

Contents lists available at www.journal.unipdu.ac.id

Register

Journal Page is available to www.journal.unipdu.ac.id/index.php/register



Research article

Data Augmentation of Sperm Images Using Generative Adversarial Networks (WGAN-GP)

I Gede Susrama Mas Diyasa ^{a,*}, Hajjar Ayu Cahyani Kuswardhani ^{b,*}, Mohammad Idhom ^c,
Prismahardi Aji Riyantoko ^d, Deshinta Arrova Dewi ^e

^{a, b, c} Department of Data Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Jl. Raya Rungkut Madya, Kec. Gunung Anyar, Surabaya 60294, Indonesia

^d Department of Information and Communication Systems, Okayama University, 3-1-1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama 700-8530, Japan

^e Center for Data Science and Sustainable Technologies INTI International University, Persiaran Perdana BBN, Putra Nilai, 71800 Nilai, Negeri Sembilan, Malaysia

email: ^{a,*} igsusrama.if@upnjatim.ac.id, ^b 21083010044@student.upnjatim.ac.id, ^c idhom@upnjatim.ac.id, ^d pnai2m3s@s.okayama-u.ac.jp, ^e deshinta.ad@newinti.edu.my

* Correspondence

ARTICLE INFO

Article history:

Received September 9th, 2025

Revised January 6th, 2026

Accepted February 6th, 2026

Available online February 14th, 2026

Keywords:

Augmentation;
Sperm Image;
Reproductive Health;
WGAN-GP;
Innovation;

Please cite this article in IEEE style as:

I. G. S. M. Diyasa, H. A. C. Kuswardhani, M. Idhom, P. A. Riyanto and D. A. Dewi, "Data Augmentation of Sperm Images Using Generative Adversarial Networks (WGAN-GP)," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 12, no. 1, pp. 1-10, 2026.

ABSTRACT

This study analyzes the use of WGAN-GP for data augmentation in the analysis of sperm morphology. WGAN-GP has been the focus in this study for generating sperm microscopy images, which in turn aims to mitigate the problem of data scarcity in medical imaging. A heterogeneous dataset with mixed object categories was initially employed, leading to an FID score of 134, which in turn reflected a high incidence of mode collapse. For this reason, the dataset was divided into subcategories of Normal, Abnormal, and Non-Sperm identifications, with the scores of the subcategories being 59.19, 74.92, and 83.56, respectively, and showing better balanced model stability. This study's primary contribution is the use of WGAN-GP for the first time for sperm image data augmentation and the generation of more realistic synthetic images. Furthermore, this study illustrates the first understanding of the intricacies of data distribution's complexity and its effect on the model's performance, indicating the possibility of improvement using class-based techniques and sophisticated architectures for the generator. The innovation of this study is the application of WGAN-GP to sperm morphology datasets, improving image quality and the stability of the results, coupled with extensive model performance analysis and providing a further understanding of the field of medical image data augmentation.

Register with CC BY NC SA license. Copyright © 2026, the author(s)

1. Introduction

The study and evaluation of reproductive health hinge *prima facie* on the analysis of sperm morphology. The ability of spermatozoa to potentially fertilize an egg is reflected in the metrics of their morphology. [1][2]. This domain of research requires sufficiently large and varied data sets pertaining to sperm image morphology. Such data sets are difficult to acquire as a result of challenges pertaining to sample collection, inherent biological variability, and the limitations of image capture [3][4].

Microscopic domains, such as imaging sperm, face the challenges of not having enough well-labeled samples and often having poor-quality images [5][6]. Moreover, due to the high morphological variability of sperm within and across individuals, deep learning models that are trained on narrow data distributions, can easily overfit [3]. More standard forms of data augmentation, such as rotation, flipping, and cropping, do not generate any fundamentally new morphological structures and do not enhance the resolution of blurry images. Conversely, Generative Adversarial Networks (GANs) show potential in image generation, in which the new images can be as close a match as necessary to the actual

images [7]. Other studies have shown that while optimizing the essential biological parameters of the problem, GANs can provide realistic images of 2D biological cells. However, there are unique challenges to the biological applications of GANs, such as high computation time, and the problems of mode collapse, which occurs when the generator produces insufficient diversity [8][9]. Other studies also in the medical imaging domain support the conclusion that when augmented with GANs, models exhibit significant improvement in performance as opposed to when traditional augmentation methods are applied, especially when the dataset used for training is limited [4]. This illustrates the importance of data augmentation as a means of addressing insufficient training data.

WGANs are considered. WGANs have good stability and synthesis quality relative to other GANs. A typical GAN has a lack of diversity, where WGANs are better because their loss function incorporates the Earth Mover's distance and can be adjusted to satisfy a Lipschitz constraint. This means WGANs can avoid mode collapse better than classical GANs [10]. This approach creates smoother and more stable gradients, which are important for high-quality images [11]. Further improving stability, WGANs with Gradient Penalty (WGAN-GP) avoid issues of exploding/vanishing gradients [12][13]. Thus, the WGAN-GP framework is highly applicable for the augmentation of sperm morphology datasets, improving the size and diversity of the dataset.

Multiple subsequent research studies have applied WGAN or WGAN-GP models within the sperm field and beyond. For instance, Huang et al. (2022) studied GAN-based data augmentation and its applicability in the medical field for the analysis of brain tumors in MRI images. They combined PGGAN (Progressive Growing GAN) with a Structural Similarity (SSIM) loss and created a set of realistic and varied images of FLAIR (Fluid Attenuated Inversion Recovery) brain tumors scans, all sized 256×256. Their PGGAN-SSIM model was noted for surpassing other variants of GANs in FID (Fréchet Inception Distance) and MS-SSIM (Multi-scale Structural Similarity) comparisons. Their study [14] was, however, confined to FLAIR type brain tumor images and did not address other types of MRI images. Also, they did not address how the synthetic images contributed to the detection of the tumors, thus further evaluations were needed.

Another study by Azamat and GeunSik (2020) suggested the development of Morpho-GAN, a model that falls within the category of unsupervised GANs and is aimed at producing and maintaining computational efficiency. The results showed that Morpho-GAN was able to generate realistic and competitive data compared to GAN models of its time and required substantially less training [15]. The major challenges remained in assessing the generated data's quality and diversity, particularly in terms of balanced modes and variability.

Nonetheless, in a separate instance, Du Xiuli (2024) employed a DCGAN-GP based approach for data augmentation of motor imagery EEG signals. EEG signals were converted to a collection of 2D time-frequency images, on which a DCGAN-GP model was trained to produce a collection of synthetic samples. During the training of the classifier, real samples were mixed with the synthetic ones in different predetermined ratios. It was concluded that the DCGAN-GP classifier, more so than others, improved the classification of the EEG signals across different subjects [16]. This shows the potential for this approach to supplement scarce research data; however, more work must be done to determine the extent to which the synthetic data can be applied to other datasets.

Likewise, in 2022, with the aid of regularization and mixed sampling strategies, data augmentation for the classification of blood cells in a dataset with class imbalance was attempted by Rana et al. In their approach, the generation of diverse synthetic images was achieved by means of the combination of Wasserstein divergence GAN, mixup, and nonlinear mixup. The training process incorporated minority class sampling. For underrepresented classes, this approach enhanced class representation for the classifier, which improved the classification of highly imbalanced datasets with high imbalance and increased the generalization of the model to novel data [17]. The intricate and expensive nature of the approach, however, leaves much to be desired in terms of cross-dataset and cross-paradigm validation in order to corroborate the approach.

From several previous studies, in the field of medical image data augmentation, various Generative Adversarial Networks (GAN) approaches have been carried out to overcome the problem of limited data. However, this study introduces several significant novel aspects that distinguish it from previous studies, highlighting the specific application of sperm morphology, the use of WGAN-GP with

better stability and quality, increased variability of synthetic data, comprehensive evaluation of model performance, incorporation of augmentation techniques that have better performance, and thorough validation of the resulting synthetic data, so that this study can make a significant contribution to medical image data augmentation, especially in sperm morphology analysis. This study also paves the way for the application of WGAN-GP in other medical contexts that require high-quality and diverse data to improve the accuracy and reliability of the analysis model. Therefore, this study aims to overcome some limitations in sperm morphology analysis by introducing a new approach using the Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP). The discussion of this research begins with a background that presents the novelty of this research, then discusses the research methods used, such as sperm data acquisition, data preprocessing and augmentation processes, trials and results, and the conclusions of this research.

2. Materials and Methods

The different stages of the research include gathering data, data preprocessing, training the WGAN-GP model, and evaluating model performance. These stages were built to ensure reliability and reproducibility of the results of the experiments. Figure 1 shows the flow of the study.

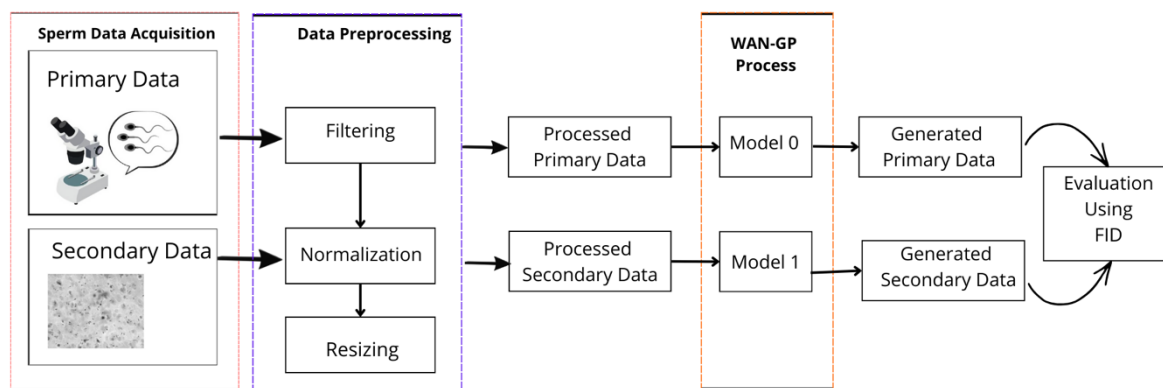


Fig. 1. Examples of Primary Dataset

In Figure 1, we can see the sperm data acquisition, data preprocessing, synthetic data generation, and quality assessment, which are the specific phases of the research process. Primary and secondary sperm image datasets are the two types of sperm image datasets used. Primary data are the images obtained directly from the acquisition process via microscopy. Secondary data are sourced from existing sperm images. Once the acquisition stage is complete, both types of data enter the data preprocessing stage. The primary activities at this stage are filtering, normalization, and resizing. Preprocessing of data helps decrease noise and visual distractions in the images, standardize pixel intensity values across the images, and achieve a consistent data distribution. The outcome of the resizing of images helps the researcher ensure that all images are in the correct format and of high quality to serve as inputs to the learning model. After primary and secondary data have been processed, the two types of data are separated and used as inputs to two different WGAN-GP models, Model 0 for primary data and Model 1 for secondary data. Each model was designed to learn to replicate the original data and generate realistic sperm images under controlled conditions.

Images of synthetic sperm were produced for both types of data, primary and secondary. The last step in this process involved assessing the range and quality of the images generated. In terms of metrics for quantitative assessments, the Fréchet Inception Distance (FID) was used, which analyzes the distance between the distributions of real versus synthetic images. The diversity and realism of the images were visually evaluated for the purposes of qualitative assessments. These two methods were combined to evaluate the performance of the WGAN-GP model for the generation of synthetic data for augmentation purposes.

2.1. Sperm Data Acquisition

The initial phase of this research entails gathering data from two distinct categories: primary and secondary sources. For primary data, prospective data collection was undertaken using an Optilab camera connected to an OptiLab Advance Plus microscope. Sperm samples were imaged at a microscopic level using a 100x objective lens. These volunteer samples were ethically obtained from the

research participants after signing informed consent. Prior to image acquisition, the sperm samples underwent a dilution process and staining using the Hematoxylin-Eosin method to enhance contrast and clarify the morphological structure of the sperm under the microscope.

The main data were collected from 12 volunteer participants at the Pathology Laboratory of the Faculty of Medicine at UPN Veteran Jawa Timur, the volunteers provided samples to be analyzed in the laboratory, in total, 12 volunteers provided 1012 sperm image samples for the study. Data of a secondary nature were acquired from preexisting sources which served to assist in the analysis and comparison with the primary data. As a result, approximately 1012 sperm image samples were obtained from this stage and became the primary dataset for the research. All the data collected in this study can be accessed via the link <http://data-Sperm>, where the images can be viewed and further analyzed. As an example, Figure 2 shows one of the sperm images used in this study, providing a visual representation of the data collected.

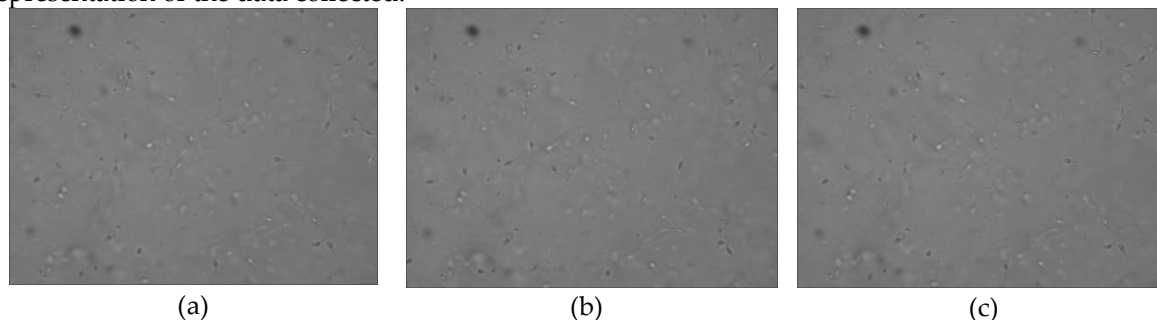


Fig. 2. Examples of Primary Dataset

Secondary data acquisition were obtained from existing sources, such as public databases, scientific publications, and previous research. In this study, the secondary data used were SMIDS (Sperm Morphology Image Data Set). SMIDS is a public dataset consisting of sperm cell images captured using a smartphone acquisition method [18]. The images from the SMIDS dataset were used in this study as a benchmark comparison. SMIDS images have been stored in the RGB color space, and the developer has given the label distributions as Normal with 1021 data, Abnormal with 1005 data, and Non-Sperm with 974 data. Figure 3 also shows examples of images collected from the SMIDS dataset.

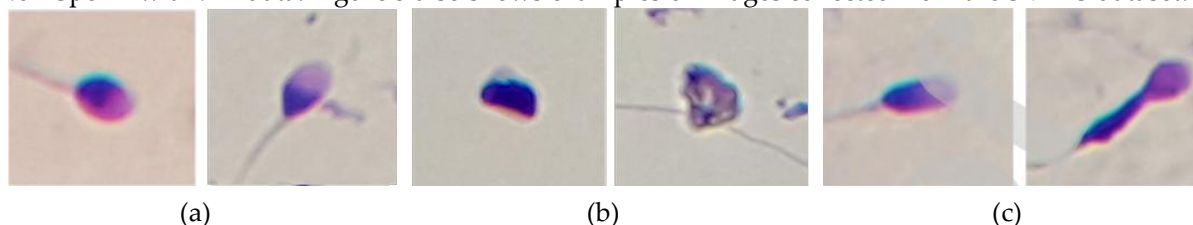


Fig. 3. Examples of SMIDS Respective Class: (a) Normal; (b) Abnormal and (c) Non-Sperm

2.2. Data Preprocessing

Before the sperm images were used for analysis or model training, a series of preprocessing steps were applied to improve image quality, standardize their format, and minimize irrelevant variability. In this study, the preprocessing workflow comprised three main procedures: filtering, pixel-intensity normalization, and resizing. The first step involved filtering with a Band-Pass Filter (BPF). Conceptually, a BPF operates in the frequency domain, where an image is transformed from its spatial (pixel) representation into its frequency representation using techniques such as the Fourier Transform [19]. This transform decomposes the image signal into a sum of sinusoidal components with varying frequencies and amplitudes [20]. Low frequencies typically represent gradual intensity changes such as, background illumination, whereas high frequencies correspond to abrupt intensity variations, such as, random noise or very fine details [21]. A BPF is designed to retain only a specific range of frequencies deemed relevant. Mathematically, if $F(u,v)$ denotes the Fourier Transform of an image $f(x,y)$ and $H(u,v)$ represents the filter's transfer function, then the filtered image in the frequency domain, $G(u,v)$, can be expressed by Equation (1).

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

In an ideal BPF, the transfer function $H(u,v)$ equals 1 for frequencies within the desired passband and 0 outside it. The frequency range was chosen to capture the essential structural information of

sperm cells. Consequently, very low and very high frequency components were suppressed, effectively isolating mid-range frequencies that carry the most relevant features [22]. This process yields cleaner images with enhanced contrast of critical features—such as edges and textures of the sperm head, because the corresponding frequencies are preserved or even accentuated.

Next, pixel-intensity normalization was performed. Specifically, each image's intensity values were scaled to the range $[-1, 1]$. This practice is standard in modern deep learning to ensure computational stability and efficiency [23]. Symmetric normalization around zero helps maintain well-distributed activations across network layers, which is crucial for preventing the vanishing-gradient problem, where gradients become too small during backpropagation and hinder learning in the earlier layers [24]. This range alignment is particularly important for WGANs, whose generator output typically uses a hyperbolic tangent (tanh) activation producing values within $[-1, 1]$ [25]. Matching the distribution of real images to the generator's output range enables the critic (discriminator) to make more direct and stable comparisons, freeing it from learning an additional rescaling task and allowing it to focus on evaluating image quality and providing more informative gradients to the generator [26].

Finally, all images were resized to a uniform resolution of 128×128 pixels. This standardization is an architectural requirement for convolutional neural network (CNN)-based models such as WGAN, which process data as tensors with fixed dimensions. Convolution and pooling operations rely on consistent spatial sizes, and the transition to fully connected layers at the network's end requires input vectors of constant length to function correctly [27][28]. By standardizing the image size, the dataset can be processed consistently within the predefined network architecture. The results of this preprocessing stage are illustrated in Figure 4.

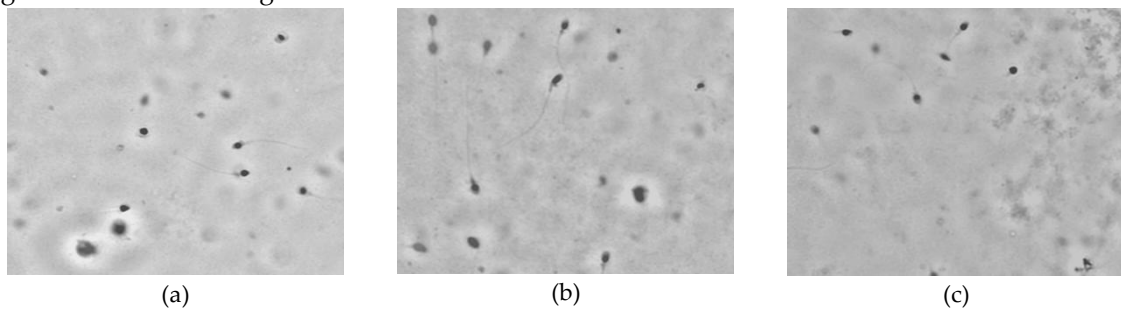


Fig. 4. Example of Preprocessing Results

2.3. Wasserstein GAN with Gradient Penalty

The Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) is a new variation of the original GAN framework that attempts to solve the problems of training instability and mode collapse [10]. The GAN framework is built on the concept of Wasserstein, which describes the amount of “effort” needed to transform the distribution of generated data into the distribution of real data, as defined in Equation (2). The computation of these metrics is difficult, and to overcome this problem WGAN uses the Kantorovich-Rubinstein duality. With this the Wasserstein Distance can be reformulated as shown in Equation (3), where the supremum is over the set of all 1-Lipschitz functions. Under this scenario, the Critic Network is trained to approximate one such 1-Lipschitz function so as to provide better and more useful gradients to the Generator.

$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|] \quad (2)$$

$$W(P_r, P_g) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{x \sim P_r} [f(x)] - \mathbb{E}_{x \sim P_g} [f(x)] \right) \quad (3)$$

The WGAN-GP mechanism is illustrated in Figure 5. The Generator receives a random noise vector and converts it into a synthetic sample, while the Critic evaluates both real samples from the dataset and generated samples from the Generator. Its outputs scores for both types of data, which, together with the gradient penalty form the loss used to update the Critic's parameters. At the same time, the Generator updates its own parameters so that it progressively produces more realistic data.

To stably enforce the 1-Lipschitz constraint, WGAN-GP introduces a gradient penalty into the Critic's loss function, defined in Equation (4), where interpolated samples between real and generated data are used as in Equation (5). The Generator's loss, shown in Equation (6), minimizes the negative of the Critic's average score on fake data, thereby encouraging it to generate samples that the Critic rates highly.

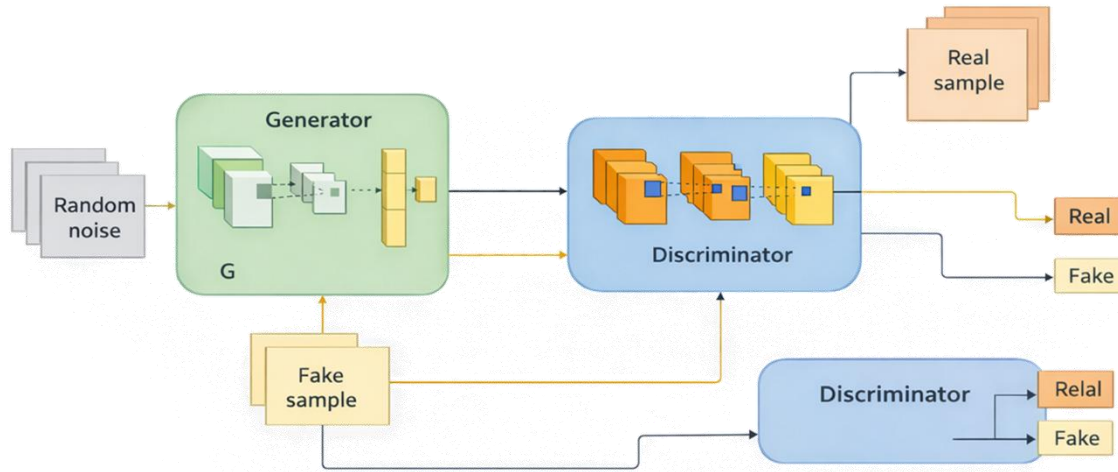


Fig. 5. WGAN-GP Mechanism

In this study, two independent WGAN-GP models were trained to generate synthetic sperm images. Model 0 was dedicated to learning the distribution of the preprocessed primary dataset, while Model 1 focused on the preprocessed secondary dataset. Each model involved an adversarial interplay between a Generator and a Critic, where the Generator attempted to synthesize samples increasingly similar to the real data, while the Critic learned to distinguish between genuine and generated images. During training, for every update of the Generator over a set number of epochs, the Critic was updated multiple times. Ultimately, this led to two trained Generators that could create new samples statistically representative of the training set.

$$L_D = \mathbb{E}_{\tilde{x} \sim P_g}[D(\tilde{x})] - \mathbb{E}_{x \sim P_r}[D(x)] + \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \quad (4)$$

$$\hat{x} = \epsilon x + (1 - \epsilon)\tilde{x}, \text{ dengan } \epsilon \sim U[0, 1] \quad (5)$$

$$L_G = -\mathbb{E}_{\tilde{x} \sim P_g}[D(\tilde{x})] \quad (6)$$

Both implementations utilized deep convolutional structures as the Generator and Critic, with the breakdown of their structures shown in Table 1 and Table 2, as well as the training parameters in Table 3. The generator followed a decoder structure with multiple stages of upsampling, beginning with a latent vector that was projected through a fully connected layer, followed by a number of transposed convolutional layers with Batch Normalization for gradient stability and LeakyReLU for the non-linearity.

Table 1. Generator Architecture

Layer	Kernel size	Output Shape
z (input)	-	128
Linear + Reshape	-	1024 x 8 x 8
ConvTranspose2d + BN + LeakyRelu	5 x 5	512 x 16 x 16
ConvTranspose2d + BN + LeakyRelu	5 x 5	256 x 32 x 32
ConvTranspose2d + BN + LeakyRelu	5 x 5	128 x 64 x 64
ConvTranspose2d + BN + LeakyRelu	5 x 5	64 x 128 x 128
ConvTranspose2d, Tanh	5 x 5	1 x 128 x 128

The output layer used a tanh activation function to constrain output pixels to the interval of $[-1, 1]$, which is in accordance with the normalized distribution of the input. The critic, following best practices for WGAN-GP, did not contain batch normalization in order to maintain the validity of the gradient penalty. The Critic serves as an encoder that downsamples input to extract features and

produce a single scalar output. No activation function is used in the output layer so that it can function as a regressor, which is a tenet of the Wasserstein GAN framework.

Table 2. Critic Architecture

Layer	Kernel size	Output Shape
Input Image	-	1 x 128 x 128
Conv2d + LeakyRelu	5 x 5	64 x 64 x 64
Conv2d + LeakyRelu	5 x 5	128 x 32 x 32
Conv2d + LeakyRelu	5 x 5	256 x 16 x 16
Conv2d + LeakyRelu	5 x 5	512 x 16 x 16
Flatten + Linear	-	1

Table 3. Hyperparameters

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

3. Results and Discussion

Monitoring of the WGAN-GP training process on the primary dataset revealed that the model was able to reach a relatively stable convergence. As illustrated in Figure 6, the generator and critic loss curves gradually moved toward an equilibrium after an initially volatile phase, which can be interpreted as a positive sign of the framework's technical soundness. Nevertheless, quantitative evaluation of the generated images produced a high Fréchet Inception Distance (FID) score of 134. This finding was corroborated by qualitative inspection of the synthetic images (Figure 7), which showed that the model had not yet fully succeeded in reproducing spermatozoa morphology. Instead, the outputs tended to exhibit repetitive background textures and very limited structural diversity.

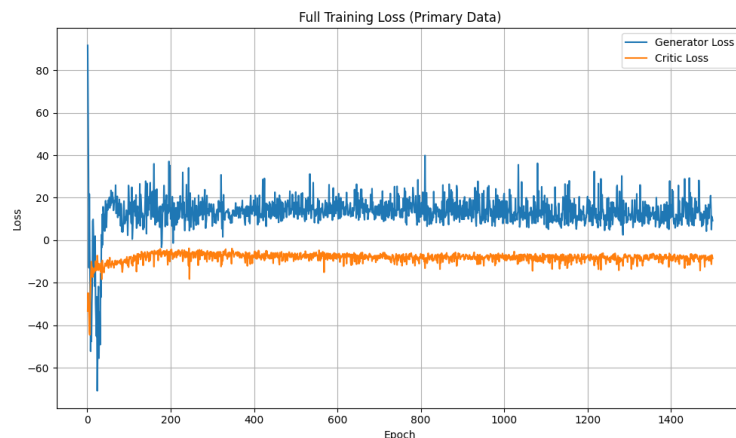


Fig. 6. Training Loss on Primary Data

This kind of behavior is a form of partial mode collapse, where the generator learns only a small set of data distributions and becomes “stuck” at a few simple modes. The most apparent reason is the complexity of the primary dataset; it has a mixture of several objects in one scene, specifically normal and abnormal spermatozoa with cellular debris, where the generator has a much harder time learning not just the appearance of every object but also the probabilities of all objects occurring together. Faced with such complexity, the generator has pretty much taken a shortcut whereby it just paints a consistent background rather than the elaborate mixtures of objects.

To test this hypothesis about data complexity, a comparative training experiment was conducted on a secondary dataset with a multi-class structure. The model was trained separately on more homogeneous subsets of data (Normal, Abnormal, and Non-Sperm). This approach yielded far superior performance, with significantly lower FID scores for each class. The Fréchet Inception Distance (FID) score assesses how similar real and generated images are for each class in the SMIDS dataset. In the

Normal class, the score is 59.19, which indicates that the real and generated images are fairly similar. For the Abnormal class, the score is 74.92, suggesting that the images are less similar than in the Normal class. The Non-Sperm class scored 83.56, suggesting the greatest divergence in quality, indicating lower image quality or a wider divergence. Compared to other classes, this class has a higher discrepancy between real and generated images. These FID scores demonstrate the effectiveness, or limitations, of the generative models in creating realistic sperm images. The synthetic images generated for these secondary datasets are shown in Figure 3. These results provide strong quantitative support for the hypothesis: by training on separate homogeneous classes, the complexity of the task faced by the generator is drastically reduced. Rather than being forced to model the joint distribution of multiple objects, the generator can focus on learning a simpler, unimodal distribution for each class.

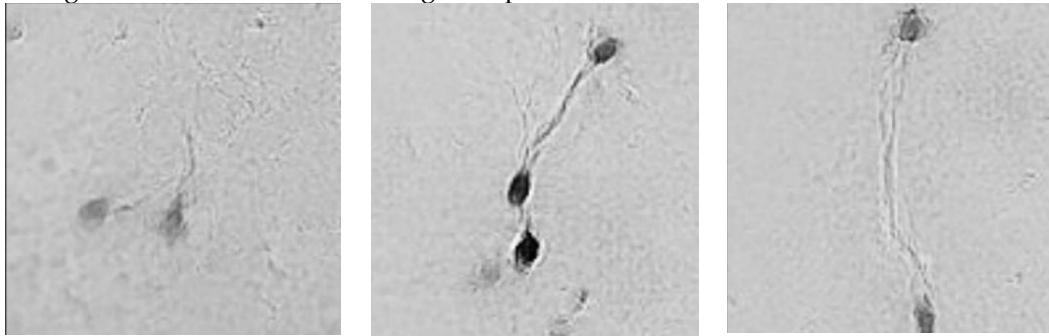


Fig. 7. Examples of Synthetic Data Generated from Primary Data

With fewer challenges it is easier to address the core concerns behind the mode collapse witnessed in the previous trial. The class-wise difference in FID also provides further insight. The Normal class likely had the lowest FID because of its more intra-class homogeneity. The other two classes, “Abnormal” and “Non-Sperm,” which inherently exhibit greater visual diversity and more variabilities in their structures or in artefacts, were harder to model, yielding higher FID scores.

A broad comparison shows a primary difference between generalist and specialist approaches. The generalist framework, a result of sparse multi-object primary dataset training, had difficulty assimilating the complex data distribution, which resulted in poor image quality and convergence. On the other hand, specialist models trained on class-specific subsets showed that task simplification on cleaner, homogeneous sub-distribution, obtains amplified gains in perceptual quality. This reaffirms the intrinsic value of data curation in the success of generative models. Apart from the primary hypothesis of data heterogeneity, a number of other factors could have contributed to the performance gap. First, the critic network may not have supplied truly helpful gradients when trained on the ambiguous primary dataset and instead relied on the low-level textures proxies for “authenticity,” which supplied unhelpful learning signal to the generator. This may have been more problematic in relation to the missing spatial, structural, and higher level features that the latent space was unable to fully encode due to the combinatorial complexity of multi-object scenes. The generator was thus left to perform a lossy compression. Finally, the training hyperparameters, unbalanced in relation to the text conditioned on the more simple landscapes, in relation to the primary data, were too aggressive for the multi modal terrain, may have stunted optimal solution yield.

Taken together, the interaction of weak gradient signals, constrained model capacity, and suboptimal training dynamics provides the best explanation for the performance difference between the two models.

4. Conclusion

This study originally aimed to establish a WGAN-GP model to create high-quality synthetic microscopic images of spermatozoa for data augmentation. The results documented in previous chapters show that this objective was only partially achieved, yet it produced several important insights. The WGAN-GP model demonstrated stable training convergence, as initially planned; however, the resulting images showed that their perceptual quality was highly dependent on the characteristics of the input data. The primary dataset contained multiple object types within each image, and its complexity and heterogeneity were a key factor in impeding the achievement of output diversity and high quality, this was evidenced by high FID scores and mode collapse. In contrast, the WGAN-GP model’s performance

on the streamlined homogeneity of the secondary data set, as evidenced by the data outcome, supports the evaluation findings. Overall, the models from this study have shown that WGAN-GP can be implemented as a stable training dynamic but will have a limited data set as a guiding principle of how to generate good images.

Applying the conclusions drawn from this study to meaningful future enhancements and applications presents opportunities in many directions. From a modeling perspective, building refined versions of the WGAN-GP architecture and then building on other more sophisticated frameworks could prove beneficial. Increasing network depth or width could strengthen feature learning, while integrating Residual Blocks and Self-Attention mechanism may improve gradient stability and spatial dependencies. Other layer-based regularization strategies could be employed to counter mode collapse, such as Layer Normalization as a more stable substitute for the gradient penalty and Consistency Regularization (CR) may also reinforce the critic's learned features. On a more basic level, future studies may be directed toward more sophisticated techniques, including the use of the StyleGAN architecture, which offers more nuanced control over the generation of features. Outside of architecture, other preprocessing techniques, such as object-detection based cropping of the base dataset, may serve to further simplify the generative task.

Considering applications, the generator models trained on the secondary dataset demonstrate practical potential as targeted data-augmentation tools. These generators can be utilized to augment deficient classes in training datasets of classification models, thereby enhancing the robust accuracy of automating diagnostic systems. The success observed with tailored datasets indicates that generative modeling can serve as a valuable support tool in medical research, particularly in domains where annotated data are limited.

Author Contributions

I. G. S. M. Diyasa: Conceptualization, formal analysis, methodology, resources, supervision, validation, writing – original draft, and writing – review & editing. H. A. C. Kuswardhani: Data curation, formal analysis, investigation, project administration, software, supervision, and writing the original draft. M. Idhom: Conceptualization, funding acquisition, methodology, software, supervision, and writing – original draft. P. A. Riyantoko: Investigation, methodology, supervision, validation, and visualization. D. A. Dewi: Funding acquisition, investigation, methodology, supervision and writing – original draft.

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

- [1] World Health Organization, WHO Laboratory Manual for the Examination and Processing of Human Semen, 6th ed. Geneva, Switzerland: WHO, 2021.
- [2] Mayo Clinic, "Sperm morphology: What does it mean?" Mayo Clinic, Mar. 2025.
- [3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019.
- [4] I. G. S. Mas Diyasa, I. K. E. Purnama, and H. M. Purnomo, "A new method to improve movement tracking of human sperms," *IAENG Int. J. Comput. Sci.*, vol. 45, no. 4, pp. 1–9, 2019.
- [5] M. Frid-Adar et al., "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *arXiv preprint, arXiv:1803.01229*, Mar. 2018.
- [6] 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*, vol. 5, no. 1, pp. 103–116, Jan. 2024.
- [7] Wikipedia, "Generative adversarial network," Jun. 2024.
- [8] M. M. Saad, R. O'Reilly, and M. H. Rehmani, "A survey on training challenges in generative adversarial networks for biomedical image analysis," *arXiv preprint, arXiv:2201.07646*, Jan. 2022.
- [9] U. Sirisha, C. K. Kumar, S. C. Narahari, and P. N. Srinivasu, "An iterative PRISMA review of GAN models for image processing, medical diagnosis, and network security," *Comput. Mater. Continua*, vol. 82, no. 2, pp. 1757–1810, 2025.
- [10] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017.

- [11] I. G. S. Mas Diyasa et al., "Improving classification accuracy of breast ultrasound images using Wasserstein GAN for synthetic data augmentation," *J. Robot. Control*, vol. 6, no. 4, pp. 1962–1971, Jul. 2025.
- [12] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *IEEE Trans. Med. Imaging*, vol. 38, no. 3, pp. 736–746, 2018.
- [13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [14] P. Huang, X. Liu, and Y. Huang, "Data augmentation for medical MR image using generative adversarial networks," *arXiv preprint, arXiv:2111.14297*, 2021.
- [15] A. Abduazimov and G. Jo, "Morpho-GAN: Unsupervised learning of data with high morphology using generative adversarial networks," in *Proc. Korean Comput. Inf. Soc. Winter Conf.*, vol. 28, no. 1, 2020.
- [16] X. Du, X. Ding, M. Xi, Y. Lv, S. Qiu, and Q. Liu, "A data augmentation method for motor imagery EEG signals based on DCGAN-GP network," *Brain Sci.*, vol. 14, no. 4, p. 375, 2024.
- [17] P. Rana et al., "Data augmentation with improved regularisation and sampling for imbalanced blood cells classification using Wasserstein divergence GAN, mixup, and nonlinear mixup," 2022.
- [18] H. Ilhan, "Sperm morphology image data set (SMIDS)," *Mendeley Data*, ver. 1, 2022, doi: 10.17632/6xvdhc9fyb.1.
- [19] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2018.
- [20] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.
- [21] W. K. Pratt, *Digital Image Processing: PIKS Inside*, 3rd ed. New York, NY, USA: Wiley, 2001.
- [22] J. C. Russ, *The Image Processing Handbook*, 6th ed. Boca Raton, FL, USA: CRC Press, 2011.
- [23] T. M. Nguyen, T. H. Le, and H. G. Lim, "A comparative study on data preprocessing in deep learning-based medical image analysis," *Appl. Sci.*, vol. 13, no. 14, p. 8103, Jul. 2023.
- [24] A. Kirillov et al., "Segment Anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 10399–10409.
- [25] 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.
- [26] J. O. Öfele, "On the effective design of generative adversarial networks," Ph.D. dissertation, Dept. Comput. Sci., Technical Univ. Munich, Munich, Germany, 2023.
- [27] 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*. Cham, Switzerland: Springer, 2023, pp. 1–23.
- [28] D. Samal, P. Agrawal, V. Madaan, and V. Goyal, "Efficient deepfake detection using convolutional neural network," in *Proc. Int. Conf. Intelligent Control, Computing and Communications (IC3)*, 2025, doi: 10.1109/IC363308.2025.10957581.