*BioMedInformatics*

MDPI

*Article*

# Generation of Musculoskeletal Ultrasound Images with Diffusion Models

**Sofoklis Katakis [1,*], Nikolaos Barotsis [2], Alexandros Kakotaritis [1], Panagiotis Tsiganos [3], George Economou [1], Elias Panagiotopoulos [4] and George Panayiotakis [2]**

[1] Electronics Laboratory, Department of Physics, University of Patras, 26504 Patras, Greece; kakotaritis.alexandros@gmail.com (A.K.); economou@physics.upatras.gr (G.E.)
[2] Department of Medical Physics, School of Medicine, University of Patras, 26504 Patras, Greece; nbarotsis@upatras.gr (N.B.); panayiot@upatras.gr (G.P.)
[3] Clinical Radiology Laboratory, School of Medicine, University of Patras, 26504 Patras, Greece; tsiganos@upatras.gr
[4] Orthopaedic and Rehabilitation Department, Patras University Hospital, 26504 Patras, Greece; ecpanagi@med.upatras.gr
* Correspondence: kat.sofoklis@gmail.com

**Abstract:** The recent advances in deep learning have revolutionised computer-aided diagnosis in medical imaging. However, deep learning approaches to unveil their full potential require significant amounts of data, which can be a challenging task in some scientific fields, such as musculoskeletal ultrasound imaging, in which data privacy and security reasons can lead to important limitations in the acquisition and the distribution process of patients' data. For this reason, different generative methods have been introduced to significantly reduce the required amount of real data by generating synthetic images, almost indistinguishable from the real ones. In this study, the power of the diffusion models is incorporated for the generation of realistic data from a small set of musculoskeletal ultrasound images in four different muscles. Afterwards, the similarity of the generated and real images is assessed with different types of qualitative and quantitative metrics that correspond well with human judgement. In particular, the histograms of pixel intensities of the two sets of images have demonstrated that the two distributions are statistically similar. Additionally, the well-established LPIPS, SSIM, FID, and PSNR metrics have been used to quantify the similarity of these sets of images. The two sets of images have achieved extremely high similarity scores in all these metrics. Subsequently, high-level features are extracted from the two types of images and visualized in a two-dimensional space for inspection of their structure and to identify patterns. From this representation, the two sets of images are hard to distinguish. Finally, we perform a series of experiments to assess the impact of the generated data for training a highly efficient Attention-UNet for the important clinical application of muscle thickness measurement. Our results depict that the synthetic data play a significant role in the model's final performance and can lead to the improvement of the deep learning systems in musculoskeletal ultrasound.

**Keywords:** diffusion models; musculoskeletal ultrasound; muscle thickness; synthetic data



**Citation:** Katakis, S.; Barotsis, N.; Kakotaritis, A.; Tsiganos, P.; Economou, G.; Panagiotopoulos, E.; Panayiotakis, G. Generation of Musculoskeletal Ultrasound Images with Diffusion Models. *BioMedInformatics* **2023**, *3*, 405–421. https://doi.org/10.3390/biomedinformatics3020027

Academic Editors: Federico Mastroleo, Angela Ammirabile and Giulia Marvaso

Received: 6 April 2023
Revised: 5 May 2023
Accepted: 6 May 2023
Published: 23 May 2023



## 1. Introduction

Generative models have attracted much attention in the recent literature for their ability to generate realistic data [1–3]; this is due to the nature of deep learning models in which the final performance is related to the amount of data you possess and the inner characteristics. The generative models provide the ability to enrich your dataset without the limitations that occur in real-world data campaigns. These limitations are particularly severe in the medical imaging domain. Data privacy and security can lead to significant patient data acquisition and distribution constraints. Therefore, obtaining

realistic data can be crucial for easily improving the computer-aided diagnosis (CAD) systems of musculoskeletal ultrasound (MSK-US) that are heavily based on deep learning.

Generative adversarial networks (GANs) [4] have been proposed for the generation of realistic synthetic images. Briefly, GANs are a class of deep learning models that can generate new data similar to the data on which they were trained. GANs normally consist of two deep neural networks: a generator and a discriminator network. The generator network takes random noise as input and generates a new sample of data similar to the training data. The discriminator network takes the generated and real training data as input and attempts to classify which is which. The generator network is trained to improve its ability to fool the discriminator network, while the discriminator network is trained to distinguish between the generated and real data correctly.

Several applications of GANs have been studied in ultrasound imaging. In particular, in [5], different GAN architectures were investigated to generate realistic breast ultrasound (US) images. Afterwards, the generated images were used to train convolutional neural networks (CNNs) to classify breast ultrasound images into three categories. Their results indicated that the generated images helped to outperform the baseline model. Furthermore, at [6], they used a GAN architecture to produce synthetic B-mode US images of bone data and their corresponding segmented bone surface masks in real time. Ref. [7] presents a pipeline for generating medical thyroid ultrasound images with an auto-encoding generative adversarial network as a data augmentation method for performance improvement. Similarly, at [8], a novel GAN architecture named Pix2Pix [9] is employed for data augmentation in bone surface segmentation in ultrasound images. Finally, another similar study is [10] in which the authors presented SpeckleGAN, a generative adversarial network with a speckle layer that can be incorporated into a neural network to add realistic and domain-dependent speckle.

Subsequently, at [11], a pipeline for generating synthetic 2D echocardiography images is presented using the Cycle-GAN [12]. Furthermore, at [13], a pipeline can synthesise realistic B-mode US images with customised texture editing features. Secondly, they enhance the structural details of generated images by introducing auxiliary sketch guidance into a conditional GAN. Finally, a study that is similar to ours is [14]. This study used Cycle-GAN to generate realistic B-mode musculoskeletal ultrasound images of longitudinal images of the gastrocnemius medialis muscle. The Cycle-GAN was fed with 100 images and a set of 100 synthetic segmented masks that featured two aponeuroses and a random number of fascicles. Their model output was a set of synthetic ultrasound images and an automated segmentation of each real input image. As a second step, they used existing software to measure muscle thickness, fascicle length, and pennation angle from the real and the generated images. The downside of their study is that they did not train a deep learning model using synthetically generated images to detect muscle architecture, so they have not evaluated how the generated images will affect the performance of such a model.

A more contemporary deep learning method that has presented exceptional results in generating synthetic images in many different applications is the denoising diffusion probabilistic model (DDPM), or simpler diffusion models [15–18]. The basic idea of diffusion models is to start with a random noise vector and then gradually transform it to produce a sample of synthetic data. The above is conducted by applying a sequence of invertible transformations to the noise vector over a series of discrete time steps. The noise vector is updated in each time step by adding a random perturbation, which helps introduce stochasticity into the model. Once the diffusion process is complete, the resulting noise vector is transformed back into a sample of synthetic data using a decoder network. Finally, the decoder network is trained to map the noise vector back to the data space, utilising a loss function that encourages the generated data to be as similar as possible to the real data. Diffusion models have several advantages over other generative models, such as GANs. They are more stable during training and do not suffer from the mode collapse problem common with GANs. They can also generate high-quality images with fine details and realistic textures.

Diffusion models have been applied in various medical imaging applications [19–21]. In [22], the authors propose a transformer-based UNet architecture to model the interaction between noise and semantic features. Furthermore, in [23], a conditional latent DDPM for medical images is proposed in different medical imaging datasets. In addition, at [24], a model which combines a synthetic diffusion-based label generator with a semantic image generator is presented and evaluated at brain magnetic resonance images. Another study worth mentioning is [25], in which the authors achieved image quality superior to the current state-of-the-art generative models in their synthetic data. They performed conditional and unconditional image synthesis and evaluated the quality of their synthetic data on different quantitative metrics.

In this study, the DDPMs are incorporated for the first time in musculoskeletal ultrasound imaging to generate realistic muscle images. We evaluate the similarity of the real and the generated images in different scenarios. Initially, qualitative and quantitative metrics that correspond well with human judgement are used to assess the proximity of the two data types. Later, Attention-UNet [26] is incorporated for the important clinical application of the muscle thickness measurement [27]. In particular, similar to [28], deep learning models are trained in various configurations to delineate the superficial and deep aponeuroses of the examined muscle. Afterwards, the muscle thickness is calculated by taking the average distance between the two aponeuroses at different muscle points.

This study aims to introduce, for the first time, the diffusion models in MSK-US imaging to generate high-quality synthetic images. Afterwards, these synthetic images would be used for training deep learning architectures in extracting muscle thickness in a novel MSK-US database. Therefore, the main contribution of this study is to present a complete methodology for reducing the amount of real data needed to be collected for achieving superior performance in the automation of clinical measurements relevant to the musculoskeletal system.

## 2. Materials and Methods

### *2.1. Database*

As mentioned earlier, the DDPM are trained with a small number of real images to model the data distribution's inner characteristics. Furthermore, the main aim of this study is to generate high-quality synthetic MSK-US images. For this reason, a relatively new MSK-US database was evaluated in this study. The database consists of ultrasound recordings of four superficial human muscles of 116 young and healthy volunteers (49 males and 67 females with a mean age of 25.33 $\pm$ 4.92 y). All the ultrasound recordings were acquired in the Rehabilitation Department of the University Hospital of Patras using a Logiq P9 system (GE Healthcare GmbH, Freiburg, Germany) and an ML6-15 linear array transducer operating at 10-MHz. The same examination protocol remained constant for all participants. In particular:

- Ultrasound scans were conducted longitudinally on the tibialis anterior (T.A.) muscle, at one-quarter of the distance from the inferior pole of the patella to the malleolus lateralis.
- Ultrasound scans were conducted longitudinally on the rectus femoris (R.F.) muscle, halfway along the line from the anterior–superior iliac spine to the superior pole of the patella.
- Ultrasound scans were conducted longitudinally on the bulkiest part of the medial head of the gastrocnemius (GCM) muscle.
- Ultrasound scans were conducted longitudinally on the anterior arm muscles (B.B.) at two-thirds of the distance from the acromion to the elbow crease. This section of the scan included the biceps brachii and brachialis anterior muscles.

To prevent any changes to the image properties caused by software processing, all image optimization modes except for harmonic tissue imaging were turned off. The dynamic range was set at 66 dB and the gain to 50 during the examination of all subjects. The imaging depth was set at 4 cm for most muscles, except for the rectus femoris, where it was set at 6 cm. For patients with large muscles, the depth was increased to include

the entire muscle in the image. Up to six focal zones were evenly distributed along the depth of the image. To ensure optimal ultrasound beam penetration and prevent soft tissue deformation due to transducer pressure, a sufficient amount of CLEAR ECO Supergel ultrasound gel was used. The beam inclination of the transducer was adjusted to obtain the brightest echo from the muscle fascia, ensuring that the images were obtained uniformly and consistently.

A total of 1223 ultrasound images of 4 different muscles were analysed. In particular, the images of the tibialis anterior were 306, the images of the rectus femoris were 299, the images of the gastrocnemius medialis were 299, and the images of the biceps brachii were 308. It must be mentioned that parts of this dataset have been previously presented in [28–31], but these studies' objectives differed from the current one. It is the first time that this database has been used for the task of synthetic image generation. In Table 1, the demographics of the dataset are depicted.

**Table 1.** Demographics of the dataset.

| **Subjects** | **116** |
|---|---|
| Examinations | 155 |
| Age (years) | $25.33 \pm 4.92$ |
| Sex (M/F) | 49/67 |
| Weight (kg) | $68.65 \pm 12.32$ |
| Height (cm) | $172.62 \pm 9.37$ |

### *2.2. Denoising Diffusion Probalistic Models (DDPM)*

#### 2.2.1. Introduction

Diffusion models are generative models that have been inspired by non-equilibrium thermodynamics. As mentioned earlier, the basic idea of diffusion models is to start with a random noise vector and then gradually transform it to produce a sample of synthetic data. First, they define a latent variable model which maps to a latent space of high dimensionality (same as the original data) using a fixed Markov chain [32]. The process of diffusion, which involves introducing random noise to data, is carried out in a series of steps. Subsequently, the system learns to reverse this diffusion process, allowing it to generate desired data samples from the added noise. Hence, the mathematical definition of the forward diffusion process can be described below.

Given a data point sampled from a real data distribution $x_0 \sim q(x)$ a small addition of Gaussian noise is added in $T$ steps producing a sequence of noisy samples $x_1, x_2, \ldots, x_T$. Notably, the variance of Gaussian noise added in each time step is controlled by the following variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$ where $\beta_t$ is a value between 0 and 1 at time step $t$. The addition of Gaussian noise creates a new latent variable $x_t$ that follows the distribution of Equation (1):

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

Since the distribution (1) is normal, the input data $x_0$ can transform to $x_T$ in a tractable way which is defined by the posterior probability (2):

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2}$$

As the time step $t$ becomes larger, the data sample $x_0$ loses its distinguishable features with the final in $t \to \infty$ result to be an isotropic Gaussian distribution. As suitable property of the above process and by using the reparameterisation trick [33], we can sample $x_T$ at any arbitrary time step $t$. In Figure 1 the forward process of the diffusion model is depicted.

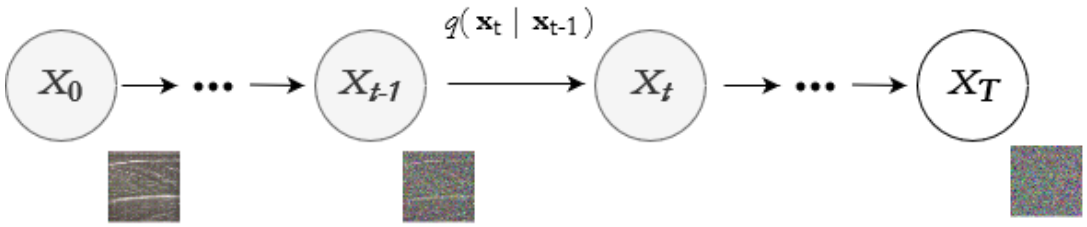



**Figure 1.** Forward process of the diffusion models.

Since the final distribution is an isotropic Gaussian, the next step is to manage to learn the reverse distribution $q(x_{t-1}|x_t)$; this is important because approximating this distribution will enable us to sample $x_t$ from $N(0, \mathrm{I})$ run the reverse process and acquire a sample from $q(x_0)$, generating a novel data point from the original data distribution. The way that we approximate the $q(x_{t-1}|x_t)$ is by a parametrised model $p_\theta$ that parameterises the mean and variance. Since the only requirement for that model is that its input and output dimensionality are identical, diffusion models are commonly implemented with U-Net-like architectures [34]. Finally, the Markov formulation asserts that a given reverse diffusion transition distribution depends only on the previous time step, described by Equation (3).

$$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (3)$$

In Figure 2, the forward and the reverse process is depicted:

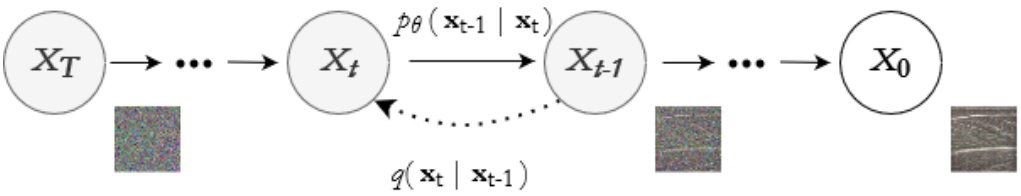



**Figure 2.** Forward and reverse process of the diffusion models.

2.2.2. Training Strategy

We trained a separate diffusion model for each muscle to generate synthetic images. Regarding the training strategy of these models, a manually curated set of approximately 300 longitudinal images for each muscle was used. As we mentioned earlier, these images were acquired from healthy individuals on the basis that the architectural characteristics of their muscles are optimal, meaning the superficial and deep aponeuroses, as well as muscle fascicles, were present and visible. The input and output size of the diffusion models was chosen to be 256 × 256 pixels. This decision was made mainly for the three following reasons:

1. The qualitative results in this image size were better than the smaller sizes (e.g., 128 × 128 and 64 × 64).
2. The input of the Attention-UNet that will delineate the deep and superficial aponeuroses, as described in the following section, is 256 × 256.
3. Larger image sizes would have required excessive computational power and training time.

The deep learning architecture utilized for the diffusion model was a modified version of the U-Net. It incorporated residual blocks (as opposed to traditional convolutional blocks) and utilized group normalization and Sigmoid Linear Unit activation functions. An extended version of the 2D Convolution layer was employed to standardize weights before the convolution step. Attention mechanisms were also employed to selectively weigh and combine different feature maps, enhancing the importance of relevant features while suppressing the influence of irrelevant or noisy features. During training, the L1 loss function was employed, and the number of timesteps that provided the best fit was determined to be 300. The batch size was set to 4, and the model was trained for 300 epochs.

### *2.3. Muscle Thickness Extraction*

#### 2.3.1. Muscle Thickness Measurement

For the muscle thickness (MT) extraction, the pipeline described in [28] was followed. This measurement involves drawing a centre line that lies midway between the superficial and deep aponeuroses. Subsequently, in five evenly distributed points along the centreline, a perpendicular chord is plotted, and the length of this chord is calculated. The muscle thickness is then measured by averaging the distances for all perpendicular chords, as is depicted in Figure 3. Finally, the measurement obtained in pixels is converted into millimetres using a scale factor obtained from the DICOM metadata of the recordings. Although this is not the standard procedure for measuring the MT, it is more robust to user dependence and easier to standardise since it eliminates the variability along the longitudinal axis.

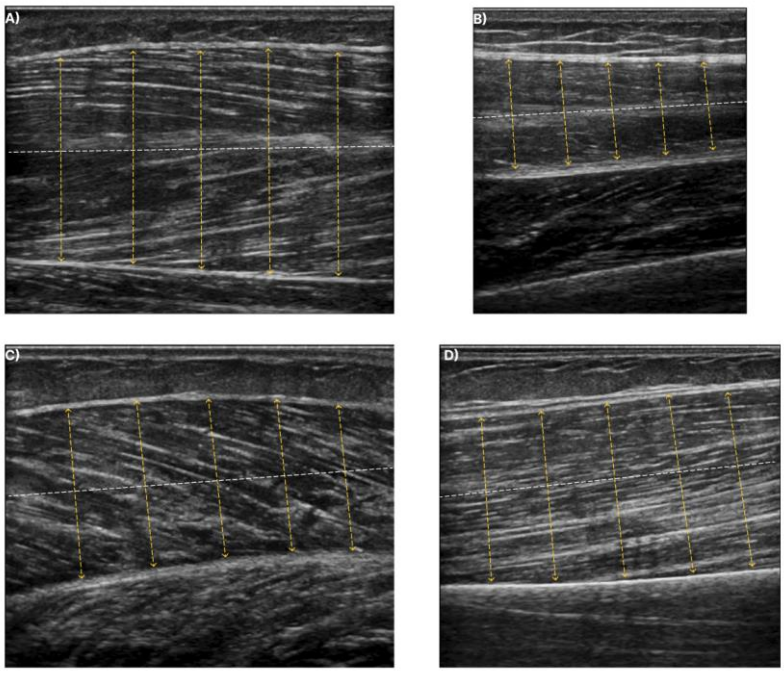



**Figure 3.** The dashed white line depicts the centreline at each muscle. The yellow lines illustrate the perpendicular chords to the centreline. Muscle thickness is measured from the average distance of the yellow dashed lines. (**A**) exhibits the T.A., (**B**) exhibits the R.F., (**C**) exhibits the GCM, and finally, (**D**) shows the B.B.

#### 2.3.2. Aponeurosis Delineation

For the MT measurement, it is crucial to delineate the deep and superficial aponeurosis in each muscle correctly. For doing so, the state-of-the-art Attention-UNet [26] has been selected. The Attention-UNet is a modified version of the original UNet architecture, which incorporates attention gates to enhance the importance of relevant features in the skip connections. The authors claim that these attention gates can filter out responses that are irrelevant in the forward and backward passes of the training process. Especially, during the backward pass, gradients arising from background regions are reduced in weight. This enables the update of the model parameters in shallower layers to be based on the spatial areas that are relevant to the task at hand.

For training a deep learning model, annotated data pairs are required. Figure 4 depicts ultrasound images of the examined muscles along with their annotation. It must be mentioned that the annotation of the aponeuroses was performed with the guidelines of a specialised doctor. For improving the final performance, image augmentation techniques were incorporated. Rotation, scaling, vertical flipping, and random erasing were used. Regarding the optimization process, the weighted dice loss and an ADAM optimizer [35]

were utilized along with a stepwise decrease in the learning rate as the learning rate policy. The input dimension of the deep learning models was 256 × 256 and lastly, the batch size was equal to 8 with 300 as number of epochs.

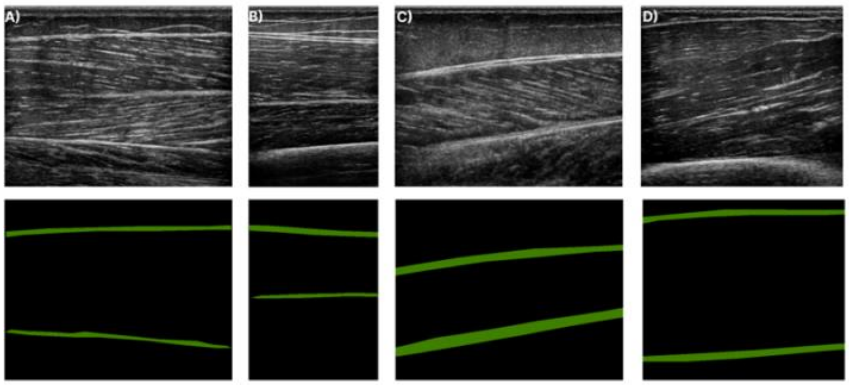



**Figure 4.** Ultrasound images for each muscle along with their corresponding annotation. The green structures in the annotation masks (second row) depict the deep and superficial aponeuroses. (**A**) shows T.A. muscle, (**B**) the R.F. muscle, (**C**) the GCM muscle, and finally (**D**) the B.B. muscle.

#### 2.3.3. Evaluation Protocol

The dataset was split into three folds (e.g., train, validation, and test set) for better evaluation. It is important to have an independent validation set because it can be used to monitor the model's performance during training and to detect overfitting. Furthermore, having a test set is another reason to avoid the final model being biased by the training or validation data. In our case, the train set consisted of 40% for each muscle examination, 20% for the validation set and the rest 40% for the test set. Furthermore, 200 synthetic images were generated and annotated for each of the four muscles. The annotation procedure is presented in Figure 4. Later, a series of experiments were performed to evaluate the impact of the generated data on the model performance. The basic intuition behind these experiments is to demonstrate that the synthetic data can be used autonomously or auxiliary for training high-performance deep learning models. In particular:

1. A model was trained with only real data in the protocol mentioned earlier (40% train set, 20% validation set, 40% test set) and recorded its performance in the test set (Real Model).
2. A model was trained with only generated data and recorded its performance in the test set of real images (Gen Model).
3. A model was trained with the real and all the generated data. Specifically, the generated data were added to the real training set, and the validation and test sets were kept the same. The final performance at the test set was recorded (Real + Gen Model).
4. Finally, a supplementary analysis was performed in which the number of the real training images was intentionally reduced in the dataset while keeping the size of the testing set constant. Specifically, the entire dataset was divided into training and testing sets and three separate experiments were conducted. At each experiment, the number of real training images was reduced and the best performance of an Attention-UNet with and without synthetic data was reported.

Regarding the evaluation metrics for the deep and superficial aponeuroses delineation tasks, five well-established indexes were incorporated [36]. Specifically, the precision and recall of the segmentation results were reported between the manual and automatic measurements. Furthermore, the Dice coefficient (DSC) and the intersection over union (IoU) were also employed. Both metrics measure the pixels overlapping between the prediction of the networks with the ground truth masks. For measuring the discrepancy in the muscle thickness measurement, the root mean square error (RMSE) between the

manual and automated readings was calculated. Finally, for the assessment of possible bias and systematic error between the two readings, the Bland–Altman analysis was also used.

## 3. Results

### *3.1. Qualitative Analysis*

Figure 5 demonstrates synthetic images that the diffusion models generated for each muscle in comparison with real images that have been trained. From a qualitative aspect, the results are exceptional since it is difficult for the human eye to differentiate them. Furthermore, it is observable that these images consist of the basic characteristics of a typical longitudinal ultrasound recording. In particular, the muscle aponeuroses formed due to the high reflectivity of the epimysium surrounding the muscle have the exact form and properties as the real recordings. Furthermore, the muscle fascicles in the synthetic images are organised in a linear, pinnate, or triangular fashion, similar to the real ultrasound images. Overall, the generated images seem consistent and realistic from a medical and visual standpoint.

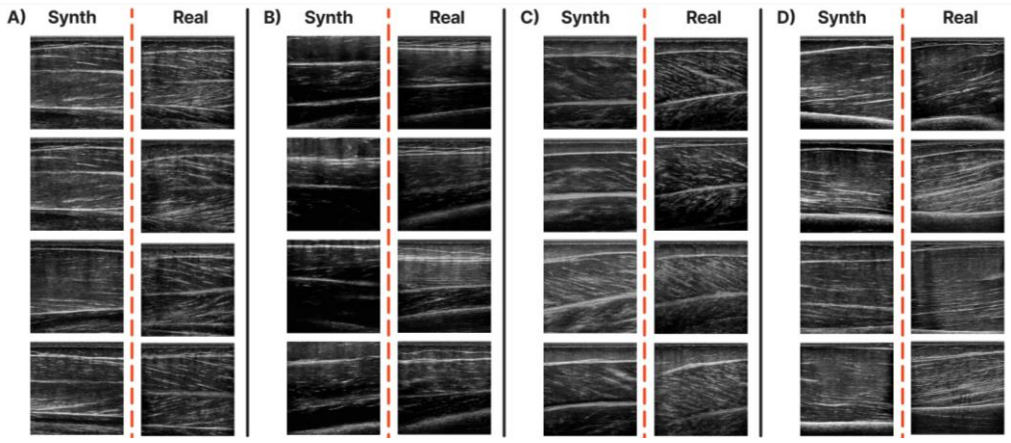



**Figure 5.** Synthetic vs. real ultrasound images from the diffusion model. (**A**) Samples from the T.A. muscle, (**B**) from the R.F. muscle, (**C**) the GCM muscle and (**D**) the B.B. muscle.

The next analysis aims to compare the pixel-level differences between the generated and the real images. Specifically, the pixel intensity distributions of 100 real and synthetic images were extracted to quantify each case's information content. Later, these distributions' shape and entropy values were compared. Finally, in Figure 6, the histograms of the real and synthetic images (all resized in 256 $\times$ 256) for each muscle are presented.

It is clear from the histograms that the distribution shapes between the two sets of images are similar (statistical similarity). More specifically, every muscle has a skewed distribution with close mean skewness and entropy between the synthetic and real image types. The biggest difference in the mean skewness is reported in the T.A. (real: 0.95, synthetic: 1.31) and similarly in the R.F. (real: 1.32, synthetic: 1.67), which can be explained by the fact that the real images are darker than the synthetic as we see in the corresponding histograms (more pixel's intensities near zero). Regarding the mean entropy values, the results are extremely close in all the muscles, which is depicted in their range of values since the synthetic (10.83–11.03) and the real images (10.64–10.92) are alike, only with a slight offset in the synthetic. Furthermore, the generated images displayed similar variation between samples, as reflected by the different $y$-axis values and a similar range of peaks on the $x$-axis.

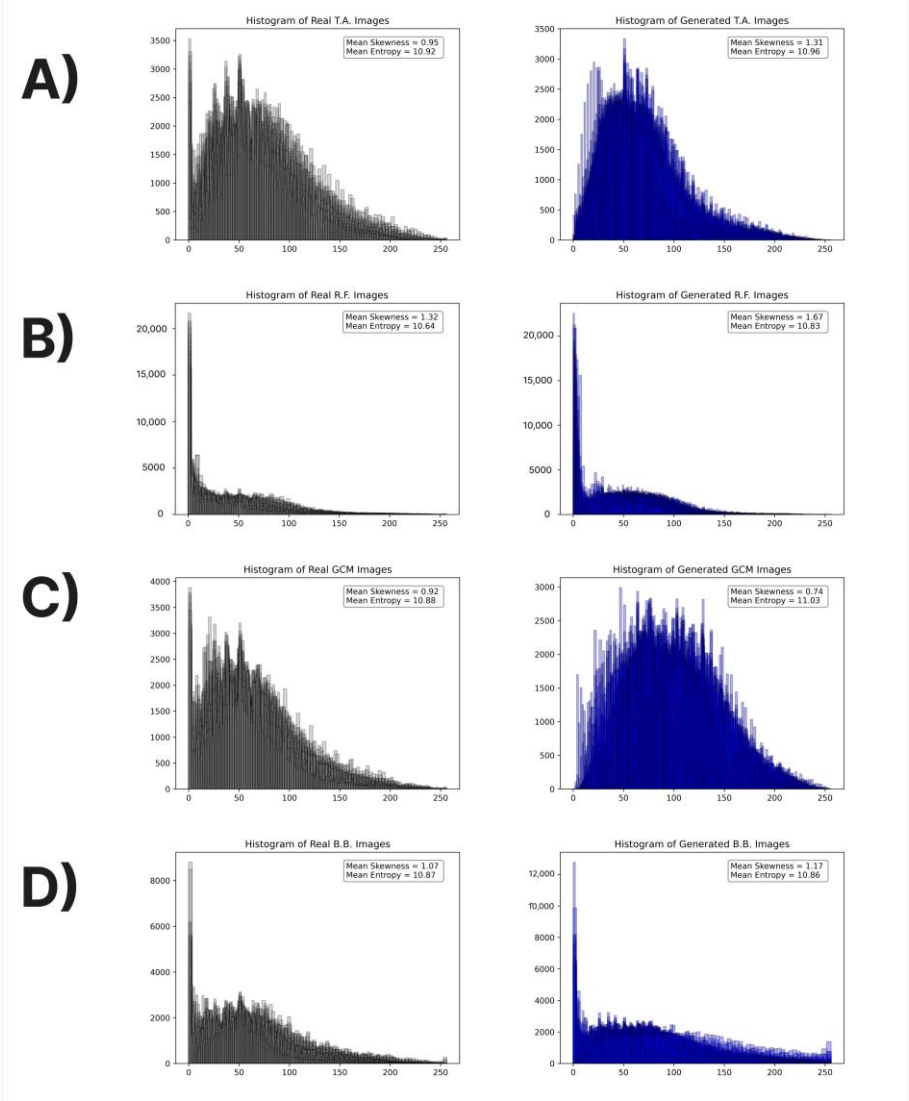



**Figure 6.** Histograms for each muscle for real and generated data for the (**A**) T.A. muscle, (**B**) R.F. muscle, (**C**) GCM muscle, and (**D**) B.B. muscle.

Another useful analysis is presented in Table 2. This table depicts four qualitative metrics between the two sets of images for each muscle. These metrics will help us quantify the similarity in the information and textural content. In particular, PSNR [37] (peak signal-to-noise ratio) has been used to measure the similarity between the two sets of images. However, since PSNR does not always correlate well with human perception of image quality, the structural similarity (SSIM) [37] metric is also incorporated in this analysis. SSIM considers their luminance, contrast, and structure by comparing local windows of pixels. Furthermore, another metric used to assess image similarity is the learned perceptual image patch similarity (LPIPS) [38]. LPIPS has the benefit that it is based on a learned model, which means it has been trained on large datasets of human judgments of image similarity; this allows LPIPS to capture the nuances of human perception more easily than the other metrics. Finally, the last quality assessment metric is the freshet inception distance (FID) [39]. The FID score measures the distance between the distributions of real and generated images, with lower scores indicating higher similarity. First, FID is calculated using a pretrained convolutional neural network to extract feature representations from the generated and real images. These features are then used to compute the mean and covariance of the feature distributions for both sets of images. The FID score is then calculated as the squared Euclidean distance between these two feature distributions.

**Table 2.** Qualitative metrics between the generated and real images for each muscle.

| | **PSNR** ↑ | **SSIM** ↑ | **LPIPS** ↓ | **FID** ↓ |
|---|---|---|---|---|
| T.A. | 61.928 | 0.996 | 0.006 | 3.395 |
| R.F. | 61.486 | 0.994 | 0.005 | 1.958 |
| GCM | 60.544 | 0.992 | 0.005 | 2.681 |
| B.B. | 60.462 | 0.993 | 0.008 | 3.819 |

From the results in Table 2, the PNSR between the two image types of all the muscles are clearly shown to be above 60 and the SSIM close to 1, indicating an increased similarity of the generated images to the real. Furthermore, the LPIPS scores are close to zero. Finally, the FID scores have small values, another indicator of the similarity in the textural content between the two types of images. Additionally, the small values of LPIPS demonstrate the high quality of the generated images with no severe artefacts. Regarding a per muscle analysis, T.A. and R.F. exhibit the best result regarding PSNR and SSIM and the R.F. and GCM regarding the LPIPS metric. However, the differences between the results are so small that they are not statistically important.

As a supplementary analysis, one more experiment has been conducted. In this experiment the average SSIM over real images from various patients (inter-patient SSIM) was calculated and compared with the average SSIM over the synthetic images (synthetic SSIM). This experiment was designed to provide insight into the level of similarity that should be expected between the same types of images. Specifically, in each image set the total images were divided in half and the SSIM index was calculated. The results of this analysis are presented in Table 3.

**Table 3.** Results of the inter-patient SSIM and synthetic SSIM.

| | **Inter-Patient SSIM** ↑ | **Synthetic SSIM** ↑ |
|---|---|---|
| T.A. | 0.995 | 0.995 |
| R.F. | 0.996 | 0.994 |
| GCM | 0.996 | 0.992 |
| B.B. | 0.995 | 0.992 |

Our results indicate that the inter-patient SSIM index is almost identical to the SSIM index of synthetic images. This finding provides further evidence that the distribution of synthetic data possesses similar textural and informational characteristics to the distribution of real images. We believe that this result supports the validity and utility of our proposed method for generating realistic musculoskeletal ultrasound images and reinforces the potential of synthetic data to supplement real data.

Continuing our analysis, visualisation of the generated and real data in a 2D space was performed. Specifically, features were extracted for each sample from the bottleneck of a pre-trained Attention-UNet. This Attention-UNet had been trained to segment all four muscles' deep and superficial aponeuroses. Afterwards, these features were normalised, and later dimensionality reduction was performed with the well-established principal component analysis (PCA). Finally, all the samples were visualised in a two-dimensional (2D) feature space for inspecting their structure. Figure 7 depicts the 2D feature space in each muscle.

Figure 7 shows that the features of the real and synthetic images in this 2D space are not easily separable in any of the examined muscles, which means that the data points share similar characteristics and are likely to belong to the same class or category. Finally, it must be noted that this result is another indicator that the textural representation of the synthetic and realistic images is similar.

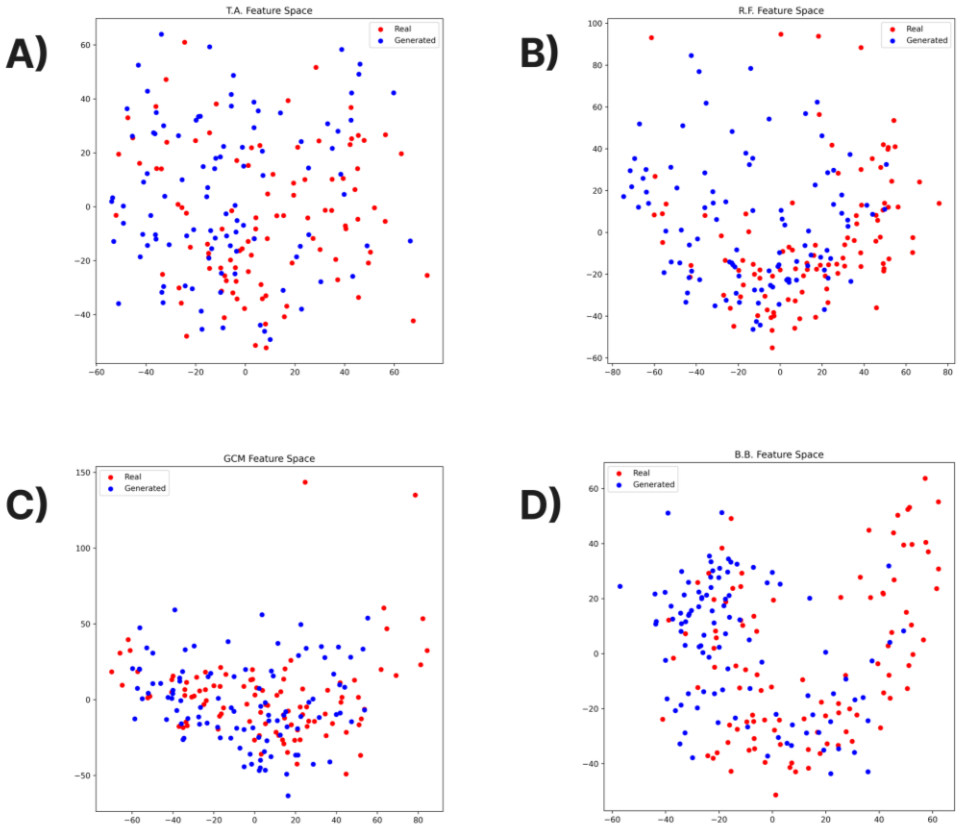



**Figure 7.** Two-dimensional visualisation for each muscle for real and synthetic data. (**A**) T.A. muscle, (**B**) R.F. muscle, (**C**) GCM muscle, and (**D**) B.B. muscle.

### *3.2. Muscle Thickness Analysis*

The image segmentation results for the experiments described in Section 2.3.3 are presented in this section. As mentioned earlier, the average performance in the test set of the real images is reported for each experiment. The results of the deep and superficial aponeuroses delineation are presented in Table 4. The model trained with the real and generated data (Real + Gen Model) outperforms the other two models in most metrics; this is important because it shows that the generated data positively impacts the final performance. Furthermore, it must be mentioned that the performance of the model trained with only synthetic data (Gen Model) is far from disappointing. Specifically, the precision reported is 0.78, while the corresponding recall is 0.85, indicating the network's ability to accurately locate both the deep and superficial aponeuroses. Additionally, the reported results for DSC and IoU are 0.80 and 0.68, respectively, further demonstrating the strong performance of the segmented masks on the test set. These two combined results prove that the generated data have a positive role in training deep learning models.

**Table 4.** Overall segmentation results for each experiment. The results are reported in mean ± std.

| | **Precision** | **Recall** | **DSC** | **IoU** |
|---|---|---|---|---|
| Real Model | 0.85 ± 0.10 | 0.87 ± 0.10 | 0.85 ± 0.07 | 0.75 ± 0.10 |
| Gen Model | 0.78 ± 0.14 | 0.85 ± 0.12 | 0.80 ± 0.10 | 0.68 ± 0.13 |
| Real + Gen Model | 0.84 ± 0.10 | 0.88 ± 0.09 | 0.86 ± 0.08 | 0.76 ± 0.10 |

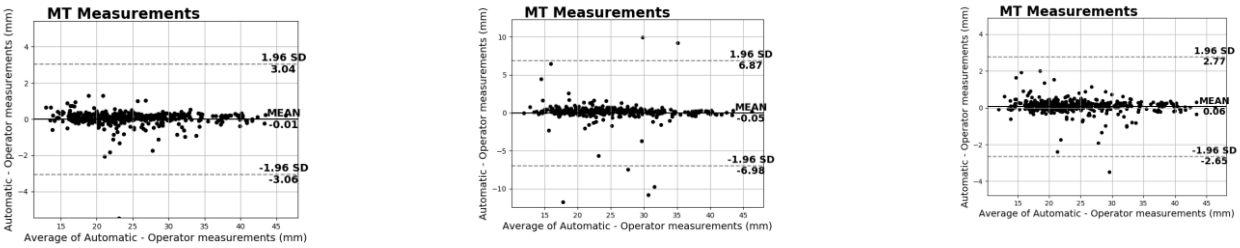
Continuing our analysis, Table 5 presents a comparison between the automated measurements obtained from the aforementioned models and the manual measurements in physical units. The table displays the mean ± standard deviation of the measurements, along with their RMSE discrepancy. All the models exhibit an extremely low RMSE in the evaluated dataset, one more indicator of the applicability of the synthetic data for this

task. Additionally, it must be highlighted that even though the Gen Model has slightly underperformed in comparison with the other two, the average discrepancy between the two readings was equal to only 1.05 mm. This provides additional evidence that synthetic data can be independently employed for training deep learning models that exhibit high performance. Finally, the other two models have achieved similar results, making it challenging to determine definitively which model is superior.

**Table 5.** Comparison results between the manual and automatic muscle MT measurements.

| | Manual (mm) | Automatic (mm) | RMSE (mm) |
|---|---|---|---|
| Real Model | 24.50 ± 6.49 | 24.51 ± 6.45 | 0.35 ± 1.52 |
| Gen Model | 24.50 ± 6.49 | 24.56 ± 6.98 | 1.05 ± 3.38 |
| Real + Gen Model | 24.50 ± 6.49 | 24.44 ± 6.51 | 0.38 ± 1.33 |

Figure 8 presents another informative analysis, namely the Bland–Altman plot of the muscle thickness measurements. All the plots indicate minimal additive bias and no evident systematic error. Moreover, a majority of the differences fall within the 95% limits of agreement, and no distinguishable patterns are discernible in the plots. Finally, it is observable that the Bland–Altman plot of the Gen model (Figure 8b) has a few points that are far from the mean values highlighting that the results in those cases are failing. Instead in the other two plots (Figure 8a) and (Figure 8c) such behaviour is not present at that extent.




(**a**) (**b**) (**c**)

**Figure 8.** Bland−Altman analysis of the manual vs. automatic MT measurements of the (**a**) Real Model, (**b**) Gen Model, and (**c**) Real + Gen Model.

Additionally, in Figure 9, a sample of the predictions of the Gen Model is presented. From the segmented masks, it is clear that the synthetic images can be applied effectively for the delineation of the deep and superficial aponeuroses.

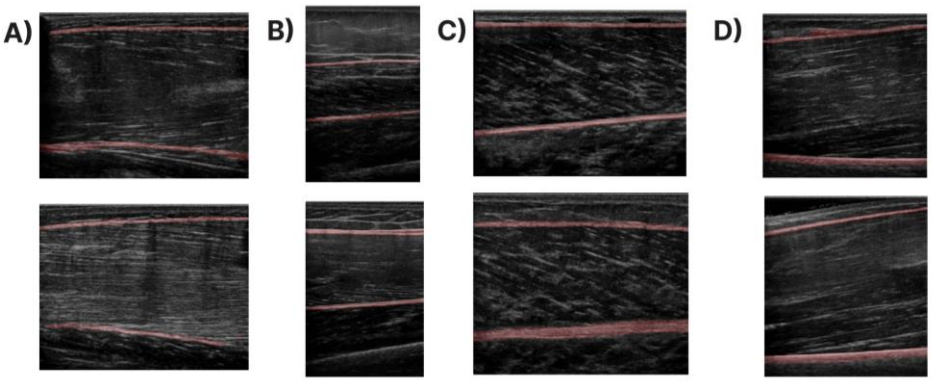



**Figure 9.** Qualitative results of the Gen Model in images of the four examined muscles. (**A**) T.A. muscle, (**B**) R.F. muscle, (**C**) GCM muscle, and (**D**) B.B. muscle.

Finally, an additional analysis was conducted to better demonstrate the impact of the generated data in training high-performance deep learning models. Specifically, the entire dataset was divided into training and validation sets and three separate experiments were performed. In each experiment, the number of training images in the dataset was intentionally reduced, while keeping the size of the validation set constant. In the first experiment, 50% of the whole dataset was considered as a training set and the rest 50% as a validation set. In the second experiment, the validation set remained constant, but the training set was reduced to 30% of the whole dataset. Finally, in the third experiment, the training set was further reduced to 10% of the whole dataset. For each experiment, two Attention-UNet models were trained: one utilizing the original dataset, and the other using the synthetic data generated by our proposed method added to the training set. Lastly, the best performance achieved by both models in terms of Dice coefficient and IoU was reported. The results of this analysis are presented in Table 6.

**Table 6.** Comparison results of the Attention-UNet trained with different number of training data in the same validation set.

| | **DSC** | **IoU** |
|---|---|---|
| 50% Train | 0.842 | 0.735 |
| 50% Train + Gen | 0.847 | 0.743 |
| 30% Train | 0.834 | 0.724 |
| 30% Train + Gen | 0.842 | 0.735 |
| 10% Train | 0.800 | 0.680 |
| 10% Train + Gen | 0.823 | 0.710 |

Based on the results, it is evident that artificial data can significantly enhance the performance of deep learning models and can be utilized to reduce the real data required to achieve exceptional performance. Notably, the results demonstrate that the model trained with a combination of 30% real images and generated data (30% Train + Gen) yields similar performance to the model trained with 50% real images (50% Train). This indicates that the generated data has significantly boosted the performance of the model in the specific problem. Finally, the experiments revealed that the impact of the generated data was more pronounced when the amount of real training data was limited. This is likely because, when there is less real data available, the generated data can provide a more valuable supplement to the training set. On the other hand, as the amount of real data increases, the deep learning architecture can generalize better leaving less space for improvement.

## 4. Discussion

This study employed state-of-the-art diffusion models to generate realistic MSK-US images of four very informative for investigating neuromuscular disorders [40] muscles. Afterwards, the synthetic image quality was assessed both qualitatively and quantitatively compared to the real data. Specifically, a histogram analysis that demonstrates that the pixel's intensity distribution is similar in both cases has been performed. Additionally, four qualitative metrics that correspond well with human perception were evaluated between the two types of images. In all these metrics, the results exhibited superior performance. In addition to that, features from a pretrained Attention-UNet were extracted and visualised in a two-dimensional space using PCA. Again, the results showed that a clear distinction does not exist between the two projections, another indicator of the similarity of the two sets of images. Finally, for evaluating the applicability of the synthetic data in a real-world scenario, an Attention-UNet was trained to automatically delineate the deep and superficial aponeuroses. Our results indicate that the synthetic data can be used autonomously or supplementarily for training high-performance deep-learning models for this task.

A significant advancement of this study compared to recent works presented in [8,14], is the use of diffusion models instead of GANs to address this problem. Generative

adversarial networks have the downside that they are hard to train because they involve a complex optimisation process that requires careful tuning of hyperparameters. Additionally, GANs may suffer from mode collapse, meaning that the generator network produces a limited set of output samples, ignoring the rest of the distribution, resulting in generated images lacking diversity and quality. Apart from that, GANs are also sensitive to data quality and quantity, requiring a large and high-quality dataset to learn meaningful patterns, or the model may not generate accurate samples. Next, another limitation of the GANs is that the generation mechanism of new samples is difficult to understand and is considered a black-box model. Instead, diffusion models can be trained efficiently without excessive tuning to produce realistic results. Furthermore, unlike other deep generative models, diffusion models have an interpretable structure based on stochastic differential equations, allowing insights into the generative process and the underlying dynamics of the data. Finally, diffusion models can be used for transfer learning by fine-tuning the model on a new dataset which is useful in scenarios where labelled data are scarce or when the model needs to adapt to new domains.

Several analyses have been performed to evaluate the quality of synthetic data properly. Initially, the histograms of the pixel intensities in 100 randomly picked generated, and real images of each muscle were extracted and compared. This analysis showed that each muscle's distribution shape and entropy are statistically similar. In particular, a right-skewed distribution exists in every muscle with close mean skewness and entropy between the two readings. The most significant difference in the mean skewness was reported in T.A. (real: 0.95, synthetic: 1.31) and in the R.F. (real: 1.32, synthetic: 1.67, explained by the fact that the real images were darker than the synthetic in both muscles. Regarding the mean entropy values, the results were extremely close in all the examined muscles. Four metrics aligned with human judgment were used to quantify the similarity of 100 generated images for each muscle with the real dataset. In all the metrics, the results demonstrated that the quality of the synthetic data is superior. PSNR was above 60, and SSIM was close to 1 in all the examined muscles. In addition, the similarity level of the real images was analysed (inter-patient SSIM) and found almost identical to the similarity level of the synthetic images. This finding provides further evidence that the distribution of synthetic data possesses similar textural and informational characteristics to the distribution of real images. Furthermore, LPIPS and FID, which also consider textural information, were close to zero in all the muscles, another indicator of the similarity of the two sets of images. Finally, the two sets of images were visualised in a common two-dimensional space. In particular, high-level textural features were extracted from the bottleneck of a pretrained Attention-UNet for each image. Afterwards, the dimensionality of these features was reduced with PCA and visualised in a common space. The results showed that the data points of the generated and real images are not forming separate classes but are mixed between them, which is one more indicator that possesses similar textural characteristics.

Furthermore, a system that automatically extracts the muscle thickness measurement was developed to evaluate the applicability of the generated data in a real-world clinical application. Specifically, the deep and superficial aponeuroses were segmented with the state-of-the-art Attention-UNet in a novel database of musculoskeletal ultrasound images. Afterwards, the MT is measured by computing the mean distance of the two aponeuroses at several points across the muscle. Since the main goal is to assess the generated data's impact on the model's final performance, different experiments were performed. From these, it is clear that the generated data are capable of producing high-performance models with (or without) the use of real images. This is a significant result that can lead to the acceleration of the integration and the improvement of the deep learning technology in MSK-US, where the acquisition process of real data is very difficult and time-consuming due to privacy restrictions. Notably, the Attention-UNet trained only with synthetic images (Gen Model) achieved over 80% with the Dice coefficient, a performance very close to the Real Model (85%) that has been trained with only real images or to the Real + Gen Model (86%) that has been trained with both types of data. In every case, these results prove that

the generated data can be used autonomously or supplementarily to train high-performance models for the specific task. This also depicted the RMSE difference between the manual and automated measurements in all the different training configurations. Specifically, the average difference between the two readings for the Real Model was only 0.35 mm and for the Real + Gen model was similar at 0.38 mm. Similarly, for the model trained only with generated data, the difference is 1.05 mm, larger than before but still deviates only 4% of the manual MT measurements. Finally, an additional analysis was conducted to better demonstrate the impact of the generated data in training high-performance deep learning models. During this analysis, it was observed that all the models trained with a combination of real and generated data outperformed the baseline models. Additionally, the model that was trained with a combination of 30% of real images and generated data (30% Train + Gen) achieved almost identical performance to the model trained with more real images (50% Train). These findings provide further support for the notion that artificial data can substantially enhance the performance of deep learning models and can serve as a supplement in situations where real data are lacking.

This study has some general limitations. Firstly, the examined muscles were only four from over 200 that the human body possesses. Secondly, the number of MSK-US images was 1223 from 116 subjects, which can be considered a relatively small number. Consequently, conducting further research involving a larger sample size and more muscles would offer a clearer understanding of the diffusion models' capability to generate MSK-US images. Another constraint is that all the actual images were obtained from a single ultrasound machine, utilizing the same software and image settings. Hence, we did not investigate multiple configuration setups that can alternate the final image. Lastly, all the recordings used in this study were acquired from young and healthy subjects, which can bias our results since the young population usually has muscles with normal echogenicity and better architectural characteristics than the elderly. However, we are confident that these challenges can be overcome with a small number of real data since the diffusion model is scalable, as we mentioned before and can be trained without excessive hyperparameter tuning.

In future work, the plan is to investigate the generation of transverse MSK-US images in these four muscles. Furthermore, we will investigate the applicability of the generated data in other clinical applications, such as the automatic extraction of the cross-sectional area (CSA) or even the extraction of the fascicle's length and pennation angle. Finally, in the future, we will investigate the generation of data acquired from older adults with higher echogenicity since ageing leads to a reduction in muscle mass and an increase in muscle fat.

**Author Contributions:** Conceptualisation, S.K. and A.K; methodology, S.K., A.K., G.E. and G.P.; software, S.K. and A.K.; validation, S.K., N.B., E.P. and G.P.; data curation, S.K., N.B., P.T. and A.K.; writing—original draft preparation, S.K. and G.E.; visualisations, S.K. and A.K.; writing—review and editing, S.K., A.K., N.B., P.T., G.P., E.P. and G.E. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the State Scholarships Foundation (ΙΚΥ), grant number MIS-5000432.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Patras University Hospital (protocol code 50/18-1-18).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** This research was co-financed by Greece and the European Union (European Social Fund—ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research—2nd Cycle" (MIS-5000432), implemented by the State Scholarships Foundation (ΙΚΥ).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Kingma, D.P.; Mohamed, S.; Jimenez Rezende, D.; Welling, M. Semi-Supervised Learning with Deep Generative Models. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Curran Associates, Inc.: Red Hook, NY, USA; Volume 27.
2. Oussidi, A.; Elhassouny, A. Deep Generative Models: Survey. In Proceedings of the 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 2–4 September 2018; pp. 1–8.
3. Turhan, C.G.; Bilge, H.S. Recent Trends in Deep Generative Models: A Review. In Proceedings of the 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 20–23 September 2018; pp. 574–579.
4. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
5. Maack, L.; Holstein, L.; Schlaefer, A. GANs for Generation of Synthetic Ultrasound Images from Small Datasets. *Curr. Dir. Biomed. Eng.* **2022**, *8*, 17–20. [CrossRef]
6. Alsinan, A.Z.; Rule, C.; Vives, M.; Patel, V.M.; Hacihaliloglu, I. GAN-Based Realistic Bone Ultrasound Image and Label Synthesis for Improved Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2020, Lima, Peru, 4–8 October 2020; Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 795–804.
7. Liang, J.; Chen, J. Data Augmentation of Thyroid Ultrasound Images Using Generative Adversarial Network. In Proceedings of the 2021 IEEE International Ultrasonics Symposium (IUS), Xi'an, China, 11–16 September 2021; pp. 1–4.
8. Zaman, A.; Park, S.H.; Bang, H.; Park, C.; Park, I.; Joung, S. Generative Approach for Data Augmentation for Deep Learning-Based Bone Surface Segmentation from Ultrasound Images. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 931–941. [CrossRef] [PubMed]
9. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks 2018. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
10. Bargsten, L.; Schlaefer, A. SpeckleGAN: A Generative Adversarial Network with an Adaptive Speckle Layer to Augment Limited Training Data for Ultrasound Image Processing. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 1427–1436. [CrossRef]
11. Gilbert, A.; Marciniak, M.; Rodero, C.; Lamata, P.; Samset, E.; Mcleod, K. Generating Synthetic Labeled Data from Existing Anatomical Models: An Example with Echocardiography Segmentation. *IEEE Trans. Med. Imaging* **2021**, *40*, 2783–2794. [CrossRef] [PubMed]
12. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks 2020. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
13. Liang, J.; Yang, X.; Huang, Y.; Li, H.; He, S.; Hu, X.; Chen, Z.; Xue, W.; Cheng, J.; Ni, D. Sketch Guided and Progressive Growing GAN for Realistic and Editable Ultrasound Image Synthesis. *Med. Image Anal.* **2022**, *79*, 102461. [CrossRef]
14. Cronin, N.J.; Finni, T.; Seynnes, O. Using Deep Learning to Generate Synthetic B-Mode Musculoskeletal Ultrasound Images. *Comput. Methods Progr. Biomed.* **2020**, *196*, 105583. [CrossRef]
15. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 6840–6851.
16. Nichol, A.Q.; Dhariwal, P. Improved Denoising Diffusion Probabilistic Models. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8162–8171.
17. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 11461–11471.
18. Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; Yoon, S. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. *arXiv* **2021**, arXiv:2108.02938.
19. Kazerouni, A.; Aghdam, E.K.; Heidari, M.; Azad, R.; Fayyaz, M.; Hacihaliloglu, I.; Merhof, D. Diffusion Models for Medical Image Analysis: A Comprehensive Survey. *arXiv* **2022**, arXiv:2211.07804.
20. Croitoru, F.-A.; Hondru, V.; Ionescu, R.T.; Shah, M. Diffusion Models in Vision: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, 1–20. [CrossRef]
21. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Shao, Y.; Zhang, W.; Cui, B.; Yang, M.-H. Diffusion Models: A Comprehensive Survey of Methods and Applications. *arXiv* **2022**, arXiv:2209.00796.
22. Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; Xu, Y. MedSegDiff-V2: Diffusion Based Medical Image Segmentation with Transformer. *arXiv* **2023**, arXiv:2301.11798.
23. Müller-Franzes, G.; Niehues, J.M.; Khader, F.; Arasteh, S.T.; Haarburger, C.; Kuhl, C.; Wang, T.; Han, T.; Nebelung, S.; Kather, J.N.; et al. Diffusion Probabilistic Models Beat GANs on Medical Images. *arXiv* **2022**, arXiv:2212.07501.
24. Fernandez, V.; Pinaya, W.H.L.; Borges, P.; Tudosiu, P.-D.; Graham, M.S.; Vercauteren, T.; Cardoso, M.J. Can Segmentation Models Be Trained with Fully Synthetically Generated Data? In *Simulation and Synthesis in Medical Imaging. SASHIMI 2022*; Springer: Cham, Switzerland, 2022.
25. Dhariwal, P.; Nichol, A. Diffusion Models Beat GANs on Image Synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
26. Oktay, O.; Schlemper, J.; Le Folgoc, L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Y Hammerla, N.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.

27. Barotsis, N.; Tsiganos, P.; Kokkalis, Z.; Panayiotakis, G.; Panagiotopoulos, E. Reliability of Muscle Thickness Measurements in Ultrasonography. *Int. J. Rehabil. Res.* **2020**, *43*, 123. [CrossRef] [PubMed]
28. Katakis, S.; Barotsis, N.; Kakotaritis, A.; Economou, G.; Panagiotopoulos, E.; Panayiotakis, G. Automatic Extraction of Muscle Parameters with Attention UNet in Ultrasonography. *Sensors* **2022**, *22*, 5230. [CrossRef]
29. Katakis, S.; Barotsis, N.; Kakotaritis, A.; Tsiganos, P.; Economou, G.; Panagiotopoulos, E.; Panayiotakis, G. Muscle Cross-Sectional Area Segmentation in Transverse Ultrasound Images Using Vision Transformers. *Diagnostics* **2023**, *13*, 217. [CrossRef] [PubMed]
30. Katakis, S.; Barotsis, N.; Kastaniotis, D.; Theoharatos, C.; Tsiganos, P.; Economou, G.; Panagiotopoulos, E.; Fotopoulos, S.; Panayiotakis, G. Muscle Type and Gender Recognition Utilising High-Level Textural Representation in Musculoskeletal Ultrasonography. *Ultrasound Med. Biol.* **2019**, *45*, 1562–1573. [CrossRef]
31. Katakis, S.; Barotsis, N.; Kastaniotis, D.; Theoharatos, C.; Tsourounis, D.; Fotopoulos, S.; Panagiotopoulos, E. Muscle Type Classification on Ultrasound Imaging Using Deep Convolutional Neural Networks. In Proceedings of the 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Zagorochoria, Greece, 10–12 June 2018; pp. 1–5.
32. Gagniuc, P.A. *Markov Chains: From Theory to Implementation and Experimentation*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017.
33. Devroye, L. Random Variate Generation in One Line of Code. In Proceedings of the Winter Simulation Conference, Coronado, CA, USA, 8–11 December 1996; pp. 265–272.
34. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
35. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
36. Bertels, J.; Eelbode, T.; Berman, M.; Vandermeulen, D.; Maes, F.; Bisschops, R.; Blaschko, M. Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Shenzhen, China, 13–17 October 2019; Springer: Cham, Switzerland, 2019; Volume 11765, pp. 92–100.
37. Horé, A.; Ziou, D. Image Quality Metrics: PSNR vs. SSIM. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Los Alamitos, CA, USA, 23–26 August 2010; pp. 2366–2369.
38. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric 2018. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
39. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv* **2018**, arXiv:1706.08500.
40. Barotsis, N.; Galata, A.; Hadjiconstanti, A.; Panayiotakis, G. The Ultrasonographic Measurement of Muscle Thickness in Sarcopenia. A Prediction Study. *Eur. J. Phys. Rehabil. Med.* **2020**, *56*, 427–437. [CrossRef] [PubMed]