BOOSTING CLASSIFICATION ACCURACY OF FERTILE SPERM CELL IMAGES LEVERAGING CDCGAN

Dipam Paul[1,2], **Alankrita Tewari**[1], **Jiwoong Jeong**[2], **Imon Banerjee**[2,3]
[1]Department of Electronics Engineering, KIIT University, Bhubaneswar, India
[2]Department of Biomedical Informatics, Emory School of Medicine, Atlanta, GA, USA
[3]Department of Radiology, Emory School of Medicine, Atlanta, GA, USA
{dipampaul17,alankritat15}@gmail.com
{jiwoong.jason.jeong,imon.banerjee}@emory.edu

ABSTRACT

Drawing inferences from a spermatozoon (Sperm Cell) image based on its morphology is ubiquitous, challenging, and of substantial practical interest. In the present study, we endeavour to deconstruct and demonstrate a framework to distinguish between the binary classes, which constitutes 'Good' (Fertile) and 'Bad' (Infertile) Sperm Cell images. We have selected the DenseNet121 architecture to train our model for this task, the reason for which is examined in Section 2.3. Furthermore, Conditional Deep Convolutional Generative Adversarial Networks (cDCGAN) was used to tackle the minority Class imbalance problem, which was heavily prominent in the dataset chosen for this task as seen in Section 2.2. We have hand-picked numerous statistical inferential tests and metrics to validate our model to accentuate the reliability of the obtained results, thus finally formulating and delineating a table based on the respective 'Quality Scores' of the test samples provided. With the cDCGAN training data augmentation, the test-set accuracy was recorded to be 86.2%, while the model without cDCGAN scored only 24.3%.

1 INTRODUCTION

It is estimated that infertility affects 48.5 million couples globally. Males are found to be solely responsible for 20-30% of infertility cases and contribute to 50% of cases overall (Agarwal et al. (2015)). Visually analyzing the sperm is a very sophisticated evaluation tool of human fertility (Lu et al. (2010)) and is a strong indicator of a man's physical and, thus, reproductive health. It provides valuable insights about sperm quality and is thus a prognostic tool for predicting male fertility potential concerning in-vivo pregnancies (Eggert-Kruse et al. (1996)). Although the evaluation and clinical significance of sperm morphology have always been regarded as a subjective aspect of semen analysis for the determination of a male's fertility potential since it has to be done by the human eye as proposed in Menkveld et al. (2011), the method's significance in predicting fertilization and pregnancy rates can not be undermined and left unnoticed as it can help in clinical decision making, for instance, to take couples directly to in-vitro fertilization/intra-cytoplasmic sperm injection (IVF/ICSI) (Donnelly et al. (1998)). Even today, most of the recent computer-assisted morphology analysis (CAMA) systems of the sperm still largely depend on human operator skills, which have a severe influence on sperm morphology evaluation results (Menkveld et al. (2011)).

MacLeod's work in 1951 (MacLeod & Gold (1951)) helped the role of sperm morphology gain more recognition as he distinguished several classes of abnormal sperm head formations, and those spermatozoa not classified into any of these classes were regarded as standard. To obtain better classification, Eliasson (1971) stated that for the complete morphological evaluation of human spermatozoa, the entire spermatozoon must be considered (including the mid-piece and the tail). The results of a comparative study between 47 laboratories conducted by Freund (1966) as based on the contemporary head type-classification system for human sperm morphology assessment showed that the method was 'personality orientated' as well as 'subjective, qualitative, non-repeatable and difficult to teach to other persons' according to KATZ et al. (1986). Furthermore, Eliasson stated that the

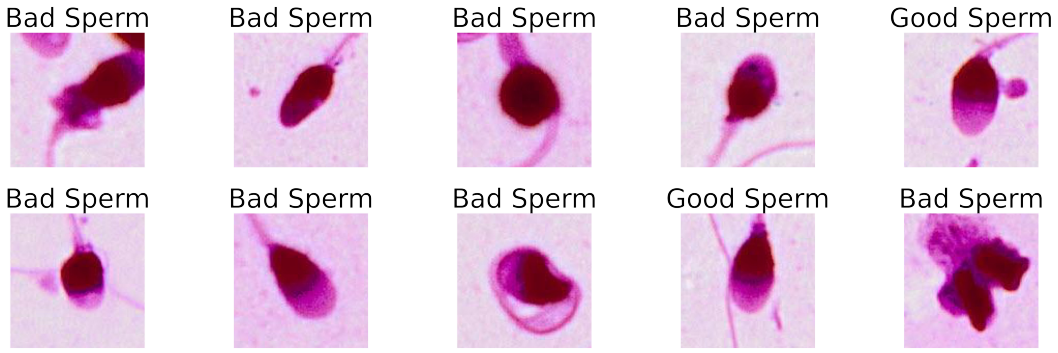

Figure 1: Distributed representation of the Morphological qualities of Sperm Images present in the dataset.

relation between 'morphological patterns' of spermatozoa in the ejaculate and the probability of fertility of the corresponding man relied on probability analyses - which implies strict and reproducible evaluation methods of sperm morphology and a realistic definition of fertility. Spermatozoa occur in infinite forms, and although metric standards have been cited for what a 'normal' human sperm head should be like Eliasson (1971), Eliasson (1980), there is not a biological or clinical basis for these 'normal values' and as per the definition for a morphologically normal spermatozoon proposed by Menkveld et al. (1990) - Borderline standard head forms with no evident anomaly may be regarded as abnormal. A healthy sperm morphology should be free of neck, mid-piece, or tail abnormalities as well, e.g. the tail must be uniform, straight, and thinner than the mid-piece. Over the years, the criteria for sperm morphology evaluation have changed substantially in search of standardization. In the World Health Organisation (1999) manual, strict criteria became the recommended method and were confirmed as the standard method of sperm morphology evaluation as reported in the new World Health Organization et al. (2010) manual. Even though there has been criticism on the concept and use of strict criteria by Comhaire et al. (1994), Morgentaler et al. (1995), Nieschlag et al. (1997), Eliasson (2010) as it lacks evidence and thus deemed unsuitable for use in the clinical laboratory, various literature such as Ombelet et al. (1995) have proven the importance of sperm morphology as a single and independent predictor of in-vivo and in-vitro fertilization. Studies such as Li et al. (2018), and Pasupa et al. (2020) shows cell dynamic morphology classification using Deep Convolutional Neural Networks and canine Red Blood Cell morphology classification using the Generative Adversarial Networks (GANs) which aims at solving the problem of insufficient labelled data, while training a deep learning classifier, respectively, inspired the authors to undertake this study and demonstrate a novel pedagogy for the classification of the sperm images distributed across two classes based on its morphology: 'Good Sperm' and 'Bad Sperm' while solving the Minority Class Imbalance problem using cDCGAN. The data is open-sourced and publicly available at https://www.kaggle.com/c/sperm-morphological-quality/overview. The authors report that the architecture was trained on high-quality annotated slide images and was tested on low-quality annotated slide images, which differed in colour distribution, magnification, and dimension. The testing of the architecture was done on ten different instances of low-quality Sperm Cell images that were provided before the ablation studies were conducted. Additionally, the authors affirm that incorporating more diversified, abnormal sperm morphology categories, and sperm morphology patterns makes this research more comprehensive by making the current model learn, perform, and generalize better. The distribution of both the classes of Sperm Cell images in the provided dataset is depicted in Fig. 1.

2 Proposed Method

In this section, we dive deep into the methods employed, and the pedagogy followed to make our architecture suitable for our classification task. We use various computationally efficient and reproducible methods while also not compromising the reliability of the obtained results. The proposed workflow of our approach is outlined in Fig. 2.

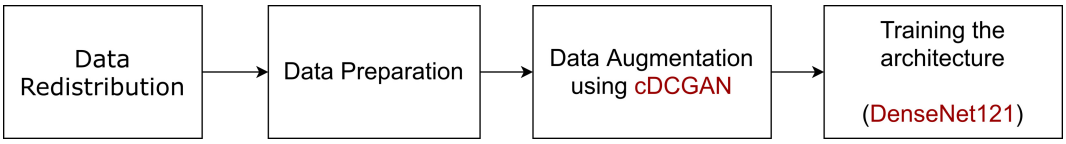

Figure 2: Proposed Workflow.

2.1 Data Re-distribution

In the beginning, we re-distributed the labelled data, which was initially divided into two folders following their respective classes into a single merged Train directory. We organized the labelled data of 'Good' and 'Bad' Sperm into one directory and created a CSV file consisting of the Train labels. The authors chose to begin with creating a data frame with the assigned labels as this would help us ease out the process of Data Replication and further split the Train data into Train and Validation sets, which are further discussed under the purview of the following sections. It is essential to note that the Train directory, after merging images from both the classes, finally consisted of 4142 images.

2.1.1 Data Preparation

We report having used various image processing techniques such as denoising, intensity normalization along with flipping techniques (*fliplr, flipud*), blurring techniques (*Gaussian Blur, Average Blur, and Median Blur*) and filtering techniques (*Sharpen, Emboss, Brightness of the whole image or subareas, Contrast Normalization, Hue, Color-Depth, and Saturation*) to enhance the training data with regards to quality and model translation (interpretability) as we proceeded to build the Train and Validation data generator in the subsequent step. We feed the enhanced images with more distinctive features into the respective data generators, which we ultimately train our DenseNet architecture with, as explained in Section 2.3. The pixels were moved around locally with random strengths assigned to them. The intricacies of how the generated samples were produced and their properties are elaborated in the following section.

2.2 Data Augmentation using cDCGAN

In the initial training set among the 'Good' and 'Bad' sperm images, we observed a significant class imbalance problem and a reasonable decrease in the number of 'Fertile' sperm, which is said to be the result of the strict criterion for the determination of normal sperm morphology (Menkveld et al. (2011)). In the given data, we calculated that approximately for every eight *'Bad Sperm' images* there was only one *'Good Sperm' image*. The authors urge to note that the motivation behind choosing this cDCGAN augmentation approach was based on practical reasons, as when we tried to run our experiments by omitting this step, we obtained inconsistent and unreliable results, which is further explored in Section 3.

In this step, before augmentation, the data was first split into a Stratified 5-Fold approach and then into a train and validation set, the distribution of which is exhibited in Table 1. Next, we execute the cDCGAN algorithm by adding the suggested DCGAN stability techniques (Radford et al. (2015)) with conditional GAN (Mirza & Osindero (2014)) architecture to boost the minority class samples for 'Good' or Fertile Sperm. The architecture of cDCGAN was trained for 100 epochs and is shown in Fig. A.2. The Generator Network is constituted of four transposed CNN layers that consist of 1024, 512, 256, 128, and 1 channel, respectively; the Discriminator is a CNN with four hidden layers. The Generator takes as inputs both the 100×1 stochastic size noise(z), which produces suitable images, and the input spectrum steers the model to generate a design that satisfies the condition. The Generator produces a 64×64 pixels probability distribution function (PDF) in a 500 nm $\times$ 500 nm physical domain. The generated design is again fed into Discriminator to be discriminated from ground-truth examples. The Generator is trained to generate a real but superficial design to deceive the Discriminator while the Discriminator is trained to discern ground-truth examples from those examples generated by Generator; i.e., Generator and Discriminator Network are concurrently trained in the course to minimize or maximize,

$$\min_{G_n} \max_{D_n} V(D_n, G_n) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log D_n(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} \log[(\mathbf{1} - \mathbf{D_n}(\mathbf{G_n}(\mathbf{z})))] \quad (1)$$

where, $D_n(x)$ represents the probability that came from real design, and $D_n(G_n(z))$ represents the probability that generated design of $G_n(z)$ came from generated design. We modify the loss function of the Generator in the cDCGAN to fit our problem to

$$l_G = (1 - \rho) \times l_{G,\text{ dsg}} + \rho \times l_{G,adv}, \quad (2)$$

where, $l_{G,dsg}$ is design loss, $l_{G,adv}$ is an adversarial loss, and ρ is the degree of adversarial loss.

$$l_{G,\text{ dsg}} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2, \quad (3)$$

is binary cross-entropy (BCE) loss between the generated samples Y_i and ground-truth samples $\hat{Y}_i$.

$$l_{G,adv} = \sum_{n=1}^{N} -\log D(G(\hat{X})), \quad (4)$$

This outlines how well the Generator deceives the Discriminator and is also quantitatively estimated by Binary Cross-Entropy (BCE) Loss. For Discriminator, as adversarial loss ID, adv, we utilize the loss function of traditional cGAN (Mirza & Osindero (2014)), which uses the BCE criterion as

$$l_D = l_{D,adv} = \sum_{n=1}^{N} (\log D(G(\hat{X})) + \log(1 - D(\hat{Z})) \quad (5)$$

We optimized ρ to make Generator Network generate high-quality realistic samples. For an adequately low ρ, a competitive outcome cannot be expected, whereas sufficiently high ρ can confuse the learning process. Consequently, an appropriate value of ρ = 0.56 was selected to maximize the Generator's efficacy to create cogent generated examples. Following training, cDCGAN suggests images on a 64 × 64 pixels PDF p(i, j), representing the probability that a 'Good' sperm image exists at the location (i, j). During every training step, the network is trained to optimize the weights to describe the correlation between the input spectrum and the PDF. A sample batch of images from the dataset that had undergone this above augmentation process is displayed in Fig. A.1. The image distributions were increased considerably and to the right extent, and the new distributions for the minority class are also represented in Table 1. We then generate a total of 1,462 'Good' sperm images that were then combined for the Training set and Validation set.

Table 1: Class-wise distribution of images before and after augmentation.

Dataset	Training (# of images)	'Good' sperm images	'Bad' sperm images	Validation (# of images)	'Good' sperm images	'Bad' sperm images
Original dataset	3,727	414	3,313	415	48	367
Dataset after Augmentation with cDCGAN	4,623	1,310	3,313	519	152	367

2.3 Fitting and Training the Model

After preparing the data, the obtained final distribution is denoted in Table 1; we use the same to train our model using a DenseNet121 architecture in 5 Cross-Validation (CV) steps, where each CV is trained for 110 epochs. This model was mainly chosen solely due to its dexterous ability and is also known to perform quite reliably in classification tasks (Huang et al. (2017)). The yardstick chosen to compare our two models with was train and test accuracy, roc_auc score, and F1-Score; the

results are further highlighted and discussed in detail in the following sections. The training method in cross-validation steps is chosen as it is known to have lesser chances of over-fitting or under-fitting and produces good results when tested on unseen examples that are shed light upon under the purview of Section 3. DenseNet121, as we know, follows a pedagogy of concatenation rather than a summation of the layers involved as represented by a sample architecture in Fig. A.3. Finally, the average training accuracy, val_accuracy, and the val_loss scores of the five Cross-Validation models are calculated and are highlighted in Table 2 for both cases of with and without application of Generative Networks in our pipeline.

Table 2: Average Cross-Validation scores obtained for Training DenseNet121 without cDCGAN (CV_{RAW}) and with cDCGAN (CV_{GAN}).

CV#	Train Accuracy	Validation Accuracy	Validation Loss
CV_{GAN}	0.948 ± 0.04	0.874 ± 0.06	0.456 ± 0.05
CV_{RAW}	0.452 ± 0.05	0.342 ± 0.04	0.185 ± 0.03

3 RESULTS

In this section, we will delve into different hand-picked statistical tests performed on the model after training the same based on the aforementioned methodology. We will also discuss how we arrived to obtain the Quality Scores of the corresponding test samples. The workflow we use to validate and draw inferences from our model is shown in Fig. A.5. The statistical tests were meticulously selected based on their suitability with the kind of data we dealt with and were in coherence with the methods that preceded it.

Figure 3: Workflow: Model Validation.

3.1 EXPERIMENTS

We report running our experiments using all the five cross-validation models on both the raw data and the processed data. The inference gave us decent results (Refer Table 2) with an adequate number of 1's against 0's. CV_{RAW} denotes the cross-validation scores on the original unbalanced dataset, while CV_{GAN} represents the scores obtained when the minority class is balanced with cDCGAN. The labels were compared and mapped with their corresponding ground truth labels. Each of the cross-validation inference results consisted of 415 and 519 images from our respective validation sets (Refer Table 1).

Next, we use the obtained CSV files of the validation set to determine their respective performances on our chosen standard metrics. We have mainly chosen F1-score to be a valid and suitable metric because of the original uneven class-distribution and mostly because it is a weighted average of precision and recall. Hence, it accurately clarifies how intuitively the model performed based on the inference drawn from the validation set. The corresponding thresholds were also obtained alongside the F1-scores to ensure that the obtained decimal labels are indexed into either 1 or 0. Table 3 outlines the F1-Scores, roc_auc score, best_threshold (best_thr) values, and the test accuracy scores obtained for the respective validation inferences manifested across each cross-validation step.

3.2 DISCUSSION OF RESULTS

From the nature of these curves and the obtained results on other metrics, the authors ascertain that the results, although not state-of-the-art but are satisfactory, indicate reliable enough to work with

Table 3: Obtained results from inferences distributed across the most optimum Cross-Validation step.

CV#	F1-Score	roc_auc score	best_thr	Test Accuracy
CV_{GAN}	0.602 ± 0.02	0.922 ± 0.03	0.715 ± 0.02	0.862 ± 0.02
CV_{RAW}	0.452 ± 0.05	0.512 ± 0.03	0.185 ± 0.04	0.243 ± 0.06

on the chosen data. Finally, we calculate the Quality Score of each of the ten test samples consisting of the low-quality annotated sperm cell images is defined by the formula outlined in Eqn. 6.

$$Quality\_Score = (\frac{n(G)}{n(G) + n(B)}) * 100 \tag{6}$$

(Where *n(G)* is the number of Good Sperm Cell images and *n(B)* is the number of Bad Sperm Cell images in a given test slide)

The authors find it very instrumental in mentioning that the test slides had a different input_shape (dimensions) than the training data (high quality) images. Therefore, we also had to resize all the individual test slides to the dimension of *130x130* before we could run our inference on the same and hence calculate their corresponding quality scores. All the ten test slides' quality scores are obtained as delineated in Table 4. These inferences were performed on low-quality annotated slide images of unlabelled sperm cells. These scores are a direct representation of the presence of Good Sperm Cell images in a given test slide and can be used to draw further insights such as the diversity of a particular model in question when tested on external examples. The authors make a note that it is evident from the results that there were multiple classes where the quality score was calculated as 0 (D926 and E902) by our proposed framework; therefore, this corresponds to the presence of high variance in the type of distribution of images with which the model was trained and the type it is being used to draw inference upon. Additionally, the scores also indicate room for the furtherance of the present study regarding refinement and enhancement of results. Lastly, comparing the obtained Quality Scores of the Test samples to their respective Gold Standard values with the same label fetched us an average root mean squared error rate of 8.13% with a standard deviation of $\pm$ 0.07 for the detected and undetected samples. Lastly, the obtained ROC and Precision-Recall curves after training the DenseNet121 with CV_{GAN} are highlighted in Fig. A.4 and A.5 respectively with their corresponding areas of overlap.

4 Conclusion

In this paper, we present a method by virtue of which we could exhibit the task of classifying Sperm Cell images based on morphology. However, in the grander scheme of things, this method also encompasses a real-world approach to train an architecture on high-quality annotated images and using the same architecture to draw inferences on low-quality annotated unlabelled images. We also explored the efficacy of our method through the means of various statistical tests. In light of this proposition, we have also made meticulous efforts to make our model learn to classify and validate the results using a different distribution (with greater variance) than the former. As a scope of improvement, the authors report that the results displayed in Table 4 (see appendix) considering 'Quality Score' as the only metric that can be extrapolated further, and other metrics can be incorporated facilitate and derive auxiliary and additional insights from the data. The extensions to validate the hypothesis, as mentioned earlier, are subject to future research.

References

Ashok Agarwal, Aditi Mulgund, Alaa Hamada, and Michelle Renee Chyatte. A unique view on male infertility around the globe. *Reproductive biology and endocrinology*, 13(1):37, 2015.

Frank Comhaire, Frank Schoonjans, Lutgart Vermeulen, and Nicole De Clercq. Methodological aspects of sperm morphology evaluation: comparison between strict and liberal criteria. *Fertility and sterility*, 62(4):857–861, 1994.

Eilish T Donnelly, Sheena EM Lewis, James A McNally, and William Thompson. In vitro fertilization and pregnancy rates: the influence of sperm motility and morphology on ivf outcome. *Fertility and sterility*, 70(2):305–314, 1998.

Waltraud Eggert-Kruse, Heike Schwarz, Gerhard Rohr, Traute Demirakca, Wolfgang Tilgen, and Benno Runnebaum. Sperm morphology assessment using strict criteria and male fertility under in-vivo conditions of conception. *Human reproduction*, 11(1):139–146, 1996.

R Eliasson. Standards for investigation of human semen untersuchungsstandards für das menschliche sperma la standardisation de l'analyse du sperme humain. *Andrologia*, 3(2):49–64, 1971.

R Eliasson. Analysis of semen. in the testis. 1980.

Rune Eliasson. Semen analysis with regard to sperm number, sperm morphology and functional aspects. *Asian journal of andrology*, 12(1):26, 2010.

Matthew Freund. Standards for the rating of human sperm morphology. a cooperative study. *International journal of fertility*, 11(1):97, 1966.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

DAVID F KATZ, JAMES W OVERSTREET, STEVEN J SAMUELS, PAUL W NISWANDER, TODD D BLOOM, and ERNEST L LEWIS. Morphometric analysis of spermatozoa in the assessment of human male fertility. *Journal of Andrology*, 7(4):203–210, 1986.

Heng Li, Fengqian Pang, Yonggang Shi, and Zhiwen Liu. Cell dynamic morphology classification using deep convolutional neural networks. *Cytometry Part A*, 93(6):628–638, 2018.

Jin-Chun Lu, Yu-Feng Huang, and Nian-Qing Lü. Who laboratory manual for the examination and processing of human semen: its applicability to andrology laboratories in china. *Zhonghua nan ke xue= National journal of andrology*, 16(10):867–871, 2010.

John MacLeod and Ruth Z Gold. The male factor in fertility and infertility: Iv. sperm morphology in fertile and infertile marriage. *Fertility and sterility*, 2(5):394–414, 1951.

Roelof Menkveld, Frik SH Stander, Theunis J vW Kotze, Thinus F Kruger, and Johannes A van Zyl. The evaluation of morphological characteristics of human spermatozoa according to stricter criteria. *Human Reproduction*, 5(5):586–592, 1990.

Roelof Menkveld, Cas AG Holleboom, and Johann PT Rhemrev. Measurement and significance of sperm morphology. *Asian journal of andrology*, 13(1):59, 2011.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Abraham Morgentaler, May Y Fung, Doria H Harris, R Douglas Powers, and Michael M Alper. Sperm morphology and in vitro fertilization outcome: a direct comparison of world health organization and strict criteria methodologies. *Fertility and sterility*, 64(6):1177–1182, 1995.

Eberhard Nieschlag, Hermann M Behre, and Susan Nieschlag. *Andrology*. Springer, 1997.

W Ombelet, R Menkveld, TF Kruger, and Omer Steeno. Sperm morphology assessment: historical review in relation to fertility. *Human Reproduction Update*, 1(6):543–557, 1995.

World Health Organisation. *WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction*. Cambridge university press, 1999.

World Health Organization, Trevor G Cooper, et al. Who laboratory manual for the examination and processing of human semen, 2010.

Kitsuchart Pasupa, Suchat Tungjitnob, and Supawit Vatathanavaro. Semi-supervised learning with deep convolutional generative adversarial networks for canine red blood cells morphology classification. *Multimedia Tools and Applications*, 79(45):34209–34226, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

A APPENDIX

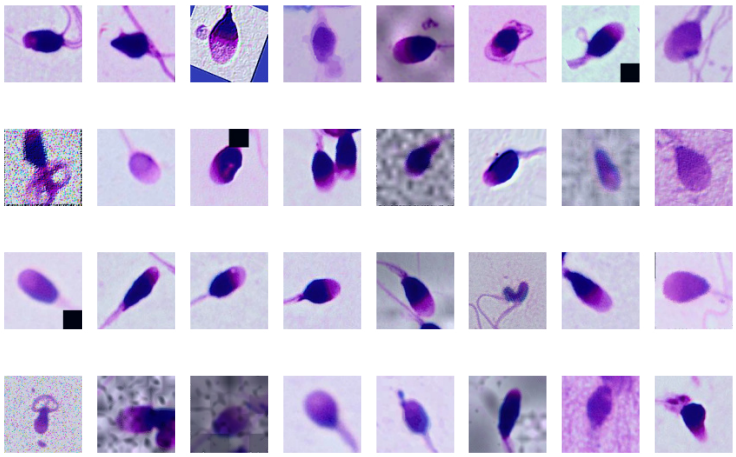

Figure A.1: Sample batch of generated augmented sperm images using cDCGAN.

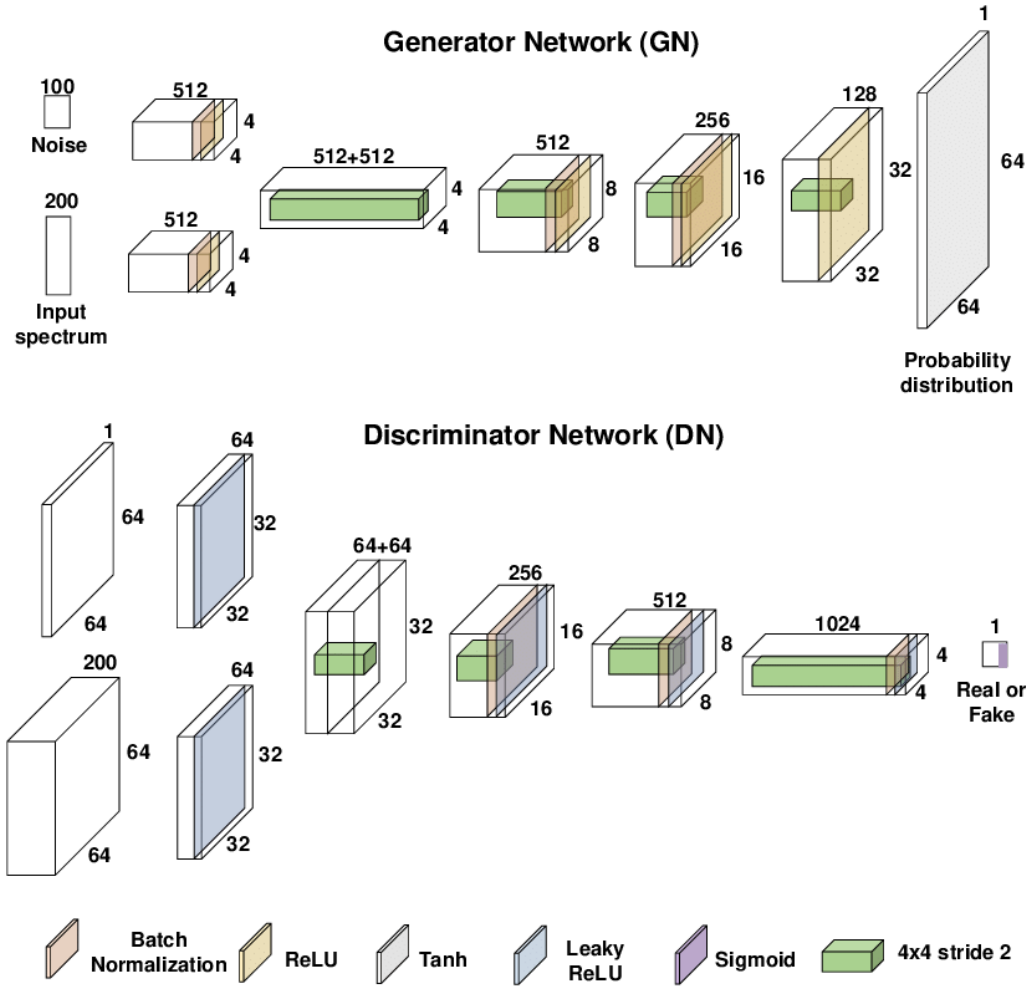

Figure A.2: Architecture of Conditional Deep Convolutional Generative Adversarial Networks.

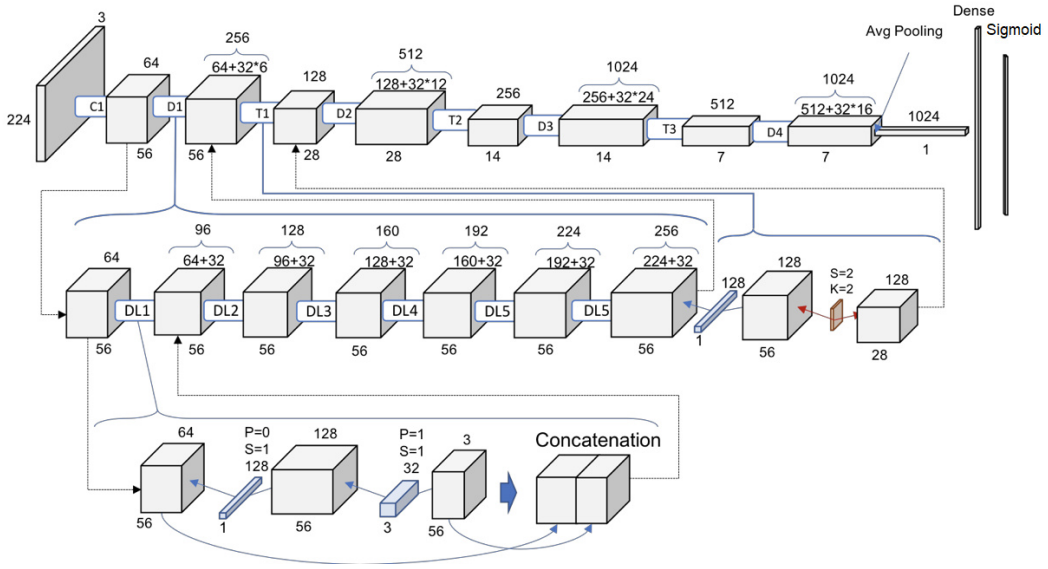

Figure A.3: Concatenation based architecture of a DenseNet121.

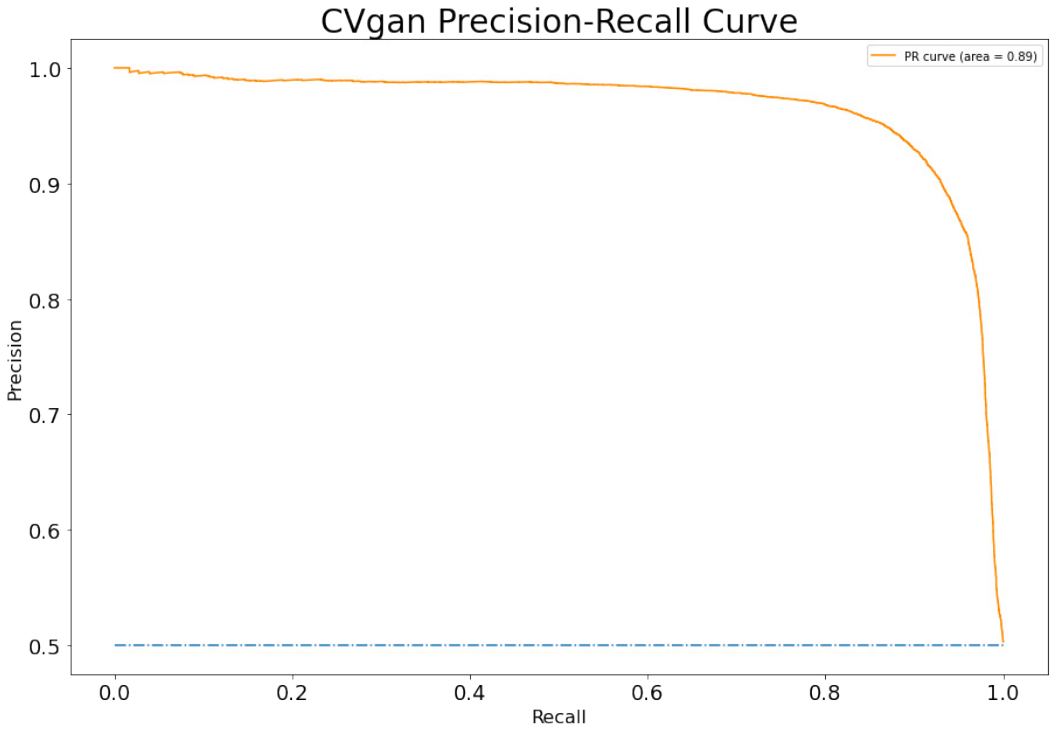

Figure A.4: Obtained Precision-Recall Curve for DenseNet121 trained with CV_{GAN}

.

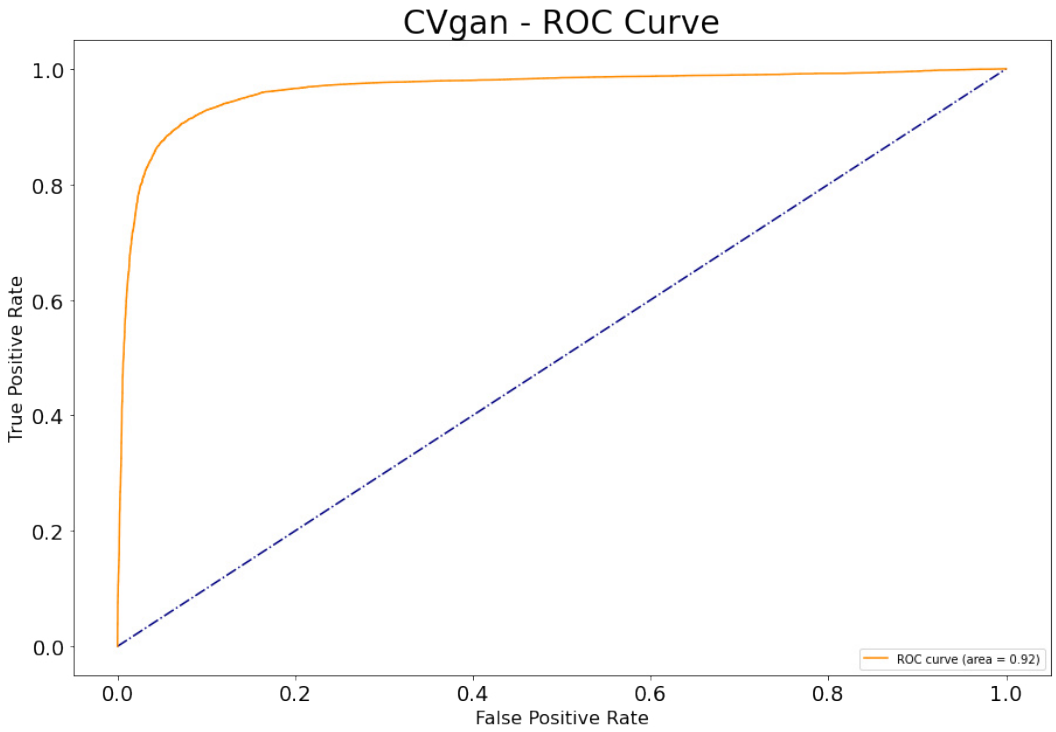

Figure A.5: Obtained Receiver Operator Characteristics curve for DenseNet121 trained with CV_{GAN}
.

Table 4: Obtained Quality Scores for each of the test samples.

Test Samples	Quality Scores
D926	0
E883	35
E902	0
E914	27
E919	17
E950	34
E952	10
E979	46
E986	38
E998	60