Evaluating diffusion model generated synthetic histopathology image data against authentic digital pathology images

T. Rai*[a,b], C. Gola[c], M. Hernández[c], S. Fingerhood[c], J. Marrero[c], P. Diaz-Santana[c], G. Giglia[d], A. Morisi[e], B. Bacci[f], S.A. Thomas[g], L. Ressel[h], N. Bacon[e,i], N. Papachristou[j], A. Cook[e,k], R. La Ragione[e,l], K. Wells[a,b]

[a]Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK; [b]Surrey DataHub, University of Surrey, Guildford, GU2 7AL, UK; [c]Veterinary Pathology Centre, University of Surrey, Guildford, GU27AQ, UK; [d]Department of Veterinary Medicine, University of Perugia, San Costanzo Street 6, 06126 Perugia, Italy; [e]School of Veterinary Medicine, University of Surrey, Guildford, GU2 7AL, UK; [f]Department of Veterinary Medical Sciences, University of Bologna, 40126, Bologna, Italy; [g]National Physical Laboratory, Teddington, TW11 0LW, UK; [h]Department of Veterinary Anatomy Physiology and Pathology, Leahurst Campus, University of Liverpool, Neston CH64 7TE, UK; [i]AURA Veterinary, Guildford, GU2 7AJ, UK; [j]Aristotle University of Thessaloniki, Thessaloniki, Greece; [k]The Veterinary Health Innovation Engine (vHive), University of Surrey, Guildford, GU2 7AL, UK; [l]School of Biosciences, University of Surrey, Guildford, GU2 7XH, UK;

*t.rai@surrey.ac.uk

ABSTRACT

This study explores the efficacy of diffusion probabilistic models for generating synthetic histopathological images, specifically canine Perivascular Wall Tumours (cPWT), to supplement limited datasets for deep learning applications in digital pathology. This research evaluates an open-source medical domain-focused diffusion model called Medfusion, where the model was trained on a small (1,000 patches) and a large dataset (17,000 patches) of cPWT images to compare performance on the different sized datasets. A Receiver Operating Characteristic (ROC) study was implemented to investigate the ability of six veterinary medical professionals and pathologists to discern between generated and real cPWT patch images. The participants engaged in two separate rounds, where each round corresponded to models that had been trained on the two different sized datasets. The ROC study revealed mean average Area Under the Curve (AUC) values close to 0.5 for both rounds. The results from this study suggests that diffusion models can create histopathological patch images that are convincingly realistic where our participants often struggled to reliably differentiate between generated and real images. This underscores the potential of these models as a valuable tool for augmenting digital pathology datasets.

Keywords: Artificial Intelligence, Deep Learning, Digital Pathology, Generative Models, Diffusion Models, Image Synthesis

1. INTRODUCTION

Deep learning typically requires large datasets for effective training of models [1]. However, this requirement can be problematic in the medical/veterinary domains due to reasons such as patient data confidentiality. These problems are further exacerbated in the case of digital pathology due to the invasive nature of obtaining tissues biopsy samples. However, generating synthetic images can potentially provide a solution to increase the size of medical datasets that would aid in training deep learning models. Until recently Generative Adversarial Networks (GANs) were considered the state of the art and demonstrated some success in creating synthetic digital pathology images. However, GANs are prone to unstable training, mode collapse and other architectural problems, potentially yielding sub-optimal synthetic images that fail to capture the true diversity of a training set [2]. Diffusion models are a recent approach that have become popular for

Medical Imaging 2024: Digital and Computational Pathology, edited by John E. Tomaszewski, Aaron D. Ward, Proc. of SPIE Vol. 12933, 129331A · © 2024
SPIE · 1605-7422 · doi: 10.1117/12.3006285

mainstream image generation tasks [3]. There has been relatively little work in the digital pathology space [e.g., 4-9], where statistical evaluation metrics have been limited to the Fréchet Inception Distance (FID) [10] and/or qualitative assessments which are often undertaken to provide feedback on the realism of these generated images. Furthermore, as with other generative methods, it has been suggested that diffusion models require large samples of data for effective training. As a result, this study proposes training an open-source medical domain focused diffusion probabilistic model [4][5] on two different sized datasets (small and large) and evaluating these models using a series of Likert scale-based subjective assessments of pathology image patches, complemented by objective Receiver Operating Characteristic (ROC) analysis to gauge user confidence in discerning generated images from real histopathological image patches.

2. METHODS

2.1 Data collection and patch extraction process

In this study we used a canine Perivascular Wall Tumour (cPWT) dataset to train our models [11][12][13][14]. A veterinary pathologist diagnosed canine Soft Tissue Sarcoma (cSTS) using histology slides collected from the Department of Microbiology, Immunology and Pathology, Colorado State University. Additionally, a Non-Animals Scientific Procedures Act 1986 (NASPA) form was submitted to and approved by the University of Surrey (approval number NERA-1819-045). Collaborating pathologists from the University of Surrey chose representative cPWT histological slides and confirmed the grades of these slides. Subsequently, these slides underwent digitisation using a Hamamatsu NDP 2.0 HT scanner. The scanning process was carried out using a magnification of 40x, resulting in an image resolution of 0.23 micrometres per pixel. The length of the scanning process was approximately 150 seconds for an area of 15 mm x 15 mm, resulting in the generation of a digital Whole Slide Image (WSI). For pre-processing, Otsu thresholding was applied to separate background slide from the tissue sample to generate tissue slide maps. Due to the large nature of WSIs, a patch-based approach was necessary and thus the tissue slide maps were generated to ensure that we extracted tissue containing patch images only. Applying a tissue threshold of 0.95 (where 95% of the patch must contain tissue) and non-overlapping patches of size 512 x 512 pixels were extracted from these slides (see examples of patch images in Figure 1). Image patches from 17 WSIs were used for training of these models. Exactly 1000 patches were randomly extracted from each of the 17 WSIs creating a total of 17,000 patch images. As previously mentioned, large datasets are often the requirement for training generative models. As a result, a random selection of 1000 patches from the training set were used to create a small subset training dataset to compare and determine the effectiveness of diffusion models trained on the two very different sized training sets.

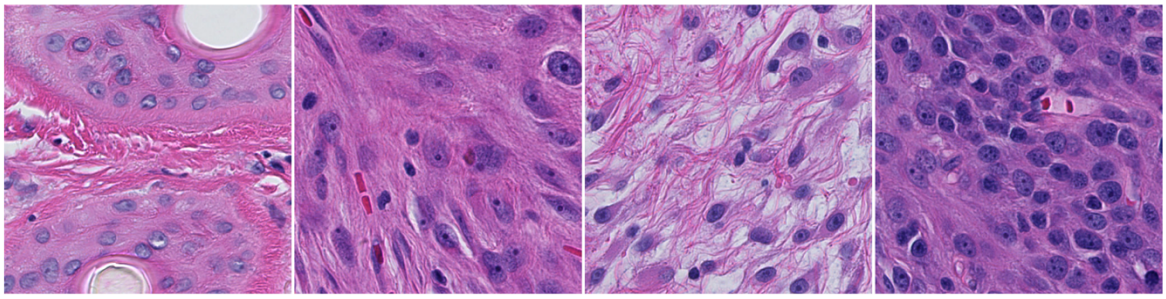

Figure 1. Examples from the canine Perivascular Wall Tumours (cPWT) dataset. All image patches extracted are of size 512 x 512 pixels.

2.2 Model training and experimental set-up

A number of architectures were considered where the model used in this study was the Medfusion model proposed by Muller-Franzes et al. [4][5]. Their Medfusion model is based on the stable diffusion architecture [15] and is made up of two parts: an autoencoder and a denoising diffusion probabilistic model (DDPM). The two parts are trained separately in two phases. The training objective of the autoencoder involved encoding the image space into a compressed latent space,

where the dimensions were reduced by a factor of 8 [4][5]. The pretrained autoencoder (with frozen weights) was then used to encode the image space into a latent space, which was then diffused into Gaussian noise using t=1000 steps as per by Muller-Franzes et al. [4]. A U-Net [16] model was used to denoise the latent space. Samples were then created using a Denoising Diffusion Implicit Model (DDIM) [17] and t=250 steps as per Muller-Franzes et al. [4][5]. For the models using the small training dataset of 1000 images, in the first phase, the autoencoder was trained for a minimum of 10 to a maximum of 50 epochs with the MSE loss function employed. A batch size of 4 was used for training the autoencoder. The second phase using the diffusion model was similarly trained for 50 epochs with a batch size of 4. For the models using the larger training dataset comprising of 17,000 patch images, the autoencoder was trained for approximately 70 epochs with a batch size of 4. The second phase diffusion model training on this dataset consisted of a batch size of 4 trained for approximately 100 epochs.

Six collaborating veterinary medical professionals and pathologists from different institutions took part in the image evaluation stage. The experiments were conducted remotely via screen sharing on Microsoft Teams, with participants being granted complete control over the host's screen for the duration of the session. A Python-based Graphical User Interface (GUI) program was developed to display a series of images to participants (see Figure 2). Each participant viewed a single image at a time and was prompted in the same window to determine whether the presented patch image was real or generated by using a Likert-scale range of options to express their level of confidence. As two different sized datasets were used for training and generating the synthetic images, there were essentially two rounds (or experiments) for the different sized datasets. The datasets consisted of carefully curated real and synthetic images, selected to maintain a consistent level of diversity between both types of images. Before the implementation of these image evaluation experiments, the participants received verbal instructions and engaged in a short 6-image preliminary sample round to familiarise themselves with the process. To minimise biases and potential learning, they were not informed of the accuracy of their selections during this preliminary round. The sequence in which the participants viewed the series of images was randomised for each individual. The first round consisted of 88 real images (n=58) and generated (n=30) images, where the real images were taken from the larger 17,000 image training set and the generated images were created by the model trained on this large dataset. Additionally, the second round consisted of 36 real images (n=22) and generated (n=14) images, where the real images were taken from the smaller 1000-image training set and the generated images were created by the model trained on this smaller dataset. As a consequence, each round was evaluated independently to determine the participants' ability to distinguish between real and generated image outputs trained on datasets of varying sizes.

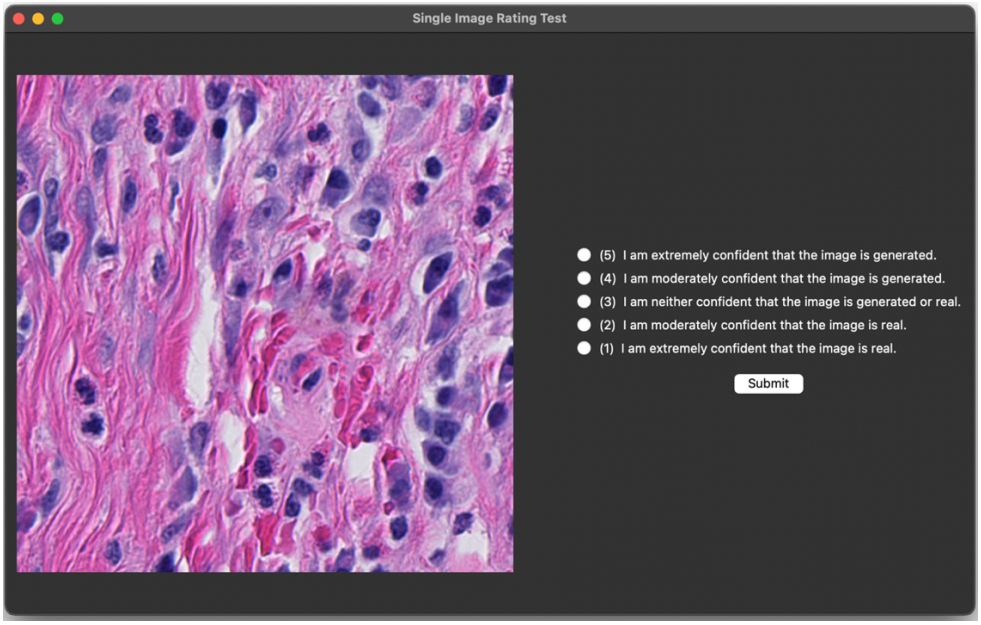

Figure 2. The Python-implemented software depicting the Graphical User Interface (GUI) presenting a single patch image to a participant, with an integrated Likert-scale prompt for assessing the participant's confidence when determining the authenticity of the image as real or generated.

3. RESULTS

3.1 Model metrics

Table 1 presents the results from phase 1 of the Medfusion model: the reconstruction quality of the autoencoder when trained on the small and large datasets respectively. The Multiscale Structural Similarity Index Measure (MS-SSIM) and mean squared error (MSE) values demonstrate the high capability of the image reconstruction, suggesting that the reconstructed images are structurally very similar to the originals. It is apparent that training on a vastly larger training set (17x larger) and for more epochs only resulted in a marginal gain in reconstruction quality. Table 2 presents a quantification of the image generation from the Medfusion model trained on the two different sized datasets, where the model's performance revealed notable differences between two dataset sizes. For the model trained on the smaller dataset of 1,000 images, the model achieved a high precision of 0.86, demonstrating a strong ability to generate relevant pathology images, however produced a lower recall of 0.25, which suggests a limited variety of the generated images. The Fréchet Inception Distance (FID) score was 118.28, which suggests a greater disparity from the real image distribution. The model trained on the large dataset of 17,000 images resulted in a reduced FID of 109.91, indicating improved image quality and a slightly closer match to the real image distribution. However, this model showed a decrease in precision value (0.60) and recall (0.18), presenting a trade-off between image diversity and relevance. Nonetheless, it must be noted that the assessment of these metrics was based on the full (small and large) training sets, comparing them to 70 and 700 generated images from the models trained on the smaller and larger datasets, respectively.

The images were also visually inspected to determine their suitability prior to being shown to the pathologist participants for the ROC study. Many of these images were strikingly realistic looking as depicted in the range of 3 generated (top) and 3 real (bottom) images in Figure 3. Nonetheless, upon visual inspection of the generated image sets combined with the quantitative generation outputs from Table 2, it was apparent that despite their realistic appearance, there was a homogeneity in the visual features of the generated images, suggesting a failure to capture the full diversity of the entire dataset. This is consistent with the quantitative findings of a high FID and low recall for the models trained on both dataset sizes.

Table 1. The Medfusion autoencoder reconstruction quality values where Mean Squared Error (MSE) and Multiscale Structural Similarity Index Measure (MS-SSIM) are presented. In this table, the model is characterised by the volume of data used in its training.

Model	Metric	Value	Mean±Standard Deviation
Small (1000 images)	MS-SSIM	0.962	±0.014
	MSE	0.001	±0.001
Large (17,000 images)	MS-SSIM	0.969	±0.011
	MSE	0.001	±0.0004

Table 2. Image generation performance using Medfusion. Presented are Fréchet Inception Distance (FID), Precision and Recall. The two models trained on the small and large datasets are presented in this evaluation.

Model	Metric	Value
Small (1000 images)	FID	118.28
	Precision	0.86
	Recall	0.25
Large (17,000 images)	FID	109.91
	Precision	0.60
	Recall	0.18

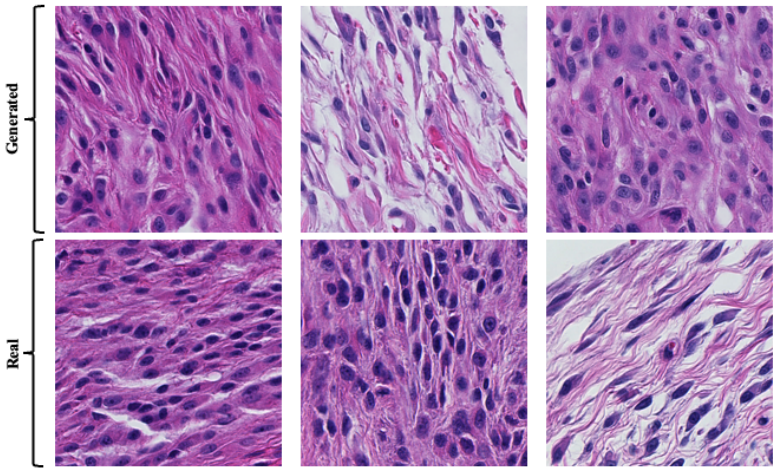

Figure 3. Generated (top) and real (bottom) images. The real images are randomly selected from the larger 17,000 image dataset and the generated images are also a random selection from the 700 outputs from the model trained on the large dataset.

3.2 ROC study

The results for the ROC study are presented in Figure 4, Figure 5 and Table 3, depicting ROC curves and Area Under the Curve (AUC) values presented for each participant. Figure 4 depicts the ROC curves for both the small (n=36) and large (n=88) rounds for each participant. It is apparent that there is great confusion when selecting choices pertaining to the authenticity of the presented images for the large round of 88 images, for all participants, as shown by the ROC curves in Figures 4 and 5 and the values in Table 3 with AUC values close to 0.5. Similar was discovered for the small round, however, there is greater variation amongst the participants, especially Participant #4 who produced an AUC of 0.77, suggesting the best performance and greater ability to discern between real and generated cPWT image patches. The mean AUC values and mean average ROC curves demonstrate an AUC of 0.47 and 0.45 for the small and large rounds respectively.

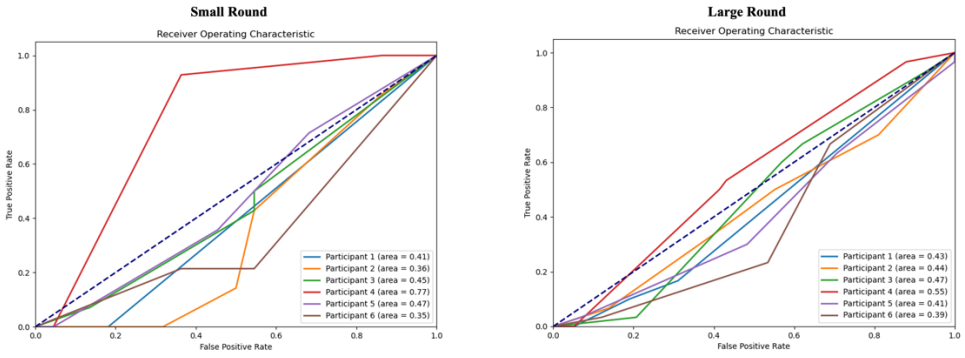

Figure 4. Receiver Operating Characteristic (ROC) curves depicted for both the small (n=36) and large (n=88) rounds for each participant.

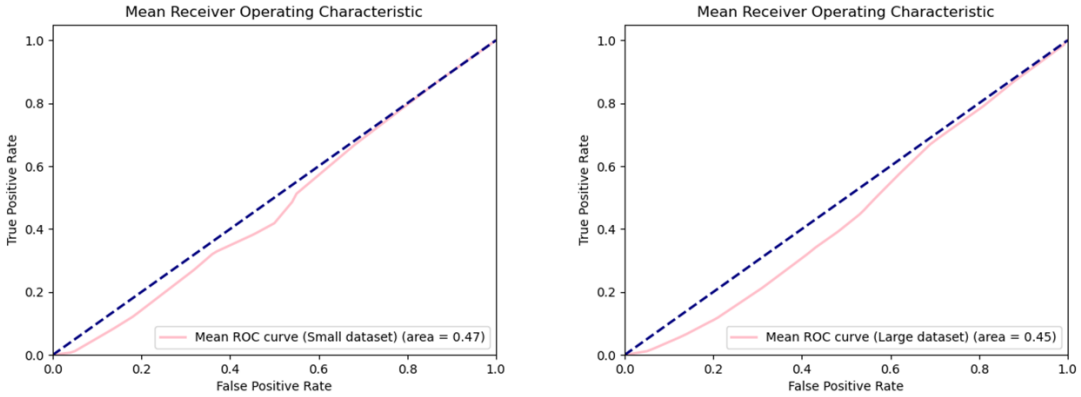

Figure 5. The mean average Receiver Operating Characteristic (ROC) curves depicted for both the small (n=36) and large (n=88) rounds.

Table 3. AUC values from the ROC plots in Figures 4 and 5, where the mean average AUC for the small round is 0.47 and for the large round is 0.45.

Participant	AUC (Small)	AUC (Large)
Participant #1	0.41	0.43
Participant #2	0.36	0.44
Participant #3	0.45	0.47
Participant #4	0.77	0.55
Participant #5	0.47	0.41
Participant #6	0.35	0.39
Mean Average	**0.47**	**0.45**

4. DISCUSSION

The objective of this study was to generate realistic looking histopathology images using the Medfusion model to assess if pathologists could be deceived by them. Results from the Receiver Operating Characteristic (ROC) analysis indicate significant difficulty for the participants in distinguishing between authentic and generated images when they were intermixed randomly. Area Under the Curve (AUC) values close to 0.5 imply that the participants were essentially guessing, demonstrating their inability to reliably identify the generated cPWT patch images from the real ones. Future work should include more participants for a broader validation of the findings from the ROC study as well as including additional meta-data that pertains to the level of experience within veterinary pathology.

ACKNOWLEDGEMENTS

We would like to extend our thanks to Zoetis and the National Physical Laboratory (UK) for their support during this study.

REFERENCES

[1] Hinterstoisser, Stefan, et al. "On pre-trained image features and synthetic images for deep learning." *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.
[2] Saxena, Divya, and Jiannong Cao. "Generative adversarial networks (GANs) challenges, solutions, and future directions." ACM Computing Surveys (CSUR) 54.3 (2021): 1-42.
[3] Moon, Taehong, et al. "Fine-tuning Diffusion Models with Limited Data." NeurIPS 2022 Workshop on Score-Based Methods. 2022.
[4] Müller-Franzes, Gustav, et al. "Diffusion probabilistic models beat gans on medical images." arXiv preprint arXiv:2212.07501 (2022).
[5] Müller-Franzes, Gustav, et al. "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis." *Scientific Reports* 13.1 (2023): 12098.
[6] Oh, Hyun-Jic, and Won-Ki Jeong. "DiffMix: Diffusion Model-based Data Synthesis for Nuclei Segmentation and Classification in Imbalanced Pathology Image Datasets." arXiv preprint arXiv:2306.14132 (2023).
[7] Moghadam, Puria Azadi, et al. "A morphology focused diffusion probabilistic model for synthesis of histopathology images." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[8] Shrivastava, Aman, and P. Thomas Fletcher. "NASDM: Nuclei-Aware Semantic Histopathology Image Generation Using Diffusion Models." arXiv preprint arXiv:2303.11477 (2023).
[9] Osorio, Pedro, et al. "Latent Diffusion Models with Image-Derived Annotations for Enhanced AI-Assisted Cancer Diagnosis in Histopathology." arXiv preprint arXiv:2312.09792 (2023).
[10] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 (2017).
[11] Rai, Taranpreet, et al. "Deep learning for necrosis detection using canine perivascular wall tumour whole slide images." Scientific Reports 12.1 (2022): 10634.
[12] Morisi, Ambra, et al. "Detection of Necrosis in Digitised Whole-Slide Images for Better Grading of Canine Soft-Tissue Sarcomas Using Machine-Learning." Veterinary sciences 10.1 (2023): 45.
[13] Rai, T., et al. "Investigating the potential of untrained convolutional layers and pruning in computational pathology." Medical Imaging 2023: Digital and Computational Pathology. Vol. 12471. SPIE, 2023.
[14] Rai, Taranpreet Singh. Deep learning for necrosis and mitosis detection in canine soft tissue sarcoma whole slide images. Diss. University of Surrey, 2023.
[15] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[16] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.
[17] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." arXiv preprint arXiv:2010.02502 (2020).