Diffusion with Adversarial Fine-Tuning for Improving Rare Retinal Disease Diagnosis

Dominika Iwanicka and Ping Lu

University of Leeds, Leeds, UK
sc20dzi@leeds.ac.uk, p.lu@leeds.ac.uk

Abstract. As machine-aided disease diagnosis becomes more common, there is a rising need for high volumes of quality data, which might be unavailable for rare diseases. Generative methods offer a solution, allowing for synthesising realistic-looking data that can improve diagnosis accuracy. We investigate the applications of diffusion to a small, imbalanced dataset of Optical Coherence Tomography (OCT) images. We propose modifying the basic Denoising Diffusion Probabilistic Model with attention mechanisms, a class-aware training strategy, and the addition of adversarial fine-tuning. We demonstrate that this model is capable of synthesising realistic-looking images with class-specific features even for diseases with as little as 22 samples. We achieve values of FID at 62.58, and CLIP Similarity at 0.96. We show that the addition of generated data in the training dataset improves the overall and class-specific performance of a ResNet18 classifier on the OCT data, offering an improvement for downstream tasks such as rare retinal disease diagnosis.

Keywords: Medical imaging · Data augmentation · Deep Learning · Generative artificial intelligence · Diffusion models.

1 Introduction

Retinal diseases are a becoming increasingly common, affecting over 2.5 million people in the UK alone [1]. Early and accurate diagnosis is crucial, and imaging technologies like Optical Coherence Tomography (OCT) provide high-resolution, cross-sectional images of the retina, allowing for a detailed visualisation of the pathological changes and enabling disease assessment.

In the recent years, machine learning has become prevalent in medical image analysis. Automated retinal disease classification based on OCT has shown promise in improving diagnostic accuracy, however the performance of these ML models relies on the availability of large quantities of labelled data [2]. Class imbalance is a particular challenge – in many existing datasets, certain diseases are underrepresented and lack the sufficient number of samples to enable accurate classification. This can lead to biased predictions, as the model can achieve good accuracy scores by generalising to the common conditions.

Most of the existing OCT classification models focus on the retinal diseases that occur frequently and therefore have a lot of data available, however the rare diseases often get left out due to the insufficient number of samples. For example, Age-Related Macular Degeneration (AMD) is a condition that affects over 25% of people over 60 years old in Europe [3]. The dataset used in this study [4] includes images of conditions such as Retinal Artery Occlusion (RAO), which affects 0.7% of people over 55 years old in Europe [5]. AMD is included in most existing OCT classifiers and is diagnosed with a good accuracy [6–9], but RAO is not included in any of the existing models. This is the case for multiple other diseases as well. In a world where machine-aided diagnosis is on the rise, such discrepancies could lead to generalisation and misdiagnosis.

To address this problem, generative data augmentation techniques have been explored to synthesise new, high-quality medical images. A model trained on such data could become more robust and learn to classify rare conditions better. Among the existing techniques, diffusion models have emerged in the recent years as a promising technique for medical image synthesis, enabling the creation of diverse samples that could be indistinguishable from real ones. However, this task is particularly challenging due to the small volume of data for certain diseases. While diffusion has achieved promising results when trained on an unconditional dataset of 1000 images [10], we are the first ones to consider its applications for a conditional dataset with some classes having as little as 22 samples.

In this work, we investigate the use of diffusion models to augment a small, imbalanced OCT dataset. A Denoising Diffusion Probabilist Model is used as a baseline, and various techniques are implemented to improve the model's ability to focus on the overall structure and learn the fine-grained details of the retina. We evaluate the performance of a classifier model trained on the original dataset, a dataset with only basic geometric augmentation, and a dataset with generative data augmentation.

We present a diffusion model modified with attention mechanisms, adversarial fine-tuning, and and a class-aware training strategy to address the challenge of generating data based on a small, imbalanced dataset. The proposed model is capable of synthesising realistic-looking data, and the inclusion of such data in the training for a classifier improves overall and class-wise performance.

The contributions of this paper can be summarised as follows:

- We present a novel diffusion model for synthesising OCT images for a small, strongly imbalanced dataset.
- In particular, we modify a DDPM with attention mechanisms and propose a multi-step training process that modulates the class embedding weight and incorporates adversarial fine-tuning.
- We demonstrate that the proposed model can synthesise realistic-looking OCT data for classes with as little as 22 images in the original dataset.
- In addition, we show that retinal disease classification based on OCT data is significantly improved with the inclusion of synthetic data in the training dataset.

2 Related work

Data quantity and distribution have been highlighted as significant limitations in the potential applications of deep learning in the medical field [11, 12]. Augmenting the dataset is a common approach for tackling this problem. Basic augmentations, such as geometric transformations or intensity operations, are used in most studies and can improve model performance [13]. Over the past years, image synthesis has been used as a form of generative data augmentation. It has shown a lot of promise and outperformed models trained on datasets with no augmentation or basic transformations only [13, 14].

Diffusion has emerged as a promising solution to medical image synthesis, allowing for generating high-quality data [15–19]. Denoising Diffusion Probabilistic Models (DDPM) were introduced in 2020 [20] as a novel approach that utilises a noise scheduler in the forward process and a UNet backbone to reverse the noise. DDPMs have since been successfully used for a variety of medical tasks, such as image segmentation [19, 21], denoising [22], or classification [23], amongst multiple others [24]. DDPMs have also performed well on small and imbalanced datasets [25]. An existing study on few-shot image synthesis shows that diffusion can generate images based on an unconditional dataset of as little as 1000 images [10]. Gupta et al. [26] demonstrate the applications of diffusion to few-shot synthesis on a conditional dataset.

The performance of diffusion models can be further improved by utilising attention mechanisms in the UNet model used for denoising [27–29]. Amongst these, Multi-headed Self Attention is a prominent variant that allows the model to learn global, long-range dependencies between input and output [30]. Another promising approach to refining the model is discriminator guided training, which incorporates the predictions of a discriminator to correct the diffusion model and helps improve its generative performance [31–33].

3 Methods

3.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models [20] learn to generate new data by gradually removing noise from an image.

The forward process adds noise to images from the dataset by following a pre-defined noise schedule. It is described by a Markov chain where Gaussian noise is added to an image x_0 over T timesteps according to a variance schedule β. The noisy image for a timestep t is given by:

$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon_T$$

where ϵ is sampled from a standard distribution. At $t = T$, the image becomes pure random noise.

The reverse process uses a model to progressively denoise the image. Starting from pure noise at $t = T$, it predicts the noise $\epsilon_\theta(x_t, t)$ at the previous timestep.

A UNet neural network serves as the backbone of this model. The prediction is then used to remove noise from the image as follows:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z$$

where $\alpha_t = 1 - \beta_t$ and z is random noise sample from a standard distribution. Over time, the model learns to generate entirely new images from pure random noise. Figure 1 illustrates the diffusion process.

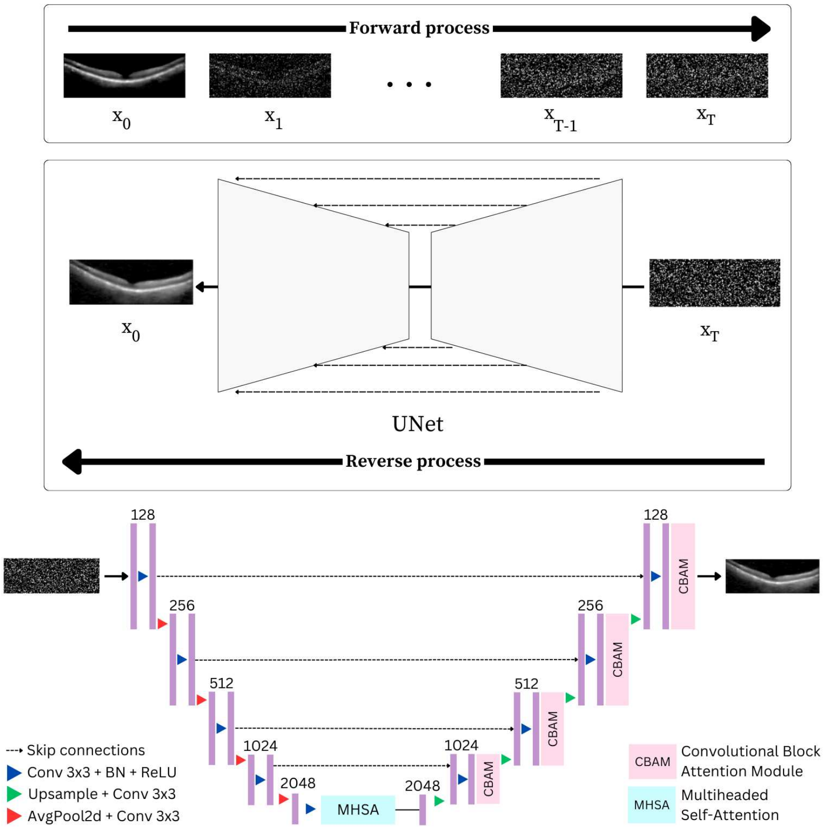


Fig. 1. An overview of the forward and reverse diffusion processes is at the top. The forward process progressively adds noise to the image, and the reverse process uses a UNet to predict the noise and remove it. The UNet architecture is further detailed at the bottom.

3.2 Attention mechanisms

To improve the model's ability to learn fine details, attention blocks were added to the UNet. At the bottleneck between downscaling and upsampling blocks, Multi-Headed Self Attention (MHSA) was added. Each attention module captures the complex long-range dependencies across different locations, and the multiple heads allow for learning distinct relationships [34]. This can be crucial for preserving the structural integrity of the synthesised images.

In addition, Convolutional Block Attention Module (CBAM) was added at the end of each upsampling block in the UNet [35]. CBAM combines channel and spatial attention, which enhances feature representation by allowing the model to focus on the most relevant locations in an image. This improves the model's ability to learn fine-grained details, leading to a sharper output. The UNet architecture is illustrated in Figure 1.

3.3 Class-aware training

To make the model conditional, class embeddings need to be introduced and passed as input for the UNet model. This allows for generating class-specific samples of each class. The class embeddings get added to the positional embeddings. A class embedding weight was added to control how strongly the model is influenced by the class vs positional embeddings.

Due to the existing class imbalances, class weights reflecting the proportions in the dataset were included and used in a weighted loss function, to ensure the model learns to represent all classes correctly. The loss is given by:

$$L = \frac{1}{N} \sum_{i=1}^{N} w_{y_i} L_{SmoothL1}(\epsilon_i, \hat{\epsilon_i})$$

where N is the batch size, w is the class weight for a given class label y_i, and $\epsilon, \hat{\epsilon}$ represent the true noise and predicted noise respectively.

The imbalance of the original dataset posed a particular challenge in making the model explainable and learning the class-specific features. While training the model without class embeddings results in a good generalisation and reflects the structure of the retina well, a good understanding of the class characteristics is needed to accurately represent the rare diseases. We attempt to counter the imbalance through class-aware training, by modifying the class embedding weight throughout the training process.

3.4 Adversarial fine-tuning

A discriminator can be used during the diffusion training process to correct and guide the diffusion model [32, 33]. This has been shown to improve the results generated by the model.

Initial results showed that the proposed diffusion model was learning the important features of the retina, but struggling to capture fine details. Similar to

Generative Adversarial Networks, the discriminator learns to distinguish between real and synthetic images. The discriminator was then used in a loss function for the diffusion model, enhancing its ability to learn fine-grained details.

4 Experiments

4.1 Dataset

The dataset used for this study is the Optical Coherence Tomography Dataset (OCTDL) [4], which contains 2064 images and represents 7 retinal diseases: Age-related Macular Degeneration (AMD), Diabetic Macular Edema (DME), Epiretinal Membrane (ERM), Retinal Artery Occlusion (RAO), Retinal Vein Occlusion (RVO), and Vitreomacular Interface Disease (VID). Samples of the classes are shown in Figure 2. The dataset presents a challenge because of its severe class imbalance, as the number of class samples range from 1231 for AMD to 22 for RAO, which reflects how common or rare the diseases are.

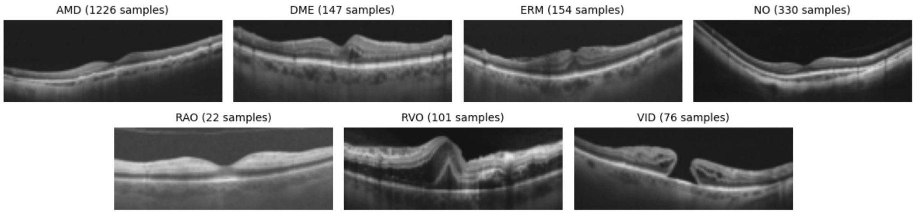


Fig. 2. Contents of the OCTDL dataset. Image captions denote class names and the number of samples.

The images in the dataset vary significantly in size. As part of the data preprocessing, 8 outliers were removed using a Z-score with a 2.5 threshold tuned experimentally. The remaining images all followed similar proportions, which were averaged to $width = 2.74 \times height$. All data points were resized according to this, using the minimum height found in the dataset, to ensure none of the images get stretched out. This resulted in a standardised dataset of 2056 images scaled to $(199, 546)$.

4.2 Implementation details

Pre-processing. Basic data augmentation (horizontal flip) was applied to all classes except AMD to double their size. Additionally, the samples from RAO were further doubled by applying random rotation of 5-15 degrees and decreasing their contrast. For training, the images were scaled down to $(94, 256)$ and normalised to $[-1, 1]$.

Hyperparameters. The baseline diffusion model utilises $T = 1000$ timesteps for the forward process and a 5-layer UNet as a backbone for the reverse process, with an 8-headed MHSA at the bottleneck. The UNet model was trained with a batch size of 8, Adam optimizer with a learning rate of 10^{-3}, and The Huber Loss (Smooth L1 Loss) with $\beta = 0.1$.

For adversarial fine-tuning, a simple discriminator was implemented consisting of 3 convolutional layers with batch normalisation, dropout with a 0.5 rate, and ReLU activation. The discriminator used a learning rate of 10^{-5} and a BCE loss function.

Training. The models were trained on the Aire HPC system at University of Leeds equipped with 3 NVIDIA L40S 48GB GPUs. All code was developed with Python 3.9 using Pytorch with CUDA support. The training time for the diffusion model was 7 hours, followed by 20 hours of fine-tuning.

The training was carried out in 3 steps:

1. The class embedding weight was set to 0 for a 100 epochs. This was done so that the model learned the overall structure of the retina without focusing on the class-specific features. Previous experiments have shown that including class embeddings from the start makes the model learn the common classes very well, but leads to poor performance for the smaller classes.
2. The class embedding weight was increased to 2 for the following 300 epochs, leading to the model learning class-specific features on top of the existing knowledge of retinal structure.
3. The diffusion model was trained alongside the discriminator for 100 epochs using an adversarial loss function, which used BCE to assess how well the discriminator distinguishes between real and fake images. This was used instead of the regular diffusion loss.

4.3 Evaluation metrics

Quantitative evaluation. We use Inception Score (IS), Fréchet Inception Distance (FID), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ration (PSNR) for evaluation. FID calculates the difference between real and generated images by comparing the feature distribution, while IS measures the diversity of the generated images by assessing the confidence and variety of predictions made by a pre-trained classifier [36]. These are used as benchmarks in most studies on image generation. We utilise the CLIP Similarity [37] as an alternative to the inception-based metrics.

Additionally, we use SSIM and PSNR are to further assess image quality [38]. SSIM measures the perceptual similarity between the images, and PSNR expresses the ratio between the maximum power of a signal and the power of the noise. We combine the generative metrics (FID, IS, CLIP Similarity) with reconstruction metrics (SSIM and PSNR) to better assess the quality of the generated images.

Downstream task evaluation. We perform further evaluation by utilising a ResNet18 [39] classifier on three instances of the dataset: basic dataset, a dataset with geometric augmentations, and a dataset with generative augmentations performed by our best diffusion model. ResNet18 has been successfully used for OCT classification tasks in previous studies [9]. Previous works in this field highlight the importance of model evaluation on downstream tasks as opposed to relying only on metrics like FID or IS [40]. The classifier was trained for a 100 epochs using a batch size of 32 and the Adam optimiser with a learning rate of 10^{-3}. Prior to training, we set aside 20% of the original dataset for testing to ensure the model is evaluated only on real, unaugmented data.

5 Results

We use a DDPM model as a baseline for comparison against the proposed modifications - addition of attention modules, class-aware training, and adversarial fine-tuning. We analyse all configurations of these modifications to investigate their impact on the model's performance.

5.1 Quantitative evaluation

Table 1 shows the quantitive evaluation for the diffusion models. The metrics were computed on 2000 images. An improvement for all metrics is visible for the final model. The lower FID value indicates a closer feature distribution between real and generated data, and the higher IS reflects the feature diversity. We noted that IS calculated for the original dataset was low, measuring 2.94 ± 0.17. While FID remains relatively high even for the best model, CLIP Similarity indicates a very good semantic similarity between the original and generated data. The increased SSIM and PSNR values show an improvement in perceptual and pixel-wise image similarity.

We can observe that in isolation, class-aware training brings the biggest improvement to the FID score, but adversarial fine-tuning influences IS, PSNR, and CLIP similarity more. SSIM only improves only with the combination of multiple additions. Including all three proposed modifications brings the best results across all metrics.

5.2 Generated images

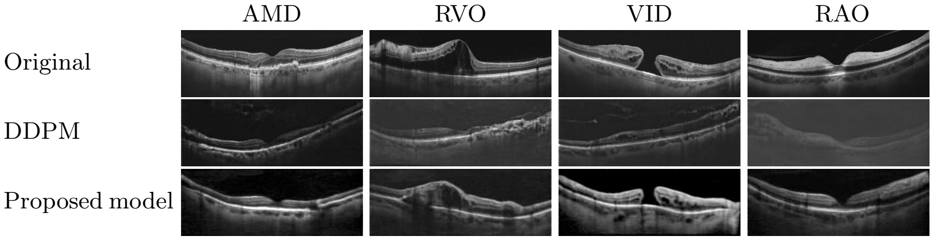
Table 2 showcases the synthesised images for four different classes, compared against images sampled from the original Dataset. In particular, we display AMD (1231 samples), RVO (101 samples), VID (76 samples), and RAO (22 samples) to demonstrate the performance of the different models on classes of varying sizes.

A visual inspection shows that a basic DDPM learns the general shape of the retina, but generalises too much. No class-specific features are visible, and the images for the smaller classes are more blurry and of a lower quality.

Table 1. Evaluation metrics of the proposed generative model with proposed modifications: attention modules (Atn.), class-aware training (CAT), and adversarial fine-tuning (AFT), compared to DDPM as a baseline. We measure the Inception Score (IS), Fréchet Inception Distance (FID), Structural Similarity Index Measure (SSIM), Peak Signal-to Noise Ratio (PSNR), and CLIP Similarity ∈ $[-1, 1]$. PSNR is given in dB.

Model configuration	IS ↑	FID ↓	SSIM ↑	PSNR ↑	CLIP Similarity ↑
DDPM	2.42 ± 0.01	160.02	0.43	7.04	0.83
DDPM + Atn.	2.72 ± 0.07	131.59	0.45	8.01	0.89
DDPM + CAT	2.36 ± 0.15	112.10	0.45	7.22	0.87
DDPM + AFT	2.85 ± 0.07	127.38	0.44	8.97	0.91
DDPM + Atn. + CAT	2.66 ± 0.09	74.26	0.47	8.04	0.93
DDPM + Atn. + AFT	2.73 ± 0.10	130.95	0.53	8.07	0.91
DDPM + CAT + AFT	2.79 ± 0.08	87.31	0.54	8.89	0.94
DDPM + Atn. + CAT + AFT	**2.91 ± 0.10**	**62.58**	**0.55**	**9.01**	**0.96**

Table 2. Images generated using the proposed model and basic DDPM, compared against the original. The images were randomly selected from the respective datasets. AMD (Age-related Macular Degeneration), RVO (Retinal Vein Occlusion), VID (Vitreomacular Interface Disease), and RAO (Retinal Artery Occlusion) represent selected classes from the OCT dataset.



The results generated by the proposed model are closer to the original, with a similar level of detail, class-specific features, and comparable contrast. However, the model was trained on half the resolution of the original images, which leads to the generated images looking more blurry and losing some detail.

5.3 Classification results

Table 3 demonstrates the results of the classifier trained on the OCTDL dataset using different data augmentation techniques. We compare the performance of the classifier on the unaugmented dataset, a dataset with basic geometric augmentations (horizontal flipping, 5-15 degree rotation), and datasets with generative augmentation done with the proposed diffusion model. We compare the overall accuracy, precision, and recall of the classifier, and the precision and recall scores for the classes.

As a baseline, we used the dataset with no augmentations trained on 1,542 data points. The classifier achieved an accuracy score of 83%. Amongst the

Table 3. Classification results for datasets with different augmentation. AMD, DME, ERM, NO, RVO, RAO, and VID denote classes in the dataset and represent different retinal diseases. Results for A - Accuracy, P - Precision, and R - Recall are given in %.

	Overall			AMD		DME		ERM		NO		RVO		RAO		VID	
Augmentation	A	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
None	83	82	83	89	94	61	45	72	64	73	84	100	50	70	37	62	42
Geometric	83	84	83	91	86	61	64	71	66	76	89	97	93	75	68	77	97
Generative	**92**	**93**	**93**	**96**	**100**	**97**	**66**	**90**	**94**	**100**	**92**	**100**	**94**	**78**	**100**	**91**	**100**

classes, AMD achieved the best results, with the remaining classes scoring significantly lower. This showcases the expected bias of a classifier trained on an imbalanced dataset, as big disparities are visible between the class-specific results. This is also reflected in the macro average of accuracy scores, which was only 75%.

The geometric augmentation was used to increase the training dataset size to 1,917. The accuracy remained the same at 84%, but there was an improvement in the recall values for the underrepresented classes. The macro average of accuracy rose to 79%.

The generative augmentation brought the total number of images for each of the classes to 500 to ensure an even split. The final training dataset size was 3,375 images. The test dataset contained exclusively the images from the unaugmented dataset to ensure an accurate evaluation. The generative augmentations have contributed to an increase of accuracy to 92%, and an increase of precision in recall for every single class. The improvement is also reflected in the macro average of accuracy scores, which reached 93% and no longer reflected the class imbalance.

6 Discussion

The implemented additions to a basic DDPM offer significant performance improvements. From Table 2, we have observed that the inclusion of the proposed modifications have allowed the model to retain class features even for diseases with very little samples. The resulting images look realistic and resemble the original well, they retain the general shape of the retina and reflect the distinct class features, although they are of a lower resolution than the original.

Each of the modifications to the basic model has also improved the evaluation metrics, however, the FID value remained high. It is important to note that FID has recently been called into question as the predominant metric for generative models, and it has been observed to frequently contradict human judgement [41, 42]. We utilised CLIP Similarity as an alternative to the Inception-based metrics and have observed a score close to 1, indicating very good semantic similarity between real and fake images. To better assess the quality and realism of the synthetic data, a human evaluation would be necessary.

While adversarial fine-tuning has improved the results, it is important to note that this was a very time-intensive process. The base diffusion model took

8 hours to train for 400 epochs, but fine-tuning it for a 100 epochs took additional 20 hours. However, this had no bearing on the time it took to generate images after the training. We determine that fine-tuning is a trade-off – better visuals and metrics can be achieved, but the training time is significantly extended.

The addition of the generated data to the original dataset had a positive impact on disease classification, as shown in Table 3. We have seen a slight increase in overall accuracy, and a significant increase in the class-specific performance metrics. This is especially visible for the smaller classes, such as VID – despite only having 76 samples in the original dataset, the geometric augmentations have increased the F1 value for this class from 35% to 95%. This shows the promise of our proposed model in downstream tasks, demonstrating that training on a combination of real and synthetic data can improve classification performance on real, unseen data.

Overall, we have achieved promising results. The OCT images synthesised using the proposed diffusion model resemble the original and have a positive impact on the classification of retinal diseases.

6.1 Future work

While our proposed modifications to DDPM offer a significant performance improvement, future work is needed to make it usable in a clinical setting. We note there is a need to train the model on full resolution images. Due to resource limitations, we used a sized-down dataset, but this could lead to critical details being masked. A possible mitigation would be fine-tuning the model on a higher image resolution. Additionally, there is a critical need for expert evaluation of the generated images to ensure the model has accurately learnt to represent the retinal pathologies. Finally, it is important to validate the performance of the proposed model on other datasets. This would demonstrate whether the model can generalise to different applications.

6.2 Ethical concerns

Outside of the existing bias towards rare diseases, it is important to consider that medical datasets often contain biases in terms of age, gender, or ethnicity. While the proposed model has successfully improved the disease imbalance, it could amplify other biases that might have existed in the data. Moreover, there could be concerns with relying heavily on synthetic data – this technically helps the classifier, but from an ethical viewpoint, it is important to consider how trustworthy that data is. If used without proper validation, it could be misleading. As a future mitigation step, we would aim to evaluate the generated images with an expert in the field to ensure proper medical representation is maintained.

7 Conclusion

In this study, we have investigated the applications of diffusion in medical image synthesis for a small, imbalanced OCT dataset. We proposed modifications to

DDPM that increase the model's ability to learn details and class-specific features, even for classes containing as little as 22 samples. We have shown that this model is capable of generating images resembling the original, and that the inclusion of these images in the training dataset improves classifier performance and enhances the diagnostic accuracy for rare diseases.

Acknowledgments. This work was undertaken on the Aire HPC system at the University of Leeds, UK. The study was carried out as part of a group project comparing the performance of different generative methods. Thank you to my teammates: Kacper Roemer, Ervans Caushi, James Brierley, Jakub Kurasz, and Oleh Chernilevskyi.

References

1. Alasdair N Warwick, Katie Curran, Barbra Hamill, Kelsey Stuart, Anthony P Khawaja, Paul J Foster, Andrew J Lotery, Michael Quinn, Savita Madhusudhan, Konstantinos Balaskas, et al. Uk biobank retinal imaging grading: methodology, baseline characteristics and findings for common ocular diseases. *Eye*, 37(10):2109–2116, 2023.
2. Aleksandar Miladinović, Alessandro Biscontin, Miloš Ajčević, Simone Kresevic, Agostino Accardo, Dario Marangoni, Daniele Tognetto, and Leandro Inferrera. Evaluating deep learning models for classifying oct images with limited data and noisy labels. *Scientific Reports*, 14(1):1–11, 2024.
3. Jeany Q Li, Thomas Welchowski, Matthias Schmid, Matthias Marten Mauschitz, Frank G Holz, and Robert P Finger. Prevalence and incidence of age-related macular degeneration in europe: a systematic review and meta-analysis. *British Journal of Ophthalmology*, 104(8):1077–1084, 2020.
4. Mikhail Kulyabin, Aleksei Zhdanov, Anastasia Nikiforova, Andrey Stepichev, Anna Kuznetsova, Mikhail Ronkin, Vasilii Borisov, Alexander Bogachev, Sergey Korotkich, Paul A Constable, et al. Octdl: Optical coherence tomography dataset for image-based deep learning methods. *Scientific data*, 11(1):365, 2024.
5. Jeany Q Li, Jan Henrik Terheyden, Thomas Welchowski, Matthias Schmid, Julia Letow, Caroline Wolpers, Frank G Holz, and Robert P Finger. Prevalence of retinal vein occlusion in europe: a systematic review and meta-analysis. *Ophthalmologica*, 241(4):183–189, 2019.
6. A Jeya Prabha, C Venkatesan, M Sameera Fathimal, KK Nithiyanantham, and SP Angeline Kirubha. Rd-oct net: hybrid learning system for automated diagnosis of macular diseases from oct retinal images. *Biomedical Physics & Engineering Express*, 10(2):025033, 2024.
7. Malliga Subramanian, Kogilavani Shanmugavadivel, Obuli Sai Naren, K Premkumar, and K Rankish. Classification of retinal oct images using deep learning. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–7, 2022.
8. Reza Rasti, Hossein Rabbani, Alireza Mehridehnavi, and Fedra Hajizadeh. Macular oct classification using a multi-scale convolutional neural network ensemble. *IEEE Transactions on Medical Imaging*, 37(4):1024–1034, 2018.
9. Depeng Wang and Liejun Wang. On oct image classification via deep learning. *IEEE Photonics Journal*, 11(5):1–14, 2019.

10. Hong Kyu Kim, Ik Hee Ryu, Joon Yul Choi, and Tae Keun Yoo. A feasibility study on the adoption of a generative denoising diffusion model for the synthesis of fundus photographs using a small dataset. *Discover Applied Sciences*, 6(4):188, 2024.
11. Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the royal society interface*, 15(141):20170387, 2018.
12. Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021.
13. Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of medical imaging and radiation oncology*, 65(5):545–563, 2021.
14. Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: a review. *Journal of imaging*, 9(4):81, 2023.
15. Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical imaging with deep learning*, 2022.
16. Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023.
17. Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
18. Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.
19. Muhammad Usman Akbar, Måns Larsson, Ida Blystad, and Anders Eklund. Brain tumor segmentation using synthetic mr images-a comparison of gans and diffusion models. *Scientific Data*, 11(1):259, 2024.
20. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
21. Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024.
22. Dewei Hu, Yuankai K Tao, and Ipek Oguz. Unsupervised denoising of retinal oct with diffusion probabilistic model. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 25–34. SPIE, 2022.
23. Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. Diffmic: Dual-guidance diffusion network for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 95–105. Springer, 2023.
24. Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2023.

25. Iman Khazrak, Shakhnoza Takhirova, Mostafa M Rezaee, Mehrdad Yadollahi, Robert C Green II, and Shuteng Niu. Addressing small and imbalanced medical image datasets using generative models: A comparative study of ddpm and pggans with random and greedy k sampling. *arXiv preprint arXiv:2412.12532*, 2024.
26. Parul Gupta, Munawar Hayat, Abhinav Dhall, and Thanh-Toan Do. Conditional distribution modelling for few-shot image synthesis with diffusion models. In *Proceedings of the Asian Conference on Computer Vision*, pages 818–834, 2024.
27. G Jignesh Chowdary and Zhaozheng Yin. Diffusion transformer u-net for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 622–631. Springer, 2023.
28. Niharika Das and Sujoy Das. Attention-unet architectures with pretrained backbones for multi-class cardiac mr image segmentation. *Current Problems in Cardiology*, 49(1):102129, 2024.
29. Behzad Hejrati, Soumyanil Banerjee, Carri Glide-Hurst, and Ming Dong. Conditional diffusion model with spatial attention and latent embedding for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 202–212. Springer, 2024.
30. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
31. Filip Ekström Kelvinius and Fredrik Lindsten. Discriminator guidance for autoregressive diffusion models. In *International Conference on Artificial Intelligence and Statistics*, pages 3403–3411. PMLR, 2024.
32. Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022.
33. Ling Yang, Haotian Qian, Zhilong Zhang, Jingwei Liu, and Bin Cui. Structure-guided adversarial training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7256–7266, 2024.
34. Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 267–276. Springer, 2021.
35. Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
36. Eyal Betzalel, Coby Penso, Aviv Navon, and Ethan Fetaya. A study on the evaluation of generative models. *arXiv preprint arXiv:2206.10935*, 2022.
37. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
38. Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.
39. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

40. Hadrien Reynaud, Qingjie Meng, Mischa Dombrowski, Arijit Ghosh, Thomas Day, Alberto Gomez, Paul Leeson, and Bernhard Kainz. Echonet-synthetic: Privacy-preserving video generation for safe medical data sharing. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 285–295. Springer, 2024.
41. Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024.
42. George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.