Diffusion Models-based Data Augmentation for the Cell Cycle Phase Classification

Zirui Chen[1*]

[1]Department of Computer Science, City University of Hong Kong, 999077, Hong Kong

[*]Corresponding author email: ziruichen6-c@cityu.edu.hk

Abstract. For biological research, sample size imbalance is common due to the nature of the research subjects. For example, in the study of the cell cycle phase, the sample size of dividing cells is also much smaller due to the extremely short duration of the mitotic phase compared to the interphase. Data augmentation using image generative models is an excellent way to address insufficient sample size and imbalanced distribution. In addition to the GAN-like models that have been extensively applied, the diffusion model, as an emerging model, has shown extraordinary performance in the field of image generation. This experiment uses the diffusion model as a means of image data enhancement. The experimental results expose that the performance of the classifier with data augmentation is significantly improved compared with the original dataset, and the positive predictive value is increased from about 0.7 to more than 0.9. The results reveal that the diffusion model has a good application prospect in the area of data enhancement and can effectively solve the problem of insufficient data or unbalanced sample size.

Keywords: Biology image classification, data augmentation, diffusion models.

1. Introduction

As an essential part of the immune system, the immune response of T cells plays a vital role in the body's anti-infection. It is not only involved in immune protection but also an essential factor leading to immune pathology. The division of the T cell cycle phase is crucial for diagnosing and clinical countermeasures to the disease [1]. Jurkat cells are a type of immortalized T lymphocyte lineage. Because it retains the characteristics of human peripheral blood T lymphocytes, the Jurkat cell model has been widely used in the simulation experiments of T cells.

Compared with traditional image classification techniques, convolutional neural network technology classifiers have superior performance in recognizing cell morphology and extracting cell features [2]. However, the characteristics of organisms lead to the fact that datasets for biomedical research are often unbalanced, with general types of data samples playing a dominant role. Due to the characteristics of CNN classifiers, when the dataset exhibits a significant imbalance, the classification accuracy also exhibits a significant reduction, which makes it challenging to apply models in the real world [1, 3].

Data augmentation is essential to obtain a dataset with a uniform number of samples (more suitable for real-world applications). Traditional data augmentation methods often enrich datasets from the data level, employing image processing techniques such as geometry and color transformation to expand the sample size of a few types. This method often needs to be designed according to human subjective thoughts, and it is also likely to lead to overfitting [4]. In recent years, the rapid development of

generative models has provided a new means for data augmentation. The high-quality images generated by the Generative adversarial network (GAN) model have been proven capable to solve the problem of low classification precision caused by the insufficient number and the unbalanced distribution of original image sets [1, 5, 6, 7]. Diffusion models are an emerging generative model inspired by considerations from nonequilibrium thermodynamics. The inference is achieved by superimposing noise during forwarding propagation and lossy removal of noise during backpropagation [8, 9]. Although it has outperformed GAN models in terms of generated image quality, image type diversity, and training difficulty, its performance has not been tested in the biomedical domain [10].

This project aims to test data augmentation of Image Classification for the Cell Cycle Phase via a diffusion model. The superior performance of the diffusion model is demonstrated by comparing the classifier performance of the primitive data, the data augmented with the GAN model, and the data augmented by the diffusion model.

The remainder of the paper is arranged as follows. The dataset applied for the experiment and diffusion model data augmentation methods is introduced in Section 2. The comparison of the classification result and analysis are presented in Section 3. Section 4 consists of the discussion and future improvement. The conclusion is in Section 5.

2. Methodology

2.1. Dataset

The cell cycle image set includes 32266 images of Jurkat cells of size 66*66 pixels, collected from the Image Stream platform [2]. The data consists of 7 stages: G1, S, G2, prophase, metaphase, anaphase, and telophase. Distinguishing between interphase (G1, S, G2) and mitotic (prophase, metaphase, anaphase, telophase) cells is relatively simple, and current studies have been able to achieve an accuracy rate of over 98%. However, because the morphological structure of interphase cells is very similar, the accuracy of distinguishing the three phases of interphase is only about 79% [1].

In addition, due to the short duration of the mitotic phase in the complete cell cycle, the data volume of Prophase, Metaphase, Anaphase, and Telophase in the original images is extremely small, as shown in Table 1. The imbalance in the amount of data can easily lead to inaccurate classification results. This study aims to use data augmentation to balance the number of images at each stage for better classification. The ratio between the train set, validation set, and test set would be 6:2:2.

Table 1. The number of images of different cell stages and the number of generated images by DDPM

Cell Cycle Stage	Original Image	Generated Image by DDPM	Total number of images used for classification
G1	14333	/	8600
G2	8601	/	8601
Mitotic	716	7160	7876
S	8616	/	8616

2.2. Model introduction

In this research, Denoising Diffusion Probabilistic Models (DDPM) are utilized as a means of data augmentation. DDPM has shown excellent performance in the field of image generation. The basic idea is to gradually add noise to the original image and denoise the sample of pure noise to get a coherent image [11]. Although slightly inferior to GAN in sampling time, the model structure of DDPM is simpler and does not require complex parameter adjustment in the GAN training process. Meanwhile, existing research has demonstrated the superiority of DDPM-generated images in terms of fidelity [9]. Therefore, it is a reasonable and feasible idea to use the diffusion model instead of the GAN model in some cases.

In the next part, the basic mechanism of DDPM would be briefly introduced from three components: forward propagation, backward propagation, and loss function.

2.2.1 Forward process

In the process of forward propagation, a certain amount of noise is added to the image at each time step, and finally, the original image x0 is converted into a pure noise image after T time steps. Since the image x_t at time t only depends on the image at the previous moment (that is, x_{t-1}), the entire forward propagation can be regarded as a Markov chain. Therefore, the process of diffusion can be expressed as (1):

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q\left(x_t|x_{t-1}\right) = \prod_{t=1}^{T} N\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathrm{I}\right) \tag{1}$$

the process of adding noise can be described by a multivariate Gaussian distribution. The mean is determined by the previous image and the variable β. The variance of this distribution is fixed as the product of β and identity. The larger the β, the greater the blurring. It is necessary to adjust the size of β to ensure that the image approaches pure noise after T steps without going so fast that too much information is lost [8, 9]. The image change process in forward propagation is shown in Figure 1. As T increases, the image gets closer to pure noise.

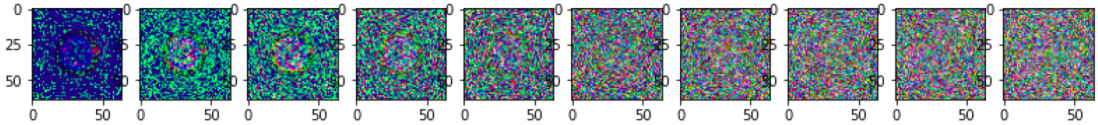

Figure 1. Sample Image of the diffusion process. From left to right, the image gets closer and closer to pure Gaussian noise as the noise is added.
(Photo Credit: Original)

2.2.2 Backward propagation

As opposed to the forward propagation process, backward propagation is the process of obtaining an image from pure noise by gradually getting rid of the noise. The process can also be represented by a Markov chain (2):

$$p(x_{T:0}) = p(x_T) \prod_{t=1}^{T} p\left(x_{t-1}|x_t\right) \tag{2}$$

The noise reduction process obtains xt-1 by predicting the noise added each time and subtracting it from x_t. Since the sum of Gaussians is also Gaussian, it is feasible to sample each time step image independently. For any given time, t, x_t can be derived from x0 and the accumulated mean and variance. By feeding the model an image annotated at time t, it is viable to obtain the mean and variance used in the noise reduction process.

Backpropagation is implemented through the U-net model. U-net is a special kind of neural network model. As the model deepens, the size of the input will gradually decrease, while the depth (number of channels) will gradually increase until the bottleneck, and then the depth will decrease, and the size will increase until it reaches the dimension of the input. Its input and output have the same dimension, which meets the requirements of the diffusion model.

2.2.3 Loss function

The formula of the Loss function is shown below. Since it involves more complex inferences, it will not be introduced here. In short, the loss function optimizes the model by minimizing the difference between the predicted noise and the target noise.

$$L_{simple} = \mathbb{E}_{t,x_0}\ \left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right] \tag{3}$$

3. Experiment result

3.1. The result of image generation

The sample images generated by DDPM are shown in Figure 2. The samples show that the images generated by DDPM have a high sense of realism, and the structural features of various types of cells are well extracted and restored on the basis of maintaining authenticity. This experiment applies common image data augmentation methods, such as rotation, colour inversion, etc., to ensure the robustness of training.

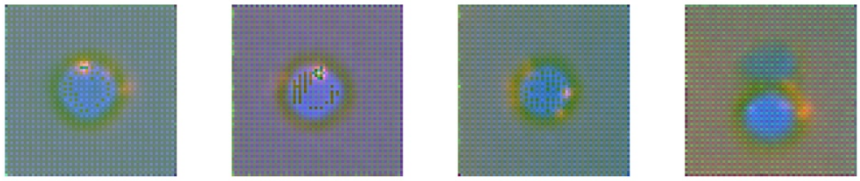

Figure 2. Sample Image of cells of mitotic stage generated by DDPM.
(Photo Credit: Original)

3.2. The result of image classification

The comparison of image classification results would be between the original image set, the image set with the image generated by the WGAN-GP model (from Jin, X., Zou, Y., and Huang, Z. (2021)), and the image set with the image generated by the DDPM model. The classifier adopts the Resnet architecture to ensure the effectiveness of the horizontal comparison of data.

Considering the data imbalance of the original dataset, the positive predictive value (PPV, or called precision, represents the percentage of real positive samples among all positive samples labeled by the classifier) is chosen to measure the quality of the classification results in this experiment. The experimental results are presented in Table 2.

Table 2. The PPV comparison of images classification generated by the DDPM mode.

Cell Cycle Stage	Original	WGAN-GP	DDPM
G1	0.8316	0.8181	0.8148
G2	0.8453	0.8456	0.8478
Mitotic	0.7183	0.9744	0.9412
S	0.6765	0.6700	0.6902

The results show that the PPV of the unmodified three types of data (G1, G2, S) is not much different from the original data set, within the margin of error. The PPV results are not degraded by the decreasing relative proportion in the training set.

In contrast, due to the small sample size of the M stage, the classifier using the original image set cannot achieve effective classification, and the PPV is only about 0.7. Compared to GAN models that have been extensively tested on various datasets, DDPM shows comparable performance, achieving a classification PPV of over 0.94 for dividing cells.

With the support of sufficient data, the reason why dividing cells can achieve better classification (PPV above 0.9, while in interphase) may be that the activity of dividing cells is more intense. Phenomenon that are easier to observe from cell structure, such as chromosome doubling, cell division, etc., contribute to the classification of the cell cycle, so a better classification effect can be achieved.

4. Discussion

In this study, only one single dataset is used to demonstrate the performance of DDPM. To obtain more comprehensive and objective experimental results, it is essential to apply diverse datasets for testing.

At the same time, the DDPM model also has room for optimization. In addition to the original version of DDPM used in this experiment, the performance of more advanced diffusion models with various improvements in practical applications deserves further study [9]. The research on DDPM optimization

has progressed, and "OpenAI" has proposed some methods that can further improve the image quality, like additional normalization layers and residual connections [10]. Since DDPM realizes the forward and reverse processes of diffusion through Markov chains, its speed is limited by the step size (since multiple forward passes are required). Current research, such as Denoising diffusion implicit models (DDIM), achieves acceleration at the expense of image diversity [12]. How to realize the acceleration of the diffusion model while ensuring the generation quality has important practical significance for the large-scale application in the future.

In addition, it is worth investigating whether the images generated by DDPM have biomedical plausibility. How to ensure that the model learns the complete cellular structure, not just from the image level, is a topic worthy of future research.

5. Conclusion

For biological research, unbalanced sample size is a very common research difficulty due to the nature of the study population. This paper applies diffusion models to the field of data augmentation for cell cycle classification. The four-stage classification of the cell cycle performed on the Jurkat dataset achieved better classification results after applying several augmentation methods. The introduction of generated images solves the problem of insufficient original images and extremely unbalanced sample size and significantly improves the performance of the classifier. Compared with the GAN model that has been widely used, the diffusion model has comparable performance and has the advantages of concise structure and good development prospects. Future improvements in diffusion models will include being able to take into account the biological properties of the image generation logic, better image quality, and faster training speeds.

References

[1] Jin, X., Zou, Y., and Huang, Z 2021 *Information*, 249. https://doi.org/10.3390/info12060249

[2] Eulenberg, P., Köhler, N., Blasi, T., Filby, A., Carpenter, A. E., Rees, P., Theis, F. J., and Wolf, F. A 2017 Reconstructing cell cycle and disease progression using Deep Learning. *Nature Communications*. https://doi.org/10.1038/s41467-017-00623-3

[3] Johnson, J. M., and Khoshgoftaar, T. M 2019 Survey on deep learning with class imbalance. *Journal of Big Data*. https://doi.org/10.1186/s40537-019-0192-5

[4] Shorten, C., and Khoshgoftaar, T. M 2019 A survey on image data augmentation for Deep Learning. *Journal of Big Data*. https://doi.org/10.1186/s40537-019-0197-0

[5] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014 *Generative Adversarial Networks*. arXiv.org. Retrieved September 15, 2022, from https://arxiv.org/abs/1406.2661

[6] Qasim, A. B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J. C., Sekuboyina, A., Kofler, F., Lipkova, J., Li, H., and Menze, B 2021 *Red-gan: Attacking class imbalance via conditioned generation. yet another perspective on medical image synthesis for skin lesion dermoscopy and brain tumor MRI*. arXiv.org. Retrieved September 15, 2022, from https://arxiv.org/abs/2004.10734

[7] Liqaa M.Shoohi1 and J. H. S. 2022 Medico Legal Update. Retrieved September 15, from https://ijop.net/index.php/mlu/article/view/514

[8] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and amp; Ganguli, S 2015 Deep unsupervised learning using nonequilibrium thermodynamics. arXiv.org. Retrieved September 28, 2022, from https://arxiv.org/abs/1503.03585v8

[9] Ho, J., Jain, A., and Abbeel, P 2020 *Denoising Diffusion Probabilistic models*. arXiv.org. Retrieved September 15, 2022, from https://arxiv.org/abs/2006.11239v2

[10] Dhariwal, P., and Nichol, A 2021 *Diffusion models beat gans on image synthesis*. arXiv.org. Retrieved September 15, 2022, from https://arxiv.org/abs/2105.05233

[11] Lucidrains. (n.d.). *Lucidrains/denoising-diffusion-pytorch: Implementation of denoising diffusion*

probabilistic model in Pytorch. GitHub. Retrieved September 19, 2022, from https://github.com/lucidrains/denoising-diffusion-pytorch

[12] Song, J., Meng, C., and amp; Ermon, S 2022 Denoising diffusion implicit models. arXiv.org. Retrieved September 28, 2022, from https://arxiv.org/abs/2010.02502