

Image Synthesis for Sperm Dataset Augmentation using WGAN-GP

Hajjar Ayu Cahyani Kuswardhani
Computer Science Faculty
UPN “Veteran” East Java
Surabaya, Indonesia
21083010044@student.upnjatim.ac.id

I Gede Susrama Mas Diyasa
Computer Science Faculty
UPN “Veteran” East Java
Surabaya, Indonesia
Igsusrama.if@upnjatim.ac.id

Mohammad Idhom
Computer Science Faculty
UPN “Veteran” East Java
Surabaya, Indonesia
idhom@upnjatim.ac.id

Abstract— This research explores the efficacy and limitations of applying a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) to generate synthetic human sperm microscopy images for data augmentation. We assessed the WGAN-GP's performance on a complex, heterogeneous dataset where images contained multiple object types. Despite achieving stable training convergence, the model's output quality was suboptimal, as evidenced by a high Fréchet Inception Distance (FID) score of 134 and qualitative signs of partial mode collapse. The generator struggled to capture the complete morphological diversity of the sperm cells. A second experiment using a dataset pre-sorted into distinct classes (Normal, Abnormal, Non-Sperm) yielded a marked improvement. This approach led to substantially lower FID scores (59.19, 74.92, and 83.56) and exhibited more robust training dynamics. Our findings underscore a critical conclusion: the success of WGAN-GP in this domain is fundamentally tied to the simplicity of the data distribution. We recommend that future efforts leverage class-conditioned models, simplified data structures, and refined generator architectures to achieve high-precision augmentation for medical imaging tasks.

Keywords— *Generative Adversarial Networks (GANs), WGAN-GP, Data Augmentation, Sperm Morphology, Image Synthesis.*



© 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution ShareAlike (CC BY SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).

I. INTRODUCTION (HEADING I)

Sperm morphology analysis is a fundamental component of fertility evaluation, where cell shape and structure are key indicators of fertilization capacity [1][2]. Research in this field relies heavily on the availability of large and diverse image datasets, yet their acquisition is often hindered by data collection difficulties, high biological variability, and technical issues [3]. The limited availability and low quality of microscopic samples [4] often cause deep learning models to suffer from overfitting [3]. While conventional data augmentation methods are incapable of creating new morphological variations, Generative Adversarial Networks (GANs) offer a solution for synthesizing realistic images [5]. Although proven effective, the application of GANs in the biological domain faces challenges such as mode collapse and

high computational demands [6][7]. Nevertheless, GAN-based augmentation has consistently shown significant performance improvements over traditional methods on limited datasets [4].

To address these challenges, this study specifically selects the Wasserstein GAN (WGAN) for its superior stability. By adopting the Earth Mover's Distance (EMD), WGAN effectively mitigates the mode collapse problem often encountered in conventional GANs [8], resulting in more stable gradients suitable for high-quality medical images [9]. The adoption of a gradient penalty (GP) in WGAN-GP further enhances training stability and prevents the issues of vanishing/exploding gradients [10], making it a highly relevant approach for sperm morphology data augmentation.

Numerous studies have implemented GAN variants across various medical imaging domains. For instance, PGGAN has been used to generate brain tumor MRI images [11], MorphoGAN for data with complex morphologies [12], DCGAN-GP for EEG signals converted into images [13], and other approaches for imbalanced blood cell classification [14]. Despite their success, these studies often face limitations such as the need for further validation on downstream tasks [11][13], challenges in evaluating synthetic data quality [12], and high computational complexity [14].

Therefore, this research provides a novel contribution by specifically applying WGAN-GP to a human sperm image dataset. Unlike previous research, which has tended to use conventional augmentation or general GANs, this study is the first to evaluate the capability of WGAN-GP to generate synthetic sperm images as an augmentation baseline. By focusing on the exploration of visual quality, this research will lay the groundwork for assessing the relevance and potential of WGAN-GP for expanding training data in the domain of spermatology.

II. METHOD

This research was conducted through several key stages, beginning with data acquisition and pre-processing, followed by the training of the WGAN-GP model, and concluding with a quantitative evaluation of the results. The complete research workflow is illustrated in Figure 1.



Fig. 1. Research Workflow

A. Data Acquisition

The initial phase of this research involved acquiring data from both primary and secondary sources. Primary data was collected prospectively using an Optilab camera mounted on an OptiLab Advance Plus microscope with a 100x objective lens magnification. Sperm samples were gathered from volunteers under informed consent and prepared using Hematoxylin-Eosin staining to enhance morphological contrast, resulting in approximately 960 images. For the secondary source, the public Sperm Morphology Image Data Set (SMIDS) was utilized [15]. SMIDS, which consists of smartphone-acquired sperm images, served as a comparative benchmark for training a separate model. The class distribution of the labeled SMIDS dataset is presented in Table 1, with representative images of both datasets shown in Figure 2.

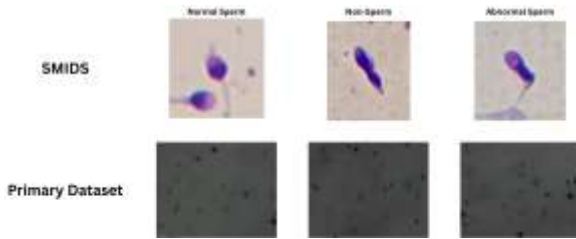


Fig. 2. The Sample of Used Dataset

TABLE I. SMIDS CLASS DISTRIBUTION

Class	Count
Normal	1021
Abnormal	1005
Non-Sperm	974

B. Data Preprocessing

All images first underwent a filtering step to improve quality. This process involved the application of a Band-Pass Filter (BPF) in the frequency domain, which operates by transforming an image into its frequency representation via the Fourier Transform, $F(u,v)$ [16, 17]. The BPF was specifically configured to suppress low-frequency components associated with background illumination and high-frequency components linked to noise [18]. By isolating the mid-range frequencies containing information about sperm structures, this step effectively enhanced the contrast of key morphological features [19]. This filtering operation is mathematically represented by Equation (1), where $H(u,v)$ is the filter's transfer function.

$$G(u,v) = H(u,v) \cdot F(u,v) \quad (1)$$

Following filtering, further standardization was applied. The pixel intensities of all images were normalized to a range of $[-1, 1]$ to promote training stability and align the data distribution with the generator's tanh output layer [20-23]. Finally, all images were resized to a uniform dimension of 128×128 pixels. This standardization is a fundamental architectural requirement for the CNN-based models used in this study, ensuring consistent tensor dimensions for processing [24]. Representative results of this pre-processing stage are shown in Figure 4.

C. Wasserstein GAN with Gradient Penalty (WGAN-GP)

The core of this research's synthetic data generation process was the training of two independent Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) models: one dedicated to the primary dataset and the other to the secondary SMIDS dataset. The WGAN-GP architecture was specifically chosen for its well-documented ability to overcome the training instability and mode collapse issues that frequently plague standard GANs [10]. Its primary advantage stems from replacing the typical divergence metric with the Wasserstein-1 distance (also known as Earth Mover's Distance). This metric provides a more meaningful measure of the distance between the real and generated data distributions, which in turn yields smoother and more reliable gradients for the generator to learn from, a critical factor for stable training [8].

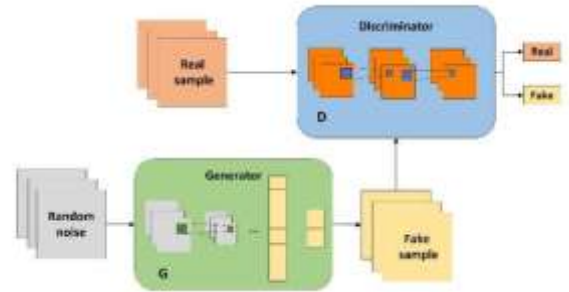


Fig. 3. WGAN's Mechanism Flowchart

The entire mechanism, illustrated in Figure 3, revolves around an adversarial process governed by the critic's loss function. Instead of simply classifying images as real or fake, the critic acts as a regressor, and its main objective is to maximize the score difference between real and generated samples. The key to the model's stability lies in how this is constrained, which is encapsulated in its loss function, shown in Equation (2). The first two terms in this equation represent the critic's effort to push the scores of real data higher than those of fake data, thereby approximating the Wasserstein distance. The third and most critical component is the gradient penalty. This term enforces the necessary 1-Lipschitz constraint by penalizing the critic if the norm of its gradients (with respect to its input) deviates from 1. This penalty is calculated on samples randomly interpolated between pairs of real and generated images. Concurrently, the generator's objective is to produce samples that receive a high score from

the critic, effectively learning to minimize the Wasserstein distance and generate increasingly realistic images. The specific architectures for the Generator and Critic are detailed in Table 2 and Table 3, with all training hyperparameters listed in Table 4.

$$L_D = \mathbb{E}_{\tilde{x} \sim p_g} [D(\tilde{x})] - \mathbb{E}_{x \sim p_r} [D(x)] + \lambda \mathbb{E}_{\tilde{x} \sim p_g} [(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2] \quad (2)$$

TABLE II. HYPERPARAMETER SETTING

Hyperparameter	Value
Epoch	1500
Learning Rate	1e-5
Optimizer	Adam
Beta1	0.0
Beta2	0.9
critic iteration	5
lambda	10

D. Frechet Inception Distance (FID)

The quantitative evaluation of the generated images was performed using the Fréchet Inception Distance (FID). FID is a robust metric that addresses the limitations of pixel-wise comparisons by assessing the quality and diversity of synthetic images in a deep feature space [25]. The underlying theory is that if two image sets are perceptually similar, the statistical distribution of their high-level features should also be similar. The mechanism involves passing both real (r) and generated (g) images through a pre-trained Inception V3 network and extracting feature activations from a hidden layer. These feature sets are then modeled as multivariate Gaussians, and the FID score is calculated as the Fréchet distance between these two distributions using Equation (3). The first term in the equation measures the distance between the mean feature vectors, while the second term (involving the trace, Tr) measures the distance between the covariance matrices, thus capturing differences in both the average features and their variations. For this study, the FID score was calculated independently for each of the two models by comparing the generated data against the corresponding pre-processed real data. A lower FID score indicates a smaller distance between the two distributions, signifying that the generated images are of higher quality and diversity.

$$FID(r, g) = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \quad (3)$$

III. RESULTS AND DISCUSSION

The experiments revealed a significant performance disparity when the WGAN-GP model was trained on the primary versus the secondary dataset. The following sections detail the findings from each experiment and provide a comprehensive discussion of the outcomes.

A. Performance on the Primary Dataset

Although training on the primary dataset achieved stable convergence, as indicated by the loss curves in Figure 4, the model failed to produce high-quality images. The quantitative evaluation yielded a very poor Fréchet Inception Distance (FID) score of 134. This was confirmed by qualitative analysis

in Figure 5, which showed that the generator did not replicate sperm morphology, instead producing repetitive background textures with minimal diversity. This outcome is a clear indicator of partial mode collapse, where the generator avoids learning the complex data distribution and defaults to a simple, easy-to-learn pattern. The root cause was identified as the inherent complexity of the multi-object dataset, where each image contained a heterogeneous mix of normal sperm, abnormal sperm, and debris. Faced with this difficult task, the generator took a "shortcut" by only modeling the simplest consistent feature—the background.

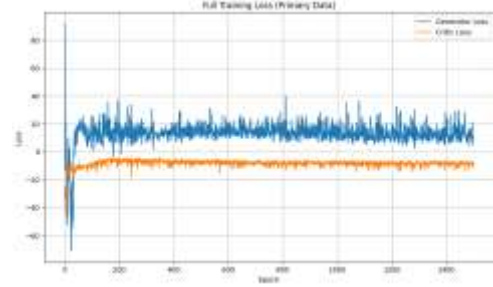


Fig. 4. WGAN-GP Loss Curve for Primary Dataset's

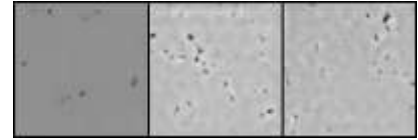


Fig. 5. Generated Image Samples from the Primary Dataset

B. Performance on the Secondary Dataset

To validate the hypothesis that data complexity was the primary issue, the model was trained independently on the homogeneous, class-segregated secondary dataset. This approach yielded vastly superior results, with significantly lower and more favorable FID scores for each class, as detailed in Table 5. This quantitative improvement confirms that simplifying the generator's task from modeling a complex, multi-object distribution to a simpler, unimodal one for each class directly solved the mode collapse problem. The inter-class FID scores also provided insight: the 'Normal' class, being more visually uniform, achieved the best score, whereas the more varied 'Abnormal' and 'Non-Sperm' classes were inherently harder to model. This success is further corroborated by a qualitative inspection of the generated images (Figure 6), which clearly exhibit the distinct and recognizable features appropriate each respective category.

TABLE III. FID SCORES ON SMIDS EACH CLASS

Class	FID
Normal	59.19
Abnormal	74.92
Non-Sperma	83.56

C. Comparative Discussion: Generalist vs. Specialist Models

The results highlight a fundamental dichotomy between a single generalist model and multiple specialist models. The generalist, trained on the complex primary data, failed, whereas the specialists, trained on simplified, homogeneous data, succeeded. This underscores the critical importance of data curation in the success of generative models. Beyond this main hypothesis, several other factors likely contributed to the performance gap. First, the critic may have provided uninformative gradients on the ambiguous primary data. Second, the model's latent space may have had insufficient capacity to encode the dataset's combinatorial complexity. Finally, the training hyperparameters may have been ill-suited for the more challenging loss landscape of the primary dataset. An interplay of these factors offers a more holistic explanation for the observed results.

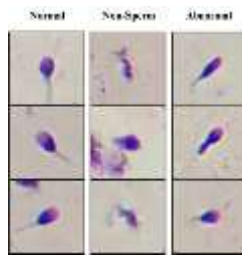


Fig. 6. Generated Image Samples from the Secondary Dataset (SMIDS)

IV. CONCLUSION

This research successfully characterized the performance of the WGAN-GP architecture for the task of synthetic sperm image generation, revealing that its efficacy is fundamentally linked to the complexity of the training data. The study demonstrated that when presented with a heterogeneous, multi-object dataset, the model adeptly learned to replicate the most consistent patterns, resulting in outputs dominated by background textures. Conversely, when trained on homogeneous, class-segregated data, the same architecture demonstrated its capability to generate high-fidelity, diverse images that accurately captured specific morphological features. The primary contribution of this work is, therefore, the clear identification of the optimal conditions for applying a standard WGAN-GP model in this domain. We conclude that while the architecture is technically robust, its potential for high-quality image generation is fully unlocked when the generative task is simplified through meticulous data curation and the use of unimodal training distributions.

The findings of this study should be considered within the context of its specific scope and methodology, which present opportunities for future exploration. First, the results are specific to the standard WGAN-GP architecture as implemented. This study highlights that this implementation is highly effective for homogeneous data distributions. Its scope did not include architectural modifications, such as adding attention mechanisms or residual blocks, which could potentially enhance its performance on more complex, multi-object scenes.

Furthermore, the research was focused on an in-depth characterization of WGAN-GP and did not extend to a comparative analysis against other state-of-the-art generative architectures like StyleGAN. Such models may employ different mechanisms that could yield distinct outcomes on the same datasets. The study also utilized a single set of hyperparameters that proved effective for the class-separated datasets; it is acknowledged that further optimization of these parameters for the more challenging primary dataset could represent an avenue for performance enhancement. Finally, the research employed a direct, end-to-end generative approach, and its scope did not include multi-stage pre-processing pipelines, such as object-detection-based cropping, which could offer an alternative strategy for simplifying the generative task.

ACKNOWLEDGMENT

The author wishes to express sincere gratitude to their academic advisors for their invaluable guidance, mentorship, and insightful feedback throughout the course of this research. Appreciation is also extended to the members of the examination committee for their constructive comments. The author thanks the colleagues within the research group for engaging discussions and support. This work was conducted using the facilities of the author's institution.

REFERENCES

- [1] World Health Organization. WHO laboratory manual for the examination and processing of human semen (6th ed.), 2021.
- [2] Mayo Clinic. "Sperm morphology: What does it mean?" Mayo Clinic, Mar. 2025.
- [3] Shorten, Connor, and Taghi M. Khoshgftaar. "A Survey on Image Data Augmentation for Deep Learning." *Journal of Big Data*, vol. 6, no. 1, 2019, p. 60.
- [4] Frid-Adar, Maayan, et al. "GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification." *arXiv preprint arXiv:1803.01229*, Mar. 2018.
- [5] Generative adversarial network." Wikipedia, June 2024; originally proposed by Ian Goodfellow et al. in 2014.
- [6] Saad, Muhammad Muneeb, Ruairi O'Reilly, and Mubashir Husain Rehmani. "A Survey on Training Challenges in Generative Adversarial Networks for Biomedical Image Analysis." *arXiv preprint arXiv:2201.07646*, Jan. 2022.
- [7] Saad, Muhammad Muneeb, et al. "Early Stopping Criteria for Training Generative Adversarial Networks in Biomedical Imaging." *arXiv preprint arXiv:2405.20987*, May 2024.
- [8] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR.
- [9] Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. *IEEE Transactions on Medical Imaging*, 38(3), 736–746.
- [10] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved Training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30.
- [11] Huang P., Liu X., Huang Y. Data Augmentation For Medical MR Image Using Generative Adversarial Networks. *arXiv 2111.14297* (2021).

- [12] Abdudzimov A., Jo G. Morpho-GAN: Unsupervised Learning of Data with High Morphology using Generative Adversarial Networks. Proc. Korean Computer Information Society Winter Conf. Vol. 28 No. 1 (2020)
- [13] Du X., Ding X., Xi M., Lv Y., Qiu S., Liu Q. A Data Augmentation Method for Motor Imagery EEG Signals Based on DCGAN-GP Network. Brain Sci. 14(4):375 (2024).
- [14] Rana P., et al. Data augmentation with improved regularisation and sampling for imbalanced blood cells classification using Wasserstein divergence GAN, mixup, and nonlinear mixup. PubMed PMID: 36302948 (2022).
- [15] Ilhan, H. (2022). Sperm Morphology Image Data Set (SMIDS) (Version 1) [Data set]. Mendeley Data. <https://doi.org/10.17632/6xvdhc9fyb.1>
- [16] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2018.
- [17] K. Jain, Fundamentals of Digital Image Processing. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.
- [18] W. K. Pratt, Digital Image Processing: PIKS Inside, 3rd ed. New York, NY, USA: John Wiley & Sons, 2001.
- [19] J. C. Russ, The Image Processing Handbook, 6th ed. Boca Raton, FL, USA: CRC Press, 2011.
- [20] T. M. Nguyen, T. H. Le, and H. G. Lim, "A Comparative Study on Data Preprocessing in Deep Learning-based Medical Image Analysis," Applied Sciences, vol. 13, no. 14, p. 8103, Jul. 2023.
- [21] A. Kirillov et al., "Segment Anything," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 10399-10409.
- [22] Z. Chen, Y. Zhang, Y. Zhang, and X. Liu, "GAN-Based Image Generation: A Survey and Taxonomy," IEEE Trans. Neural Netw. Learn. Syst., pp. 1-22, 2024, doi: 10.1109/TNNLS.2024.3387820.
- [23] J. O. Öfele, "On the effective design of generative adversarial networks," PhD Dissertation, Dept. Comput. Sci., Technical University of Munich, 2023.
- [24] A. S. Lundervold and A. L. T. A. T. H. Anonsen, "An overview of deep learning in medical imaging," in Deep Learning in Medical Imaging and Multimodal Learning for Clinical Decision Support, A. S. Lundervold and A. L. T. A. T. H. Anonsen, Eds. Cham: Springer International Publishing, 2023, pp. 1-23.
- [25] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2017.
- [26] Diyasa, I. Gede Susrama Mas, et al. "Data Augmentation of Breast Ultrasound Images Using Wasserstein Generative Adversarial Networks." (2024).
- [27] I. Gede, Dwi Arman Prasetya, Hajjar, and C. Halim, "Detection of Abnormal Human Sperm Morphology Using Support Vector Machine (SVM) Classification," Information Technology International Journal, vol. 2, no. 2, pp. 57-63, Nov. 2024, doi: <https://doi.org/10.33005/itij.v2i2.36>.
- [28] I. G. S. Mas Diyasa, W. S. J. Saputra, A. A. N. Gunawan, D. Herawati, S. Munir, and S. Humairah, "Abnormality Determination of Spermatozoa Motility Using Gaussian Mixture Model and Matching-based Algorithm", J Robot Control (JRC), vol. 5, no. 1, pp. 103-116, Jan. 2024.