Drug Delivery Monitoring," Comput. Methods Programs Biomed., **213**, p. 106505.
[8] Gao, Y., Maraci, M. A., and Noble, J. A., 2016, "Describing Ultrasound Video Content Using Deep Convolutional Neural Networks," IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, Czech Republic, Apr. 13–16, pp. 787–790.
[9] Atrey, K., Singh, B. K., and Bodhey, N. K., 2023, "Multimodal Classification of Breast Cancer Using Feature Level Fusion of Mammogram and Ultrasound Images in Machine Learning Paradigm," Multimed. Tools Appl., **83**(7), pp. 21347–21368.
[10] Shijie, J., Ping, W., Peiyi, J., and Siping, H., 2017, "Research on Data Augmentation for Image Classification Based on Convolution Neural Networks," Chinese Automation Congress (CAC), Jinan, China, Oct. 20–22, pp. 4165–4170.
[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., 2002, "SMOTE: Synthetic Minority Over-Sampling Technique," J. Artif. Intell. Res., **16**, pp. 321–357.
[12] Cao, H., Tan, C., Gao, Z., Chen, G., Heng, P.-A., and Li, S. Z., 2024, "A Survey on Generative Diffusion Model," IEEE Transac. Knowled. Data Eng., **36**(7), pp. 2814–2830.
[13] Devi, M. K., Alias, A., and Suganthi, K., 2021, "Review of Medical Image Synthesis Using GAN Techniques," ITM Web Conf., **37**, p. 01005.
[14] Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D., 2023, "Diffusion Models in Medical Imaging: A Comprehensive Survey," Medical Image Anal., **88**, p. 102846.
[15] Qin, X., Bui, F. M., Nguyen, H. H., and Han, Z., 2022, "Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GANs," IEEE Access, **10**, pp. 91663–91677.
[16] Rapoport, N., Kennedy, A. M., Shea, J. E., Scaife, C. L., and Nam, K.-H., 2010, "Ultrasonic Nanotherapy of Pancreatic Cancer: Lessons From Ultrasound Imaging," Mol. Pharmaceutics, **7**(1), pp. 22–31.
[17] Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., and Torralba, A., 2019, "Seeing What a GAN Cannot Generate," Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, Oct. 27–Nov. 2, pp. 4502–4511.
[18] Hajij, M., Zamzmi, G., Paul, R., and Thukar, L., 2022, "Normalizing Flow for Synthetic Medical Images Generation," IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT), Houston, TX, Mar. 10–11, pp. 46–49.
[19] Teo, C., Abdollahzadeh, M., Cheung., and N.-M., (Man), 2023, "On Measuring Fairness in Generative Models," Adv. Neural Inf. Process. Syst., **36**, pp. 10644–10656.
[20] Yangue, E., Fullington, D., Smith, O., Tian, W., and Liu, C., 2024, "Diffusion Generative Model-Based Learning for Smart Layer-Wise Monitoring of Additive Manufacturing," ASME J. Comput. Inf. Sci. Eng., **24**(6), p. 060903.
[21] Ho, J., Jain, A., and Abbeel, P., 2020, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Red Hook, NY, pp. 6840–6851.
[22] Song, J., Meng, C., and Ermon, S., 2022, "Denoising Diffusion Implicit Models," arXiv:2010.02502v4.
[23] Akbar, M. U., Wang, W., and Eklund, A., 2023, "Beware of Diffusion Models for Synthesizing Medical Images—A Comparison With Gans in Terms of Memorizing Brain MRI and Chest X-Ray Images," arXiv:2305.07644v3.
[24] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X., 2016, "Improved Techniques for Training GANs," *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Red Hook, NY.
[25] Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A., 2021, "Demystifying MMD GANs," arXiv:1801.01401v5.
[26] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., 2020, "Generative Adversarial Networks," Commun. ACM, **63**(11), pp. 139–144.
[27] Arora, S., Ge, R., Liang, Y., and Ma, T., 2017, "Generalization and Equilibrium in Generative Adversarial Nets (GANs)," Proceedings of the 34th International Conference on Machine Learning, PMLR, Sydney, Australia, Aug. 6–11, pp. 224–232.
[28] Nagarajan, V., Raffel, C., and Goodfellow, I. J., "Theoretical Insights Into Memorization in GANs," Neural Information Processing Systems Workshop, Vol. 1, p. 3.
[29] Arjovsky, M., Chintala, S., and Bottou, L., 2017, "Wasserstein Generative Adversarial Networks," Proceedings of the 34th International Conference on Machine Learning, PMLR, Sydney, Australia, Aug. 6–11, pp. 214–223.
[30] Radford, A., Metz, L., and Chintala, S., 2016, "Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks," arXiv:1511.06434v2.
[31] Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P., 2017, "Least Squares Generative Adversarial Networks," Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802.
[32] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T., 2020, "Training Generative Adversarial Networks With Limited Data," *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., Curran Associates, Inc., Red Hook, NY, pp. 12104–12114.
[33] Doersch, C., 2021, "Tutorial on Variational Autoencoders," arXiv:1606.05908v3.
[34] van den Oord, A., Vinyals, O., and Kavukcuoglu, K., 2017, "Neural Discrete Representation Learning," Advances in Neural Information Processing Systems, Curran Associates, Inc., Red Hook, NY.
[35] Mi, L., Shen, M., and Zhang, J., 2018, "A Probe Towards Understanding GAN and VAE Models," arXiv:1812.05676v2.
[36] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O., 2018, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Salt Lake City, June 18–22, pp. 586–595.
[37] Sheikh, H. R., and Bovik, A. C., 2006, "Image Information and Visual Quality," IEEE Trans. Image Process., **15**(2), pp. 430–444.
[38] Wang, Z., Simoncelli, E. P., and Bovik, A. C., 2003, "Multiscale Structural Similarity for Image Quality Assessment," The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, Vol. 2, Pacific Grove, CA, Nov. 9–12, pp. 1398–1402.
[39] Thum, C., 1984, "Measurement of the Entropy of an Image With Application to Image Focusing," Optica Acta: Int. J. Opt., **31**(2), pp. 203–211.
[40] Shahapure, K. R., and Nicholas, C., 2020, "Cluster Quality Analysis Using Silhouette Score," IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, Oct. 6–9, pp. 747–748.
[41] VanOsdol, J., Ektate, K., Ramasamy, S., Maples, D., Collins, W., Malayer, J., and Ranjan, A., 2017, "Sequential HIFU Heating and Nanobubble Encapsulation Provide Efficient Drug Penetration From Stealth and Temperature Sensitive Liposomes in Colon Cancer," J. Controlled Release, **247**, pp. 55–63.