

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Cervical Cytology Imaging Generation with Latent Diffusion Models

Tiago Alves



Master in Informatics and Computing Engineering

Supervisor at FEUP: Luís Filipe Teixeira

Supervisor at Fraunhofer: Luís Rosado

February 21, 2023

Cervical Cytology Imaging Generation with Latent Diffusion Models

Tiago Alves

Master in Informatics and Computing Engineering

February 21, 2023

Abstract

The high incidence of cervical cancer in women has considerably accelerated research toward efficient screening techniques over the past decade. If detected early and managed effectively, it is one of the most successfully treatable forms of cancer, making screening tests one of the main reasons why its mortality rate has decreased [21]. However, manual screening is subjective with poorly reproducible criteria, requiring highly specialized technicians, and therefore motivating the research of computer-aided diagnosis [18]. These algorithms require a large amount of cytological images, which are often associated with fewer source patients and a significant imbalance between classes [89]. To address this issue, this research proposes the use of Latent Diffusion models for the synthetic generation of cytological images. Using Dreambooth [72], a Stable-Diffusion model [69] was fine-tuned to generate single cervical cancer cells with different types of neoplastic changes from an input textual prompt. The model was then used to inpaint multi-cellular images with new abnormality types. The generated images were evaluated by two cytopathologists for realism and neoplastic changes, and were also used to retrain two distinct AI models for cervical lesions detection, with the objective of expanding the overall volume of training data and balance the data volume between the different classes. The generated images were found to be highly realistic by specialists, who were unable to distinguish them from real cytological images. Furthermore, the combination of real and synthetic images improved the performances of one of the detection models, achieving state-of-art results for this dataset and showcasing the potential of current generative AI approaches to enhance deep learning object detection models.

Keywords: *Cervical Cancer Screening, Artificial Intelligence, Latent Diffusion Models, Image Inpainting, Text-To-Image Generation, Stable Diffusion*

Resumo

A alta incidência de cancro cervical nas mulheres acelerou consideravelmente a pesquisa de técnicas de deteção deste na última década. Caso seja detetado numa fase inicial e tratado corretamente, é uma das formas de cancro mais eficazmente tratáveis, tornando os algoritmos de deteção uma das principais razões pelas quais a sua taxa de mortalidade tem vindo a diminuir [21]. No entanto, a deteção manual é subjetiva, com critérios pouco reprodutíveis, exigindo técnicos altamente especializados e motivando consequentemente o diagnóstico auxiliado por computador [18]. Estes algoritmos requerem uma grande quantidade de imagens citológicas, que geralmente estão associadas a poucos pacientes-fonte e a um desequilíbrio significativo entre as classes [89]. Este trabalho propõe o uso de modelos de difusão latente para a geração sintética de imagens citológicas. Usando Dreambooth [72], um modelo de Stable-Diffusion [69] foi ajustado para gerar células singulares de cancro cervical com diferentes tipos de alterações neoplásicas a partir de texto. O modelo foi de seguida usado para alterar partes de imagens multicelulares com novos tipos de anormalidades. As imagens geradas foram avaliadas por dois citopatologistas quanto ao seu realismo e alterações neoplásicas. Foram ainda usadas para retrainar dois modelos de Inteligência Artificial distintos de deteção de lesões cervicais, com o objetivo de expandir o volume dos dados de treino e balancear o volume entre as várias classes. As imagens geradas foram consideradas altamente realistas pelos especialistas, não conseguindo distingui-las das imagens citológicas originais. Além disso, a combinação entre imagens reais e sintéticas melhorou o desempenho de um dos modelos, atingindo resultados estado-da-arte para esta base de dados e mostrando assim o potencial das atuais abordagens generativas para melhorar modelos de deteção baseados em *deep learning*.

Palavras-Chave: *Deteção de Cancro Cervical, Inteligência Artificial, Modelos de Difusão Latente, Geração Texto-a-Imagem, Stable Diffusion*

Acknowledgements

I would like to take this opportunity to express my profound gratitude to my supervisors, Doctor Luís Rosado and Professor Luís Teixeira, for their invaluable support and guidance throughout the research process. Their expertise and knowledge in the field were invaluable in shaping the direction of this research, and their feedback and suggestions were instrumental in the development and refinement of this work. I am deeply appreciative of the opportunity to have worked under their mentorship and for their unwavering support and encouragement. I would also like to extend my sincere appreciation to the staff at Fraunhofer AICOS for providing a welcoming and supportive environment. Special thanks are due to Ana Sampaio, Vladyslav Mosiichuk, and Tomás Noronha for their assistance and expertise in solving the tasks required for this dissertation.

I would like to extend my sincerest gratitude to the doctors from IPO for their invaluable assistance in evaluating the results of this dissertation. Their willingness to generously donate their time and expertise was greatly appreciated and instrumental in the success of this research.

Lastly, I am deeply grateful to my family for their unwavering support and encouragement. To my mother, for her emotional support and technical help in the analysis of this work. To my father, for serving as a role model and consistently providing an example on how to live a good life. And to my sister, for always bringing a good mood wherever she goes.

This work was done under the scope of the project Transparent Artificial Medical Intelligence (TAMI), co-funded by Portugal 2020 framed under the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), Fundação para a Ciência and Technology (FCT), Carnegie Mellon University, and European Regional Development Fund under Grant 45905. The authors would like to thank the Anatomical Pathology Service of the Portuguese Oncology Institute – Porto (IPO-Porto).

Tiago Alves

*“Science is not only a disciple of reason but, also,
one of romance and passion.”*

Stephen Hawking

Contents

1	Introduction	1
1.1	Objectives	2
1.2	Document Structure	2
2	Cervical Cancer	4
2.1	Overview	4
2.2	Cervical Cancer screening	4
2.2.1	Identification of cervical lesions	5
2.3	Bethesda System	6
2.3.1	The Atypical squamous cells	7
2.3.2	Cytomorphological features	8
3	Datasets and Previous work	12
3.1	Overview	12
3.2	μ SmartScope	12
3.3	AI-based Cervical Lesions Detection	13
3.4	Region-Based Approach	14
3.4.1	Dataset	14
3.4.2	Detection model	16
3.5	Nuclei-Based Approach	17
3.5.1	Dataset	17
3.5.2	Detection Model	18
3.6	Deep Feature Consistent Variational Autoencoder	20
3.6.1	Single Cell Dataset	21
4	Generative AI	24
4.1	Overview	24
4.2	Variational Auto Encoders	24
4.3	Generative Adversarial Networks	25
4.4	Diffusion models	26
4.4.1	Latent Diffusion Models	28
4.5	Stable Diffusion	30
4.5.1	Stable Diffusion Versions	30
4.5.2	Textual Inversion	31
4.5.3	Dreambooth	32
4.5.4	Inpainting	34

5	Methodology	36
5.1	Hardware	36
5.2	Fine tuning stable diffusion model	37
5.2.1	Model selection	38
5.2.2	Dataset preparation	39
5.2.3	Model training	39
5.2.4	Validation	42
5.3	Multiple Cell inpainting	43
5.3.1	Dataset preparation	43
5.3.2	Hyperparameter tuning	47
5.3.3	Validation	48
6	Results and Discussion	52
6.1	Fine tuning stable diffusion model	52
6.1.1	Model selection	52
6.1.2	Hyperparameter tuning	56
6.1.3	Cythopathologist validation	62
6.2	Multiple Cell Inpainting	64
6.2.1	Cythopathologist validation	67
6.3	AI-based Cervical Lesions Detection Algorithms	69
6.3.1	Region-Based Detection Model	69
6.3.2	Nuclei-Based Detection model	73
6.3.3	AI-based Cervical Lesions Detection Algorithms Comparison	76
7	Conclusion and Future work	78
	References	80
A	AI-based Cervical Lesions Detection Algorithms Detailed Results	87
A.1	Region-based Detection Model	87
A.2	Nuclei-based Detection Model	88
B	Cythopathologists Questionnaire	90
B.1	Single Cell Questionnaire	90
B.1.1	Screenshots	90
B.1.2	Results by Specialist	91
B.2	Multiple Cell Questionnaire	92
B.2.1	Screenshots	92
B.2.2	Results by Specialist	94

List of Figures

2.1	Comparison between normal cellular components and squamous intraepithelial lesions [5].	8
2.2	Key characteristics of LSIL [5].	9
2.3	HSIL has a checkerboard pattern with black nuclei, smudged chromatin, nucleoli, and unevenly spaced apoptotic bodies [5].	10
3.1	(A) μ Smartscope with phone and LBC sample (B) Smartphone application screenshots [73].	13
3.2	Overall Scheme comparing the Region-based approach to the Nuclei-based approach.	14
3.3	Examples of the five lesion classes considered in the mobile Region-based dataset [12].	15
3.4	Patch extraction process [73].	16
3.5	Example of the Nuclei-based dataset annotations [53].	18
3.6	Model overview. The left is a deep CNN-based Variational Autoencoder and the right is a pre-trained deep CNN used to compute feature perceptual loss [36].	20
3.7	Results of the reconstructed images with 256x256 for the different numbers of epochs (images on top are the reconstruction).	21
3.8	Transformation of ASC-H in ASC-US via interpolation.	21
3.9	Illustrative examples of single cell dataset regions instances for each class.	22
3.10	Illustrative examples of single cell dataset nuclei instances for each class.	23
4.1	Variational Auto Encoder global architecture [76].	25
4.2	Simplified GAN Architecture in skin lesion synthesis [12].	26
4.3	Diffusion Model simplified Architecture [35].	27
4.4	Latent diffusion models [69].	30
4.5	Dreambooth training procedure [72].	33
4.6	Comparison between Textual Inversion and Dreambooth [72].	33
4.7	The scheme of the proposed method for large-mask inpainting (LaMa) [80].	35
5.1	Overall pipeline of the proposed methodology.	37
5.2	Comparison between images which were (a) included and (b) discarded from the training set on the experiments with 30 hand-selected images.	41
5.3	Example of a patch (a) and corresponding inpainting mask (b).	44
5.4	Comparison between Nuclei-based dataset average values and the literature.	45
5.5	Comparison between two resized masks of different classes, for the same patch.	47
5.6	Comparison between two generated cells. The ASC-US image (a) is clearly bigger than the HSIL image (b).	49
5.7	Synthetic multi-cellular image presented in the questionnaire to the cytopathologists.	51

6.1	Comparison between (a) XavierXiao Dreambooth model and (b) JoePenna. . . .	53
6.2	JoePenna model results for the Nuclei dataset. Trained with 2000 steps (a) 5000 (b) and 10000 (c).	53
6.3	Comparison between generated images (a) with the model and the original images from the training dataset (b).	54
6.4	Comparison between generated nuclei with the ShivamShrirao model (a) and original images from the nuclei dataset (b).	54
6.5	Comparison between fine-tuning the model over the sd1.4 model (a) or the sd1.5 (b).	55
6.6	Generated images using all the regions instances of the single cell database as regularization images for the Region-based approach.	56
6.7	Generated images using all the nuclei instances of the single cell dataset as regularization images.	57
6.8	Generated images using all the regions instances of the single cell dataset, apart from the class that is being trained as regularization images.	57
6.9	Generated images using the unrelated class as regularization images.	58
6.10	Generated images using the unrelated class as regularization images for the Nuclei dataset.	58
6.11	ASC-US generated images using all the regions instances of the single cell(a) and only 30 (b) for training.	59
6.12	Generated images using all the region instances of the single cell (a), only regions (b), and single cells (c).	60
6.13	Generated ASC-H cells for the Region-based approach using 5000 steps (a) and 10000 steps (b).	61
6.14	Generated nuclei using different step counts.	61
6.15	Illustrative synthetically generated images obtained after parameters tuning for the Region-based Dreambooth model.	62
6.16	Illustrative synthetically generated images obtained after parameters tuning for the Nuclei-based Dreambooth model.	63
6.17	Questionnaire results regarding the realism of single cell images.	63
6.18	Inpainted Region-based patches for each abnormality class with fixed inpainting area size.	65
6.19	Inpainted Nuclei-based Patches for each abnormality class with fixed inpainting area size.	66
6.20	Inpainted Region-based patches for all the abnormality classes with variable inpainting area size and respective masks.	67
6.21	Questionnaire results regarding the realism of multiple cell images.	67
6.22	Agreement between ground truth and cytopathologist in evaluating abnormality class of multiple cells.	68
6.23	Results of the Region-based detection model. Both graphs present the values for each abnormality class and for each test. Graph (a) presents the mAP@.50IOU values and graph (b) presents the AR10.	70
6.24	Results of the Nuclei-based detection model. Both graphs present the values for each abnormality class and for each test. Graph (a) presents the mAP@.50IOU values and graph (b) presents the AR100.	73

6.25	Comparison between data augmentation and synthetic images for the Nuclei-based approach. Both graphs present the values for each abnormality class and for each test. Graph (a) presents the mAP@.50IOU values and graph (b) presents the AR100.	75
6.26	Comparison between Region-based approach and Nuclei-based approach. Both graphs present the values for the performance obtained in the original works [73, 53], and the best performance obtained with the generated images. Graph (a) presents the mAP@.50IOU values and graph (b) presents the AR10.	76
B.1	Introduction of the Single cell questionnaire.	90
B.2	Example of two questions included in the single cell questionnaire. In (a) its represented a real image, while in (b) its represented a synthetic image.	91
B.3	Questionnaire results of specialist 1 regarding the realism of single cell images. (a) Results regarding the synthetic images. (b) Results regarding real images. . .	91
B.4	Questionnaire results of specialist 2 regarding the realism of single cell images. (a) Results regarding the synthetic images. (b) Results regarding real images. . .	92
B.5	Introduction of the Multiple cell questionnaire.	92
B.6	Example of two questions included in the Multiple cell questionnaire. Each image has two corresponding questions, the first which asks regarding the realism of the image, and the second wich asks for the lesion class. In (a) its represented a real image, while in (b) its represented a synthetic image.	93
B.7	Questionnaire results of specialist 1 regarding the realism of multiple cell images. (a) Results regarding the synthetic images. (b) Results regarding real images. . .	94
B.8	Questionnaire results of specialist 2 regarding the realism of multiple cell images. (a) Results regarding the synthetic images. (b) Results regarding real images. . .	94

List of Tables

2.1	Risk estimates supporting the 2019 ASCCP risk-based management consensus guidelines [3].	8
3.1	Region-based dataset sample and annotation distribution (training, test and total) [73].	15
3.2	Class-wise performance of the Faster R-CNN C Resnet50 model [73].	16
3.3	Distribution of the Nuclei-based dataset by class [53].	18
3.4	mAP@0.50 and AR10 of the Nuclei and the Regions detection models [53, 73].	19
3.5	Number of single cell dataset regions instances for each cell class.	22
3.6	Number of single cell dataset nuclei instances for each cell class.	23
5.1	High power computing hardware provided by Fraunhofer.	37
5.2	Region-based Dataset distribution after data preparation procedures.	39
5.3	Cell class and corresponding textual prompt for the Text to Image model.	40
5.4	Relative average size comparison between different cervical lesions classes (in percentage).	46
5.5	Relative average size of cervical cell nucleus lesions compared to an Intermediate Squamous nucleus (in percentage).	46
5.6	Number of patches created for data augmentation for the Region-based detection model in the original paper [73].	46
5.7	Number of annotations created for data augmentation for the Nuclei-based detection model in the original work[53].	46
5.8	Total number of patches for the different Region-based tests.	50
5.9	Total number of training patches for the different tests of the Nuclei-based approach.	50
A.1	Region-based approach detection model mAP@.50IOU for the different classes and for each test.	87
A.2	Region-based approach detection model AR10 for the different classes and for each test.	87
A.3	Nuclei-based approach detection model mAP@.50IOU for the different classes and for each test.	88
A.4	Nuclei-based approach detection model AR100 for the different classes and for each test.	88
A.5	The highest mAP@.50IOU values for each class independently of the number of epochs and the test. Its also presented the corresponding loss and number of epochs	88
A.6	Total number of training annotations for the different tests of the Nuclei-based approach.	89

Abbreviations

AGC	Atypical glandular cells
ASC	Atypical Squamous Cell
ASC-H	Atypical Squamous Cells, HSIL cannot be excluded
ASC-US	Atypical Squamous Cells of Undetermined Significance
CAD	Computer-Aided Diagnosis
CDA	Conventional Data Augmentations
CFG	Classifier-free Guidance
CNN	Convolutional neural networks
CPU	Central Processing Units
FFCs	Fast Fourier convolutions
FID	Fréchet Inception Distance
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HPV	Human papillomavirus
HR-HPV	High risk HPV
LaMa	Large mask inpainting
LBC	Liquid based cytology
LSIL	Low-grade SIL
NILM	Negative for intraepithelial lesion or malignancy
PAP	Papanicolaou
SCC	Squamous cell carcinoma
SIL	Squamous intraepithelial lesion
TAMI	Transparent Artificial Medical Intelligence
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

Cervical cancer has affected more than 600 thousand women in 2020, being the fourth most diagnosed cancer in women [58]. The high incidence of cervical cancer has greatly influenced research efforts in recent years, leading to a significant acceleration in the development of efficient screening techniques for early detection and prevention. However, the manual completion of cervical cancer screening is arduous, requiring highly trained specialists to obtain good results, which resulted in an uprising interest in automated systems [18]. Despite the high levels of performance achieved by current cervical cancer screening techniques, there are still certain limitations that need to be addressed. One of the main challenges is related to the algorithmic methods employed, particularly the reliance on deep learning techniques. The use of these methods requires large amounts of high-quality, well-annotated training data, which can be scarce, imbalanced, and of low quality, particularly in low-resource settings. This poses a significant challenge for the development and deployment of accurate and reliable cervical cancer screening systems.[89].

Generative modeling is a machine learning technique that automatically identifies and learns the regularities and patterns in the input data, such that it can produce new examples that might have been reasonably derived from the original dataset [19]. In particular, Latent Diffusion Models have emerged as a promising approach in the field of image processing over the past several months and have been shown to be particularly effective in the areas of image synthesis and image inpainting, even outperforming more traditional methods such as Generative Adversarial Networks [23].

The purpose of this master's thesis is to explore the potential of utilizing Latent Diffusion Models to enhance cytological imaging collections for cervical cancer screening. The validity of the generated images will be evaluated by two cytopathologists based on realism and accuracy. These images will then be used as training data for two previously developed cervical screening detection models [73, 53], alongside the original images, to assess the potential improvement in performance.

This dissertation is part of the Transparent Artificial Medical Intelligence (TAMI) project, which intends to make AI methods used by clinical practitioners more understandable and interpretable with particular emphasis on decision support systems that use image data [31].

1.1 Objectives

The primary objective of this research is to explore and compare various methods for creating cervical cytological images using Latent Diffusion Models. The study will focus on the use of the open-source latent text-to-image diffusion model Stable-Diffusion and its application in text-to-image generation and image inpainting [69]. Additionally, the investigation will center on evaluating the metrics and techniques that ensure the synthetic images not only look realistic but also retain the reliability and important features of the original images. The ultimate goal is to develop a model that can generate cytological images of different cervical cancer abnormalities, accurately depicting the characteristic neoplastic changes of each class. More specifically, it has the objective of improving the cervical image datasets that are currently available by:

1. Expanding the overall volume of training data.
2. Increasing the number of samples of underrepresented classes.
3. Balance the data volume between the different classes.

In order to evaluate the impact of the generated images, this work will report the improvements made to two of the already developed decision support systems developed in the TAMI [31] project, with the augmented dataset. Furthermore, the AI-generated images will also be examined in terms of realism and relevance by medical experts, for each one of the respective classes. This thesis will make use of private cytological datasets gathered as part of the CLARE [29] and TAMI [31] initiatives.

1.2 Document Structure

The objective of this chapter is to provide an overview of the document structure of the thesis. The structure is designed to clearly and effectively communicate the research conducted, including the background, methods, results, and conclusions. The thesis is divided in five main sections: Introduction, Background, Methodology, Results and Conclusion and Future work.

The **Introduction** section provides an overview of the research topic and the objectives of the thesis. Following next, the Background part provides information regarding previous work, state-of-art approaches to similar problems, and an overview of the datasets which are going to be used in this research. This particular part is divided into various chapters: The **Cervical Cancer** section delves into the specific topic of cervical cancer, including overviews, screening methods, and the Bethesda System. The **Datasets and Previous work** section discusses the various datasets used in the research and previous work in the field of cervical cancer detection, such as the detection models that are going to be used to validate this work. The **Generative AI** section compares different generative AI techniques, such as Diffusion models and Generative Adversarial Networks, and provides an in-depth analysis of the technologies used in this work.

The **Methodology** chapter presents the proposed methodology for the research conducted in this work. The results of the research are presented in the **Results and Discussion**, as well as deeply

analyzed and discussed. The results are then discussed in relation to the research objectives and existing literature in the field. Finally, **Conclusion and Future work** presents the main conclusions drawn from the research, highlighting the significance of the findings, as well as recommendations for future research, which could build on the work presented in this paper.

Chapter 2

Cervical Cancer

2.1 Overview

Cervical cancer is a type of cancer that develops in the lower part of the uterus, in the cells of the cervix. It is usually caused by a sexually transmitted infection called human papillomavirus (HPV), which is responsible for over 99% of the cases [59]. Although this infection resolves spontaneously in 90% of the times, if the infection persists for 10 to 20 years, it can contribute to the process of transforming cervical cells in cancer cells [59]. When diagnosed early, it is one of the most successfully treatable forms of cancer, if handled correctly. Still, if the person receives no treatment, the cancer will most likely result in death, making it essential to have an effective way of detecting this cancer in an early stage [18]. It is the fourth most prevalent female malignancy in the world, posing as a major global health challenge. In 2020 there were 604 000 new cases and 342 000 deaths from cervical cancer in women worldwide, around 90% of which took place in low and middle-income nations [59].

2.2 Cervical Cancer screening

Cervical Cancer screening has an essential role, allowing to find pre-cancer or cancer in an initial phase while it is more curable. In order to reduce cervical cancer, the best option in high income countries is to vaccinate girls against the HPV, and regularly screen for this type of cancer in woman [59]. In low and middle-income countries however, due to fewer tools and access to vaccines, cancer screening emerges as the secondary alternative preventive measure [59]. The three screening tests currently recommended by the World Health Organization are HPV testing, visual inspection with acetic acid (VIA), and cervical cytology (Pap smear or liquid-based cytology) [59].

An HPV test is a laboratory procedure used to examine DNA or RNA for specific HPV infections. To determine whether there is an HPV infection, cells from the cervix are taken and tested using PCR or hybrid capture [38]. This test can be done alone or in combination with a Pap smear test, which is used to detect abnormal cervical cells and may prompt the need for an HPV test. Pap

and HPV co-testing or HPV testing alone are more sensitive than Pap testing alone for women over the age of 30. In particular, HPV testing has shown increased sensitivity for detecting high-grade cervical intraepithelial neoplasia and has been shown to offer 60% to 70% more protection against invasive cervical cancer [38].

With VIA, or visual inspection with acetic acid, doctors may see the abnormal cervical lesions directly. This procedure must be performed by a skilled healthcare professional, and it calls for the use of a speculum and a light source. The procedure consists of administering a liberal amount of a 3%–5% acetic acid solution to the cervix with a sizable cotton swab [27]. Areas that turn somewhat white due to inflammation, such as metaplasia, typically go away within the first minute, whereas damaged tissues stay white and are more likely to be linked to cervical pre-cancer or cancer. Cervical lesions typically develop near the squamocolumnar junction. Once the white patches are found, the doctor can use cryotherapy or other methods to remove damaged tissues [27]. VIA is not recommended for postmenopausal women because the squamocolumnar junction becomes harder to visualize with age, which can negatively impact the accuracy of the test [18]. VIA doesn't rely on specialized laboratory services, in addition to using supplies purchased locally and being fairly affordable. However, quality control and quality assurance for VIA is especially crucial because of the subjective nature of the exam [27].

Cervical cancer screening with cervical cytology has been instrumental in reducing the mortality rate of cervical cancer by over 90% [18]. The two most common cervical cytology tests are the Papanicolaou (PAP) and the Liquid-based cytology (LBC) [34]. The Pap smear test is the most widely used procedure to detect cervical cancer, being simple, non-invasive, cost-effective, and easy to perform. This test detects abnormal changes in cervical cells, which suggest that a cancer is developing [39]. One limitation of the Pap smear test is that some of the cellular material is lost during sample collection, and results can be compromised by the presence of blood or inflammation [34]. The LBC, on the other hand, creates a thin monolayered smear with a clear background, reducing the amount of false results and increasing the screening speed significantly. The Pap smear is still more prevalent in developing countries, since LBC is less affordable [34].

2.2.1 Identification of cervical lesions

Identification of cervical lesions in microscopic fields is one of the key objectives covered by cervical sample analysis. Researchers often make an effort to complete this quest by segmenting the image's cells and further classify them according to their abnormalities. Although the primary goal of this stage is to identify the cells that require detailed analysis, it can also serve as the foundation for computing cell characteristics that have clinical relevance, such as the shape and size of cells and their associated inner structures. Some researchers used conventional image analysis methods for this purpose. In order to identify the outlines of the cell's nuclei, Byju *et al.* [13] uses a customized Laplacian of Gaussian (LoG) filter, and in [51] uses a strategy that aims to segment individual cells with a focus on separating any that may appear aggregated and partially overlapping in clumps. In recent years, machine learning is also used as an alternative cell segmentation technique, to assess whether each image pixel is part of the cell region or the

background. In [86] and [33], it is used a pixel-wise classifiers to distinguish nuclei, cytoplasm, and background pixels with varying degrees of success. It is worth mentioning these techniques were only used to analyze single-cell images and not larger microscopic fields, limiting their standalone applicability.

Another important aspect of cervical cell analysis is the classification of cells based on the presence or severity of abnormalities or lesions. To this end, many algorithms focus on extracting relevant information from the images of the cells. The global significant value (GSV), which is a representation of the overall intensity variation of the cell's picture, is merged with the cell's perimeter and a rotation-invariant feature in [26] to create the feature vector used to input the classifier. Among the various works, k-NN (k-nearest neighbors) produced a superior state-of-the-art performance of 88.45%.

Cell categorization, classification and analysis have been significantly improved by the use of artificial intelligence and deep learning in recent years [2]. These technologies have been shown to be more effective than traditional methods, such as Pap smears and cervicography, in identifying and distinguishing between different types of cells [18]. Convolutional neural networks, in particular, have been used to learn multilevel characteristics of cells and classify them into different categories. In the work of Krizhevsky *et al.* [46], for example, RestNet [85] was used to do a binary separation of normal and cancerous cells. In order to distinguish between the various classes of abnormal cells, Plissiti suggested the annotated cervical cell image collection SIPaKMeD and used once again CNN [64]. In recent years there has also been a growing interest in developing detection-oriented frameworks that aim to improve the overall diagnostic performance of the resulting systems. One such approach is presented in [93], which utilizes a YOLO v3 detection model and incorporates several post-detection classifiers that take into account the characteristics of the surrounding regions and the nuclei of cells. By doing so, this framework is able to reduce the number of false detections and also identify other potential infectious diseases. This method, which was developed using a large multi-center dataset, is considered to be one of the most comprehensive and effective approaches for cervical lesion diagnosis in the literature [93].

2.3 Bethesda System

There are various classification schemes for cancerous and precancerous lesions of the cervix. These can be cytologically or histologically based, and their clinical and reporting objectives may differ. The Bethesda System (TBS) is a widely used classification scheme for reporting the results of cervical cancer screening tests, including liquid-based cytology tests and conventional Pap smears. This technique provides findings for both cytology results and sample adequacy [55]. The evaluation of the sample's quality is a crucial stage in ensuring the validity of the results; hence TBS considers the sample's quality when conducting the screening by setting objective rules and minimal requirements that must be adhered to:

- Cellularity: There should be a minimum number of discernible squamous cells in the produced samples. A liquid-based preparation needs at least 5000 cells that are clearly visible

and in good condition. A normal Pap-Smear preparation should have between 8000 and 12000 cells, at the very least [55].

- **Obscuring Factors:** A specimen's interpretation may be hampered by the presence of too much blood, inflammation, bacteria, or lubricant, which prevents the correct visualization of the relevant cells [55].
- **Evidence of Transformation Zone:** It is recommended that the samples contain 10 well-preserved endocervical cells or 10 squamous metaplastic cells [55].

If the sample meets all requirements, it is considered adequate and can be interpreted. If the cells studied indicate malignant tumors or abnormalities, an abnormal result will be recorded. Specific results, such as benign infections or inflammations, are presented as typical findings and are not thought to be problematic. Abnormal cells can be classified as atypical squamous cells (ASC) or atypical glandular cells (AGC), and a different grading method is used for each type [55]. In this paper, the datasets only deal with the ASC system.

2.3.1 The Atypical squamous cells

The ACM system classifies the cells into the following categories, listed by increasing level of abnormality:

- Atypical squamous cell of undetermined significance (ASC-US)
- Low-grade squamous intraepithelial lesion (LSIL)
- Atypical squamous cell, cannot exclude high-grade lesion (ASC-H)
- High-grade squamous intraepithelial lesion (HSIL)
- Squamous cell carcinoma (SCC)

The TBS category "Epithelial Cell Abnormality: Squamous" contains the squamous intraepithelial lesion (SIL) category, which includes a range of squamous cell lesions from low-grade SIL (LSIL) to high-grade SIL (HSIL) to invasive squamous cell carcinoma. Atypical Squamous Cells (ASCs), which are divided into two categories based on the suspected underlying lesion LSIL versus HSIL, respectively, are subdivided into "Atypical Squamous Cells of Undetermined Significance" (ASC-US) and "Atypical Squamous Cells, HSIL cannot be excluded" (ASC-H), respectively [5]. Depending on the relative risk of developing cervical cancer with high-risk HPV (HR-HPV) status and other clinical information used as ancillary support, the first cytological findings that suggest gray zone interpretations, such as ASC-US and ASC-H, may be downgraded or upgraded to one of the final interpretations. With the availability of HPV test results as part of co-testing, this is now achievable in some circumstances [5].

Citology	Frequency (%)	Freq of HR-HPV+
NILM	94.0	4
ASCUS	3.6	54
LSIL	1.7	87
ASC-H	0.3	82
HSIL	0.3	95

Table 2.1: Risk estimates supporting the 2019 ASCCP risk-based management consensus guidelines [3].

2.3.2 Cytomorphological features

This section outlines the main cytologic characteristics that set Bethesda squamous groups apart from other significant entities and the key mimickers that may be taken into account during the differential interpretation of SIL. One of the most crucial diagnostic markers utilized to interpret epithelial cell abnormalities is nuclear expansion relative to the size of intermediate cell nuclei (ICN), as seen in figure 2.1.

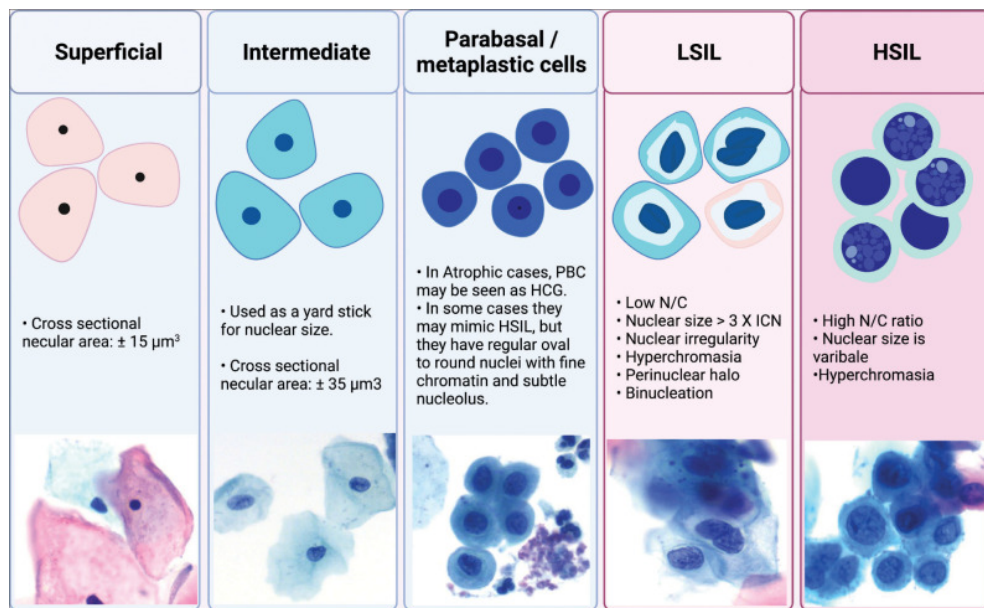


Figure 2.1: Comparison between normal cellular components and squamous intraepithelial lesions [5].

2.3.2.1 LSIL

LSIL describes morphologic alterations at the lower end of the SIL spectrum. Approximately 1.7% of all PAPs are classified as LSIL, and more than 80 % of these are HR-HPV positive [5].

Some key characteristics which define LSIL are [5]:

- Nuclear enlargement (nuclear size greater than three times ICN) with cytoplasm resembling intermediate/superficial cells and a low N/C ratio in comparison to intermediate cells.
- Nuclear hyperchromasia with occasional binucleation.
- Sharp angulations and indentations with varying degrees of nuclear irregularity.
- Coarse chromatin.
- May have extensive orangeophilia (atypical parakeratosis), a sign of enhanced keratinization.
- The HPV cytopathic effect is characterized by sharply delineated perinuclear cytoplasmic clearance (koilocytosis) with uneven outline and localized angulation.
- It may be challenging to identify ASC-H or HSIL from LSIL cells with immature metaplastic cytoplasm (small atypical parakeratotic [SAPK] cells). This could be caused by an acanthotic condylomatous lesion or eosinophilic dysplasia.

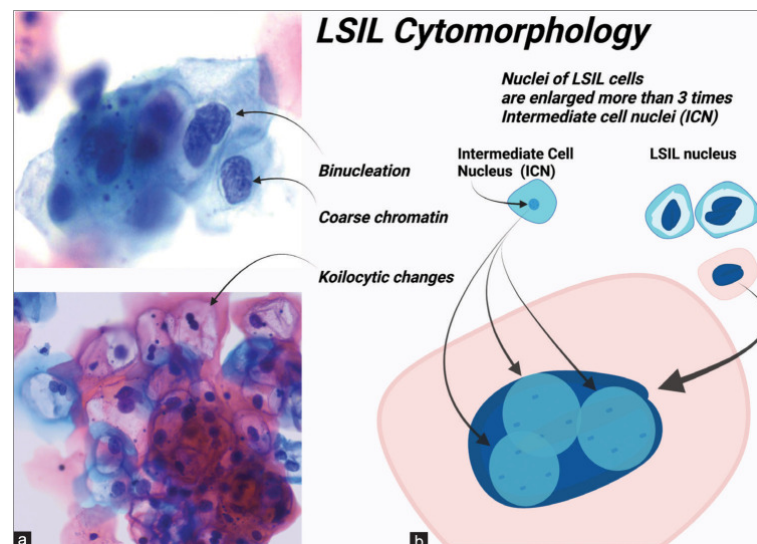


Figure 2.2: Key characteristics of LSIL [5].

2.3.2.2 HSIL

The term HSIL describes morphological alterations connected to the higher end of the SIL spectrum. An estimated 0.3% of all PAPs are classified as HSIL, and 95% of them are HR-HPV positive. HSIL has a lower rate of regression and a higher rate of cancer advancement. For 30 years, the anticipated long-term progression to invasive cancer is 30% [5]. The most important characteristics are [5]:

- HSIL cells are dispersed randomly, in sheets with a checkerboard pattern, or in hyperchromatic crowded groups.

- Compared to LSIL cells, HSIL cells are smaller and have less cytoplasm.
- Higher nuclear cytoplasmic ratio than LSIL.
- Nuclear outlines that are irregularly shaped with numerous indentations and longitudinal nuclear grooves.
- Nuclei are often hyperchromatic, but occasionally can be normochromatic or even hypochromatic.
- Normally, coarse chromatin is spread uniformly, although occasionally it can be fine.
- In cases of endocervical glandular extension, nucleoli are occasionally present but are typically absent.
- From "immature" thick "metaplastic" with focal vacuolation to sporadic heavily keratinized cytoplasm, the cytoplasm varies.

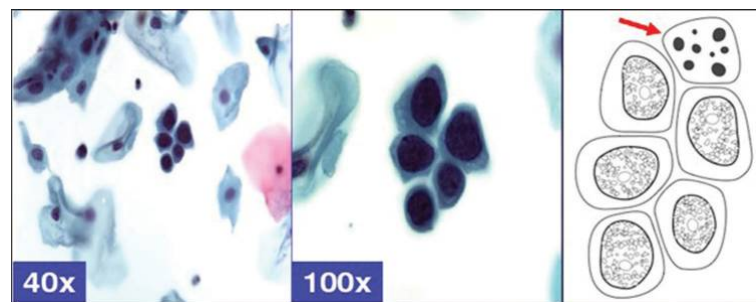


Figure 2.3: HSIL has a checkerboard pattern with black nuclei, smudged chromatin, nucleoli, and unevenly spaced apoptotic bodies [5].

2.3.2.3 ASC-US

ASC-US is considerably more common with cytomorphological traits showing more nuclear atypia than reactive alterations. Due to qualitative factors with or without quantitative restrictions, such as abnormal changes occurring only in a small number of cells, the atypical traits are ambiguous for definite dysplasia. There aren't any specified ASC-US requirements. The characteristics might range from a slight increase in nuclei size to others like nuclear hyperchromasia and uneven nuclear membranes [5]. Some key characteristics are [5]:

- 2.5–3 times increase in nuclear size ICN.
- Slight rise in the nuclear cytoplasmic ratio.
- Mild hyperchromasia and uneven nuclear structure.
- Atypical parakeratosis with equivocally atypical nuclei and orangeophilic cytoplasm.
- Unusual repair with a "School of Fish" streaming pattern and pronounced nucleoli.

2.3.2.4 ASC-H

Cases with cellular alterations that are ambiguous for high-grade dysplasia due to either qualitative or quantitative constraints are given the designation ASC-H. ASC-H instances must be properly classified because they may go downstream and be managed in a completely different way [5]. Some key characteristics which define ASC-H are [5]:

- 2.5–3 times the ICN nuclear size.
- N/C ratio slightly rising.
- Nuclear irregularity and mild hyperchromasia.
- Other characteristics overlapping with HSIL but not clearly.

2.3.2.5 SCC

Depending on the severity, squamous cell carcinoma or HSIL with characteristics suspicious of invasive squamous cell carcinoma should be defined according to the presence of necrotic debris, dysplastic tadpole-shaped cells, or fiber cells. The cytoplasm can be "immature" dense "metaplastic" with focal vacuolation or sporadically densely keratinized [5].

Some key characteristics which define SCC are [5]:

- Nuclear atypia: Nuclei are often larger and more irregular in shape than normal cells.
- Hyperchromasia: The nuclei may also be darker in color than normal cells.
- Mitotic figures: May contain more mitotic figures, which are cells in the process of dividing.
- Necrosis: May show evidence of necrosis, which is the death of cells due to a lack of blood supply.
- Inflammation: SCC may also be associated with inflammation, which is the presence of immune cells in the tissue.

Chapter 3

Datasets and Previous work

3.1 Overview

With the development of computer-aided technology, cervical cancer screening has improved significantly over the years, allowing to better identify abnormal cells and make correct treatment decisions [88]. These algorithms, mostly rely upon machine learning, which depends on large datasets which are often not balanced or equally distributed among the different categories. In addition, the misclassification cost is not typically considered in the general classification process [88]. The existing datasets of cervical cancer images, are often unbalanced and have different image qualities, decreasing the judgment of the machine learning algorithms [91]. On the other hand, studies have shown that cervical cells have intrinsic similarities. For example, intermediate and superficial cells both have nuclear margins, clear cytoplasm, and small nuclei, while the cytoplasmic and nuclear borders of dyskeratotic and metaplastic cells overlap. These findings suggest that there may be a connection between cervical cell images [44].

3.2 μ SmartScope

There is a need for more affordable and effective solutions to improve cervical cancer screening. This is because current methods, such as the ThinPrep Imaging System (TIS) and the BD FocalPoint GS Imaging System (FocalPoint), are expensive and not widely accessible [42]. These automated microscopy systems can provide high-quality images for the diagnosis of cervical cancer, but there is still a need for the development of more affordable computer-aided diagnosis systems to assist in the identification of cervical lesions [73]. By finding new and innovative solutions that are both effective and affordable, it is possible to improve the screening and diagnosis of cervical cancer.

Fraunhofer AICOS has been developing the μ SmartScope [71], a completely automated 3D-printed smartphone microscope, which has the objective of being a cheap and automated replacement for conventional microscopes. This prototype serves as a cost-effective alternative to traditional microscopes and is specifically designed to facilitate microscopy-based diagnosis in areas

with limited access to healthcare services [73]. The μ SmartScope enables autonomous microscopic image capturing by using a motorized stage that is totally powered and controlled by a smartphone, with the ultimate goal of relieving the strain of manual microscope inspection. The μ SmartScope also intends to lessen reliance on on-site professionals in microscopy diagnostics by making it simple to integrate with artificial intelligence (AI). The tool is now being used for automated analysis of blood smears contaminated with malaria [70] and is already redesigned to accommodate cervical cancer screening [61]. This allows it to achieve a solution that satisfies the requirements for the accurate microscopic examination of liquid-based cervical cytology samples [73].

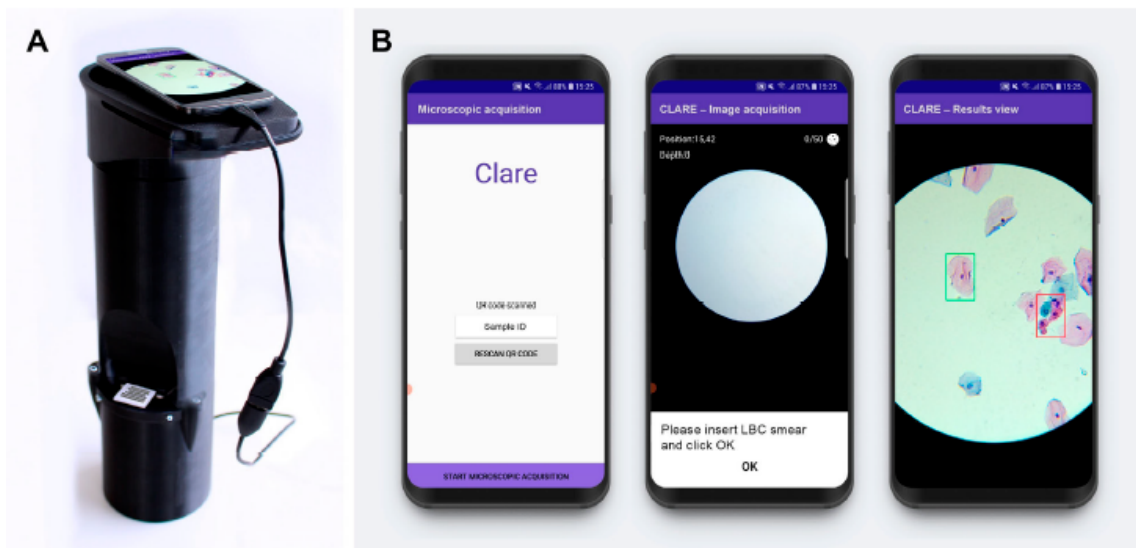


Figure 3.1: (A) μ Smartscope with phone and LBC sample (B) Smartphone application screenshots [73].

3.3 AI-based Cervical Lesions Detection

The literature extensively covers a variety of computer vision problems, including cell detection, segmentation, and counting. The most effective strategy for addressing a particular issue depends on the desired outcome. For example, density estimation simply provides the total number of objects, while detection methods provide localization through bounding boxes and the respective class. In contrast, segmentation techniques allow for the acquisition of object masks and classifications. Following the use of the μ SmartScope for cervical cytology, there are two main approaches that are being taken regarding the cervical cancer detection models: the Region-based approach and the Nuclei-based approach. These two methods utilize cytological images as a dataset but differ in their method of analysis.

The Region-based approach divides each patch into regions, which are defined as cells or clusters of cells with the same abnormality class according to the Bethesda system. This methodology takes into account how the different classes aggregate, as seen in section 2.3.2, as well as the

specific cytomorphological features of the cells, such as the size and the nuclei-cytoplasm ratio, which have been shown to be important indicators of different classes of anomalies [73].

The Nuclei-based approach only takes into consideration the nuclei of the cells. This method takes advantage of the fact that the nuclei are often the decisive factor when classifying a neoplastic change. It also takes into consideration that the nuclei are not as variable as the different regions, making it easier for the model to understand what are the specific characteristics which define each abnormality class [53]. Figure 3.2 presents an overall scheme summarizing both approaches.

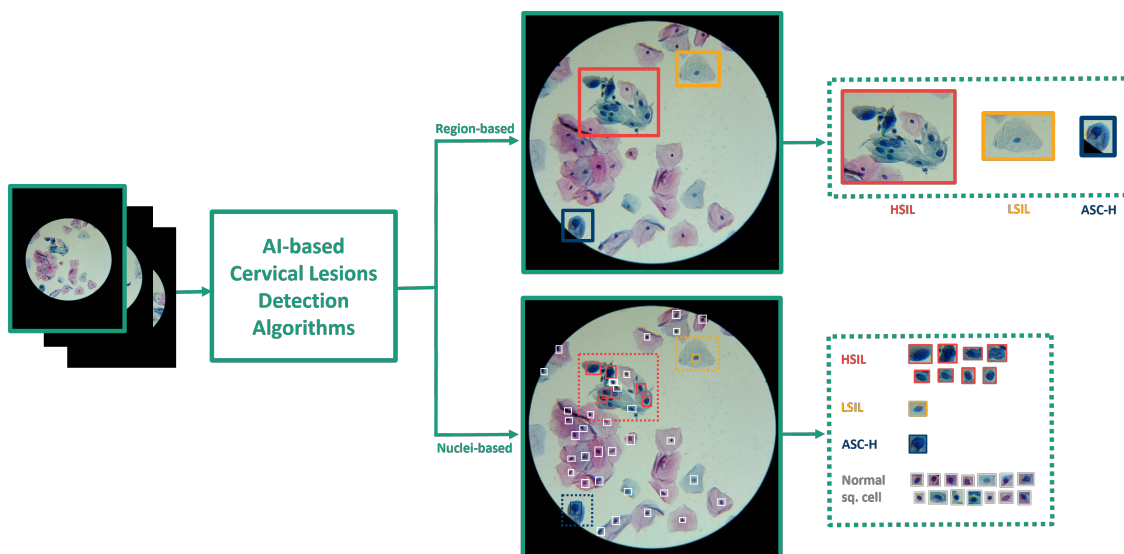


Figure 3.2: Overall Scheme comparing the Region-based approach to the Nuclei-based approach.

In the next two sections, both approaches are going to be examined in depth, giving a detailed description of each dataset, and the corresponding detection model.

3.4 Region-Based Approach

In this section, it is provided an in-depth description of the dataset utilized in the Region-based approach and a comprehensive examination of the detection model employed.

3.4.1 Dataset

The dataset used for the Region-based approach is a mobile-acquired picture database called the mobile HFF regions dataset [73] (hereinafter referred to as Region-based dataset). It is made up of 21 LBC samples, provided by Hospital Fernando Fonseca [22], from various clinical cases. The images were captured using a smartphone and then manually annotated by a specialist, which identified the presence of abnormal cells or cell aggregates of cervical regions. Annotations are made according to the Bethesda system convention, being shown as bounding boxes surrounding the abnormal regions with a classification label reflecting each region's lesion level [73]. Figure

3.3 demonstrates the diversity of structures that could be related to the same lesion level as well as the similarity between some cells of subsequent lesion levels.

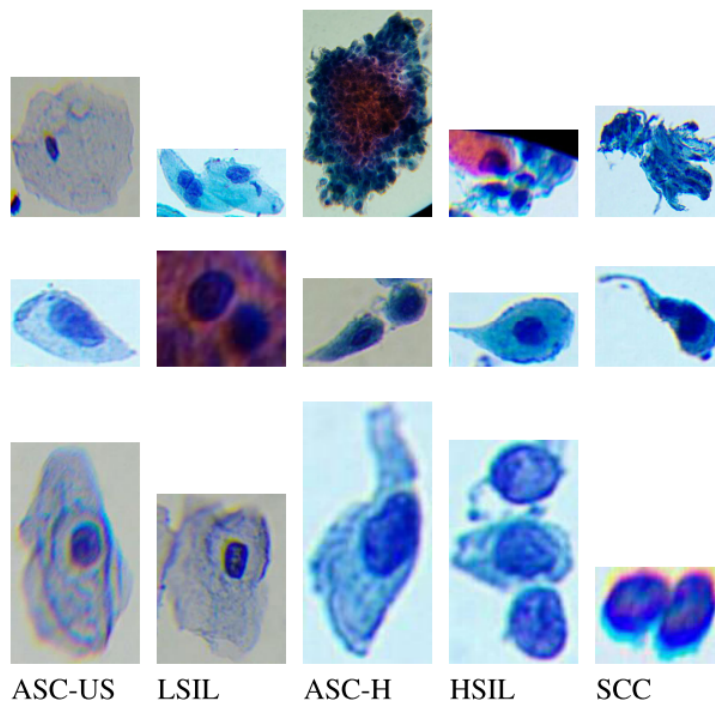


Figure 3.3: Examples of the five lesion classes considered in the mobile Region-based dataset [12].

The images obtained with the μ SmartScope were separated into neighboring patches to obtain images of fixed dimensions, which are necessary for the application of some of the detection models and to limit the computational resources used during training. Each patch was extracted taking into account the identified regions present in its area. Procedures were also applied to segment the optic disc and crop the main field of interest in accordance with the segmented region after pre-processing the obtained images [73].

Although the proportion of samples for each diagnosis outcome is generally even in the mobile Region-based dataset, Table 3.1 shows that the distribution of abnormal regions is not balanced for all lesion levels and that there are few clinical instances of each class [73].

Number	ASC-US	LSIL	ASC-H	HSIL	SCC	Total
Samples	4	3	4	3	2	16
Train annot.	352	58	79	203	13	705
Test annot.	125	38	30	29	0	222
Total annot.	477	96	109	232	13	927

Table 3.1: Region-based dataset sample and annotation distribution (training, test and total) [73].

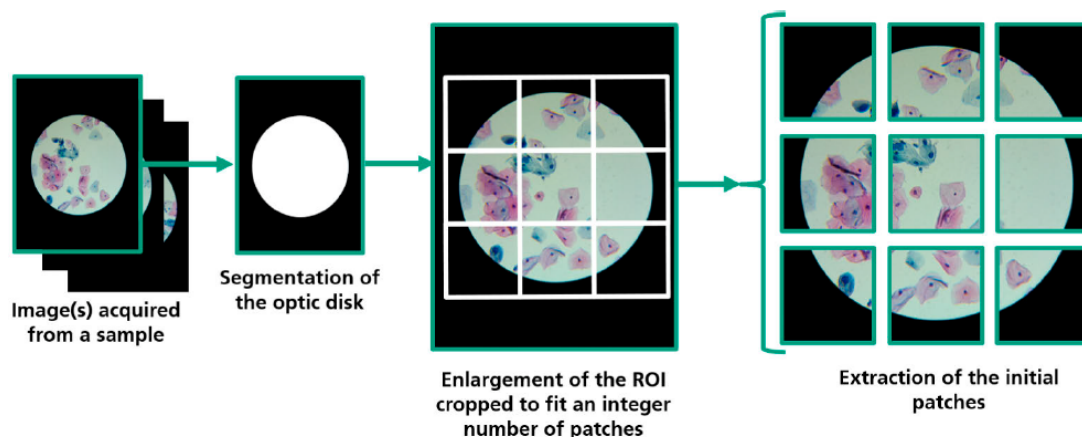


Figure 3.4: Patch extraction process [73].

3.4.2 Detection model

The Region-based dataset was gathered with the objective of training a cervical cancer screening detection model. The Faster R-CNN with a ResNet50 backbone achieved the best performance for the Region-based approach, which is a two-stage detector with slower inference speeds and complexity that may be a deterrent for mobile settings, but it exhibits the most dependable performance in terms of mean average precision (mAP) and detection sensitivity, especially when applied to small objects [67]. This study was based on networks pre-trained in the public Common Objects in Context (COCO) dataset [47], which was tailored to the cervical cytology context.

In this model, the most prevalent class was ASC-US, which had the best classification performance. Some ASC-US cases, on the other hand, were misclassified as other lesion levels, presumably due to the variety of cellular alterations that fall under this level. The null values for LSIL were probably caused by the lack of detections for this class, indicating that the network was unable to learn its usual characteristics due to its underrepresentation in the training data [73]. The majority of the instances that were incorrectly classified as belonging to adjacent lesion levels were, however, misidentified as belonging to the ASC-H and HSIL-SCC classes, despite the model’s reasonable ability to discriminate between these regions. This may be because cervical lesions are progressive, which leads to common characteristics. These factors allow us to conclude that the model’s classification performance was adequate [73]. Table 3.2 gives more details on the robustness of the model for each class.

Class	ASC-US	LSIL	ASC-H	HSIL-SCC	Avg.
mAP@.50IOU	0.008573	0.000744	0.00706	0.03836	0.01368
AR@10	0.25754	0.13125	0.27692	0.32973	0.24886

Table 3.2: Class-wise performance of the Faster R-CNN C Resnet50 model [73].

Overall, the provided metrics were below average, with the LSIL class exhibiting the worst

performance due to its inadequate representation in the sample. Due to its more distinguishing characteristics, along with the larger dimensions of the regions of interest and the greater amount of training data available, the HSIL-SCC lesion level was the one that the network was able to detect most effectively. This model performed worse on the test data compared to the cross-validation results, achieving mAP@:50 and AR@10 values of only 0.01368 and 0.24886 respectively [73].

3.5 Nuclei-Based Approach

The Region-based approach for the mobile detection of cervical cancer lesions achieved promising results for cervical cancer screening. However, the authors acknowledged limitations in the study, including the limited data volume and high structural variability of the Region-based dataset. The Nuclei-based approach aims to address these limitations by utilizing digitalized LBC samples obtained through the μ SmartScope instrument to develop a system for a nuclei-based cervical lesion detection algorithm, which would also adhere to TBS requirements [53].

This approach is a derivation of Mosiichuk's work [54], which focuses on creating an automated system to evaluate the suitability of cervical cytology samples. The work employs deep learning models to perform the recognition and classification of various classes of nuclei and other objects, taking into account the state-of-the-art object detection methods. In this work, the class of squamous nucleus achieved an AP of 82.4%, an accuracy of 79.8%, and an F1 score of 81.5% at the picture level. A total of 5216 predictions were made for 5483 annotations in terms of raw detection, resulting in a true positive rate of 74% and 473 false positives. The primary goal of the Nuclei-based approach is to investigate the use of nuclei localization algorithm techniques applied in the adequacy assessment system, to enhance the detection of cervical squamous lesions [53].

3.5.1 Dataset

Several datasets with cervical cell annotations such as Cervix93 [62], Herlev [40], SIPaKMeD [64], and ISBI Challenges [51] are publicly available. However, these datasets have limitations in terms of their usefulness for nuclei and lesion detection tasks. For example, the Herlev dataset only contains isolated images of cells with annotated abnormalities, while datasets such as Cervix93, ISBI Challenges, and SipakMeD include images of microscopic fields with annotated nucleus regions, but lack annotations for cervical lesion locations. The more recent CRIC dataset [68], in contrast, contains images of microscopic fields with annotated cervical lesion locations, but does not provide information about nucleus structures.

The Nuclei-based dataset was made to enable the creation of an efficient lesion identification module and to address the shortcomings of existing publicly available datasets. The dataset consists of microscopic images acquired using the " μ SmartScope" equipment with a 40x amplification of LBC samples. It comprises 139 samples, each represented by approximately 100 images [53]. Out of the 139 samples that were collected and digitized, a mere 21 samples had their lesion locations identified in accordance with the TBS guidelines by trained professionals, which are the samples used in the Region-based dataset mentioned above 3.4.1. The annotations are provided

in the form of bounding boxes and cover the cytoplasm, the nucleus, and in some cases, multiple cells. However, the annotations of the regions were modified to focus solely on the nuclei of cervical lesions [53]. The main objective of this modification was to create a dataset that would be suitable for the development of a cervical lesion detection module based on the identification of nuclei lesions. This would allow for improved recognition of cervical nuclei with lesions through the use of transfer learning techniques and by drawing from the findings of the adequacy evaluation module [53]. Figure 3.5 demonstrates an output of the transformation process, in which all the lesions from all the squamous nuclei were considered to possess a lesion level corresponding to the class of the lesion in which it was enclosed. The bounding boxes of Figure 3.5 are as follows: In the Adequacy Assessment Dataset, squamous nuclei are denoted by green bounding boxes, lesion annotation of a cell is represented by white bounding boxes, and the transformation of squamous nuclei to ASC-US nucleus annotation is represented by an orange bounding box [53].

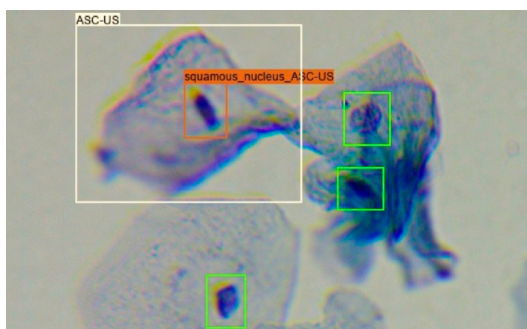


Figure 3.5: Example of the Nuclei-based dataset annotations [53].

The final dataset is a combination of the Region-based dataset and the one used in the automated adequacy work [54], and Table 3.3 illustrates the final number of nuclei annotations per TBS lesion class in the Cervical Nuclei-based dataset.

Class	ASC-US	LSIL	ASC-H	HSIL	SCC	Normal	Total
Nuclei Annotations	768	144	132	329	22	31698	33093

Table 3.3: Distribution of the Nuclei-based dataset by class [53].

3.5.2 Detection Model

The cervical lesion detection module has the objective of identifying and categorizing cervical lesion nuclei. Being an adaptation of the previous work made by Mosiichuk’s [54] it followed a pipeline that closely resembles the one from the nuclei detection for adequacy evaluation. The model aims to specifically locate squamous nuclei within regions of squamous cervical lesions and classify them based on their corresponding lesion grade, in contrast to the previous nuclei model which focused on detecting all classes of nuclei and classifying them according to their respective cell class. This constitutes the primary differentiation between the two modules [53].

Regarding the train/test division, since the Cervical Lesions Dataset was previously utilized in the Region-based approach [73], the division used in this study was also the same. This includes the implementation of the patch-slicing procedure, where images were divided into patches of fixed dimensions. The training data was downsampled to balance the number of empty patches utilized for training. Despite the fact that the annotation type was changed from areas to nuclei, the images were the same, enabling a fair comparison of the results produced by the two methods [53].

Using the Adequacy Assessment model weights for transfer learning and the dataset with normal squamous nuclei annotations, the optimum configuration was obtained by having an LR of $4.862e-5$ and a Batch Size of 16 and using the model SSD ResNet50 V1 FPN [53].

In general, the model’s performance suggests that it is able to detect and classify instances of ASC-US and ASC-H relatively well, but it struggles to detect instances of LSIL and has a trade-off between precision and recall for HSIL. The model’s performance for the Normal class is relatively good but has a high rate of misclassification. The performance of the model on the individual classes and overall can be improved by using a larger and higher-quality dataset, fine-tuning the model’s parameters, and using more advanced architectures.

It is also feasible to compare the model developed in this study with that of the Region-based model [73], as both utilize the same dataset and the split of the test set comprises identical images. However, it should be noted that the images in the two datasets are divided into smaller patches, and the annotations for the two datasets differ, thus comparisons must be made with caution. The results are presented in Table 3.4.

Metric	Model	ASC-US	LSIL	ASC-H	HSIL	Squamous
mAP@0.50	Region-based	0.0086	0.0007	0.0070	0.0384	
	Nuclei-based	0.0160	0.0097	0.0177	0.0305	0.8649
AR10	Region-based	0.2575	0.1312	0.2769	0.3297	
	Nuclei-based	0,6763	0,4820	0,6968	0,6457	0,7034

Table 3.4: mAP@0.50 and AR10 of the Nuclei and the Regions detection models [53, 73].

The model performed better for all classes except for HSIL regarding the mAP@0.50. However, the paper mentions the class imbalance of the dataset had an impact on the network’s recognition of the nuclei of the underrepresented classes, resulting in a bias towards the dominating class (ASC-US). Although this class had the most predictions, a majority of them were false positives, leading to a decrease in the model’s precision [53]. In Table 3.4, the Nuclei-based model’s performance is measured using the AR10 metric for the purpose of comparison with the Region-based approach. While the AR100 metric may provide a more accurate representation of the Nuclei-based model’s performance, the AR10 metric demonstrates a significant improvement in all classes when using the Nuclei-based approach.

The methodology employed in this study demonstrated superior performance compared to that of [73], maybe as a result of the greater similarity of the images that were employed. Furthermore, the use of annotations of nuclei instead of regions present in the original lesion dataset may have resulted in reduced intra-class morphological variability, contributing to the improved performance [53]. Despite the apparent success of the suggested methodologies, it is important to note that the quality of the dataset plays a crucial role in the development of deep learning systems. It is believed that the use of higher-quality, larger datasets would lead to enhanced performance and potentially enable cervical lesion identification to achieve significant advancements [53].

3.6 Deep Feature Consistent Variational Autoencoder

This chapter presents a research study conducted at Fraunhofer AICOS on the use of a Deep Feature Constant Variational Autoencoder (CNN-VAE) [36] for synthetic image generation.

The aim of this experiment was to evaluate the capabilities of the CNN-VAE in generating realistic synthetic images of single cells. To do this, it was created a modified version of the Region-based dataset, which consists of single cell images cropped from the multi-cellular images of the Region-based dataset. For each single cell image is also provided the respective annotation regarding the TBS abnormality level. The goal was to assess the ability of the CNN-VAE to generate synthetic images that capture the details and diversity of single cells as well as their context within larger structures. The architecture of the model can be seen in Figure 3.6.

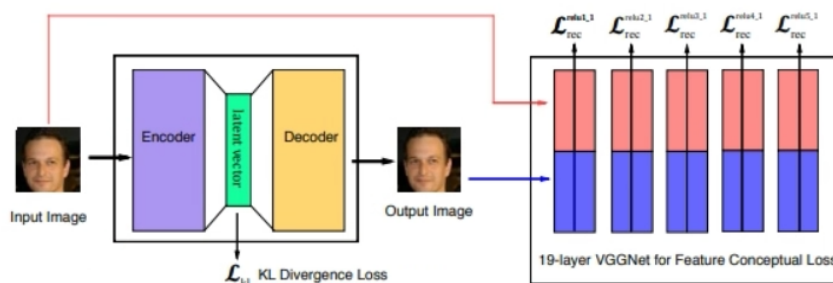


Figure 3.6: Model overview. The left is a deep CNN-based Variational Autoencoder and the right is a pre-trained deep CNN used to compute feature perceptual loss [36].

The results of the experiment showed that the CNN-VAE was able to successfully reconstruct single cell images. Specifically, the CNN-VAE was able to capture important features of the cells such as shape and texture but was not able to synthesize novel images that were not present in the original dataset. Some illustrative examples of the reconstructed images can be seen for a different number of epochs in Figure 3.7.

In addition to evaluating the CNN-VAE’s ability to generate synthetic images, it was also explored its potential for modifying images through latent space manipulation. Specifically, it was attempted to transform images of some abnormal class, such as ASC-H cells into images of another class, such as ASC-US cells via the linear interpolation of the latent vector, as seen in Figure 3.8.

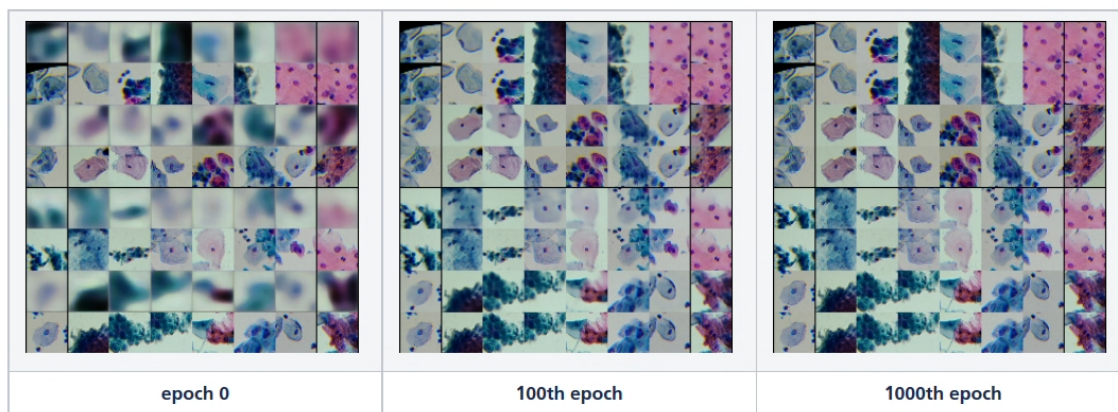


Figure 3.7: Results of the reconstructed images with 256x256 for the different numbers of epochs (images on top are the reconstruction).

While the results of this experiment were not as successful as it was hoped, it did provide insight into the limitations of the CNN-VAE and suggested avenues for future improvement.

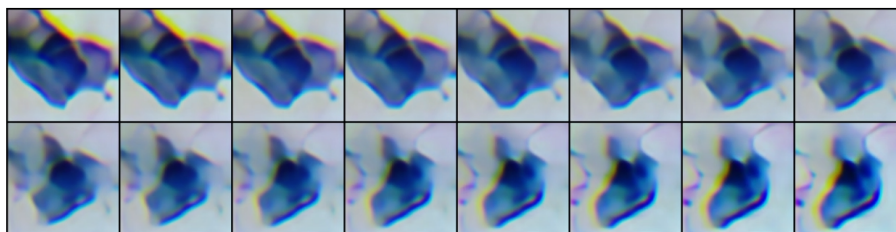


Figure 3.8: Transformation of ASC-H in ASC-US via interpolation.

Overall, this study demonstrated that the CNN-VAE is able to reconstruct images well from the latent space generated by the encoder. However, it was unable to generate images from a random latent space. These findings suggest that the CNN-VAE has limitations in its ability to generate novel synthetic images. Nonetheless, this dissertation will use the dataset from this research in order to generate single cells and by building upon this previous study, we hope to produce more realistic and diverse synthetic images of single cells.

3.6.1 Single Cell Dataset

The dataset used in this research is composed of single cells and nuclei, cropped from the Region-based dataset. This dataset contains images with resolutions of 128 and 256 pixels and is divided into two categories: Regions and Nuclei instances. As shown in Table 3.5, the single cell dataset regions instances include a diverse set of cell classes, including ASC-US, ASC-H, LSIL, HSIL, SCC, glandular cells, pavimentosa, and transformation zone cells. The total number of single cell dataset regions instances is 1,547.

It is worth noting that, while the dataset includes a variety of cell classes, the number of instances for each cell class is not evenly distributed. For example, the number of ASC-US cells is more than 26 times larger than the number of SCC instances. This clear unbalancing needs to

Cell class	Instances
ASC-US	352
ASC-H	79
LSIL	58
HSIL	203
SCC	13
Glandular cells	3
Pavimentosa	831
Transformation zone	22

Table 3.5: Number of single cell dataset regions instances for each cell class.

be taken into account during the development of models trained with this dataset. Three instances of each class can be observed on image 3.9

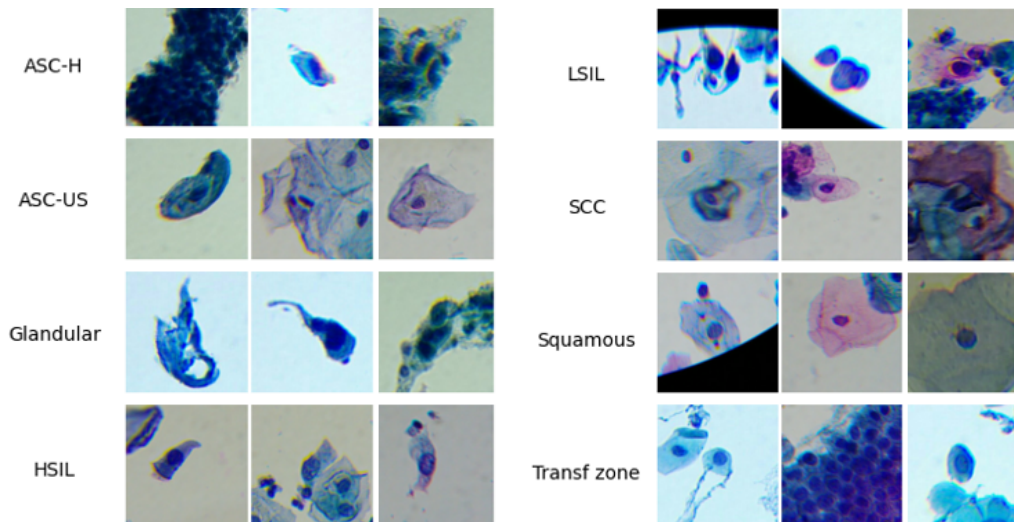


Figure 3.9: Illustrative examples of single cell dataset regions instances for each class.

In Figure 3.9 it is clear that the regions instances do not contain only single cells but also regions. This should be taken into consideration when generating a model of single cells.

Regarding the single cell dataset nuclei instances, the class distribution can be seen in Table 3.6.

From Table 3.6 we can conclude that the Nuclei-based dataset is also not well distributed between the classes, and has a total of 1098 instances. From image 3.10 it is possible to infer that the nuclei images have lower resolution since the nucleus represents a very small portion of each patch. The images are also much more homogeneous, being harder to understand the differences between the different abnormalities classes.

Cell class	Instances
ASC-US	598
ASC-H	101
LSIL	83
HSIL	316

Table 3.6: Number of single cell dataset nuclei instances for each cell class.

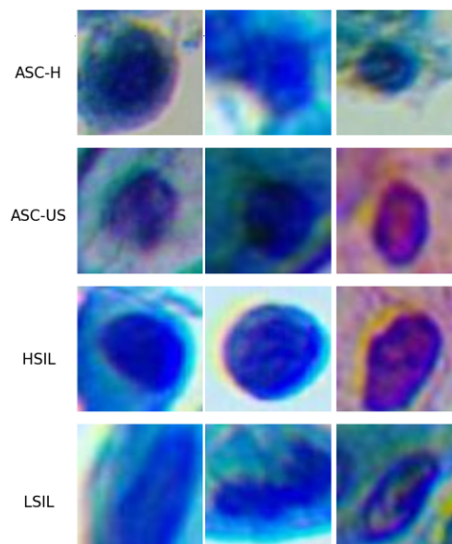


Figure 3.10: Illustrative examples of single cell dataset nuclei instances for each class.

Chapter 4

Generative AI

4.1 Overview

Generative Modelling was created in 1989, with the goal of utilizing neural networks with the goal of learning data without supervision, potentially benefiting standard classification tasks. In recent years, the use of generative modelling has accomplished results in various areas, being able to synthesise videos, images, text-to-image, summarising, liquid simulation, among many others [83]. In the initial years, generative models achieved their results by defining an energy function on data points, proportional to likelihood. These models struggled to scale to complex, high dimensional data, and required the laborious Markov Chain Monte Carlo (MCMC) sampling method, which is an iterative slow process. More recently, generic deep learning architectures and generative models have advanced, setting new standars for visual fidelity and sampling speed, leading to an increasing interest in the scientific community of what they can accomplish [28]. There are many different approaches to generative models, such as generative adversial networks, variational auto encoders and diffusion models. Each one has its own advantages and disadvantages, and they try to reach an optimum in speed and quality, among other factors.

4.2 Variational Auto Encoders

Variational autoencoders are a type of generative model that are particularly suited for the task of modifying images via concept vectors. They were concurrently proposed by Kingma and Welling in December 2013 [45] and Mohamed, Rezende and Wierstra in January 2014. They combine concepts from deep learning and Bayesian inference to create a contemporary version of autoencoders, a class of network that seeks to encode input into a low-dimensional latent space and then decode it back. [45] A traditional autoencoder starts with an input image, maps it to a latent vector space using an encoder model, and then uses a decoder to return the output to the original image's dimension. Posteriorly, the autoenconder learns to recreate the original inputs by being trained using target data that contains the same images as the input images. The output can be subjected to a variety of limitations to force the encoder to learn more-or-less interesting latent

representations of the data. Usually the code is constrained to be low-dimensional and sparse, so the encoder functions as a compressor of the data [14]. VAE converts the input image into the mean and variance of a statistical distribution rather than compressing it into a fixed code in the latent space. In essence, this indicates that its presuming that a statistical process, which produced the input image, was random and that this randomness should be taken into consideration while encoding and decoding [15]. The VAE then randomly selects one element from the distribution using the mean and variance parameters, decodes that element, and reconstructs the original input from that element. Every point sampled in the latent space is decoded to a valid output, which increases robustness and forces the latent space to encode meaningful representations everywhere. [15]

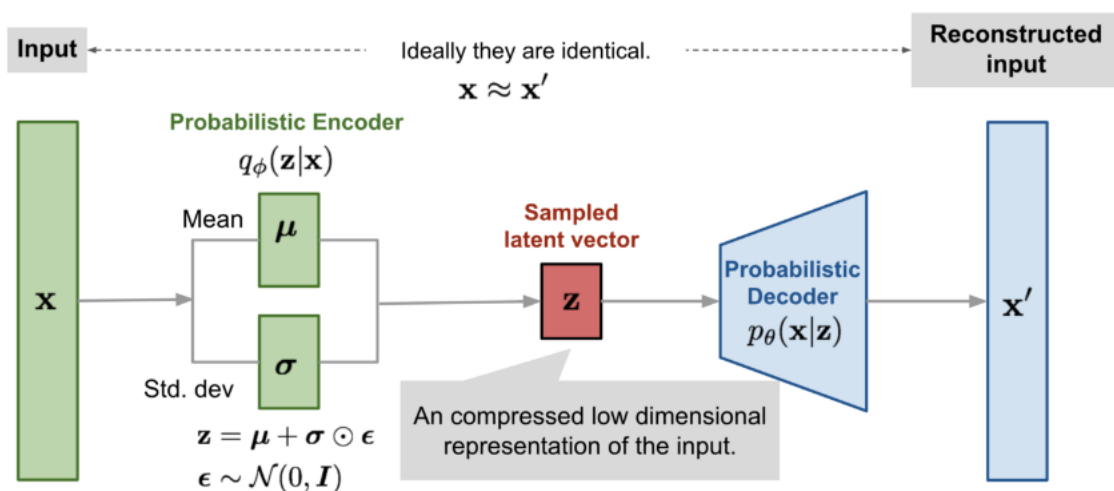


Figure 4.1: Variational Auto Encoder global architecture [76].

4.3 Generative Adversarial Networks

Generative Adversarial Networks, have been raising popularity in the scientific community, being able to solve different problems such as generative tasks, image synthesis or even classification and super-resolution [12]. They accomplish this by using a competitive approach utilizing two networks to derive back-propagation signals. Generative Adversarial Networks are divided in two, a Discriminator, whose goal is to classify the input image as synthetic or real, and the generator, which generates the most realistic synthetic images possible [19]. These two networks compete with each other, and the generator is never shown the training dataset, only having as input the gradient of the discriminator decision. [8]

Generative adversarial networks have been used to balance the weight and increase the scale of datasets before, as well as number of samples in each category [12]. These networks have the power to generate completely synthetic realistic images, such as landscapes, human faces, and more specifically to this work, medical images. Pollastri introduced a novel method for the task of segmenting skin lesions that makes use of DC-GAN to enhance the data, which creates

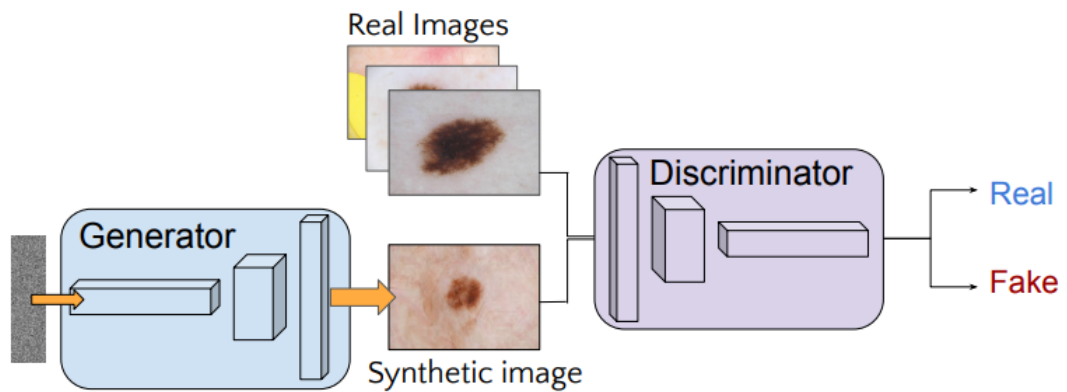


Figure 4.2: Simplified GAN Architecture in skin lesion synthesis [12].

synthetic images of skin lesions and its segmentation masks [65]. Karras created a new design which automatically learned and separates high level attributes and stochastic variance, completely unsupervised [43]. To enhance classification performance, they suggested fusing deep learning, transfer learning, and generative adversarial networks [81].

4.4 Diffusion models

Another set of methodologies which are based on likelihood estimation methods and take inspiration in physical phenomenon are Diffusion Models. It is based on a concept of thermodynamics, which describes how molecules diffuse from dense locations to less dense ones, mostly known as heat death or increase of entropy [75]. According to information theory, this is equivalent to information loss brought on by gradual intervention of noise. In 2020 some seminal papers have shown that this new technology is capable of beating Generative adversarial networks on image synthesis [57]. Diffusion models can be used to create data that is comparable to the data they were trained on, by adding gaussian noise and then learning to recover the data by undoing this noise-adding process. When the model is trained, it is possible to generate new images, by applying the mastered denoising technique to randomly sampled noise [84]. After learning the distribution, the model may generate useful data from random noise, by converting a latent encoded representation of image data into a more insight full representation. In this point of view, Diffusion models and autoencoders are comparable.

The complete process can be summed up as a two-step phenomenon, as shown in 4.3, with the forward pass (X_i to X_T) transforming the data distribution into noise and the reverse pass converting the noise distribution into the data distribution (X_T to X_i). A diffusion model must be trained in order to learn the reversing process, or $p(x_{t-1}|x_t)$. A neural network can be used to implement the forward and reverse training phases for the diffusion model. However, the input and output dimensions of the design must be identical [4].

The forward process begins with the data $x \sim p(x)$ and gradually adds noise to obtain a noisy version of the data $z = \{z_t | t \in [0, 1]\}$. The reverse process reverts the forward process by predicting

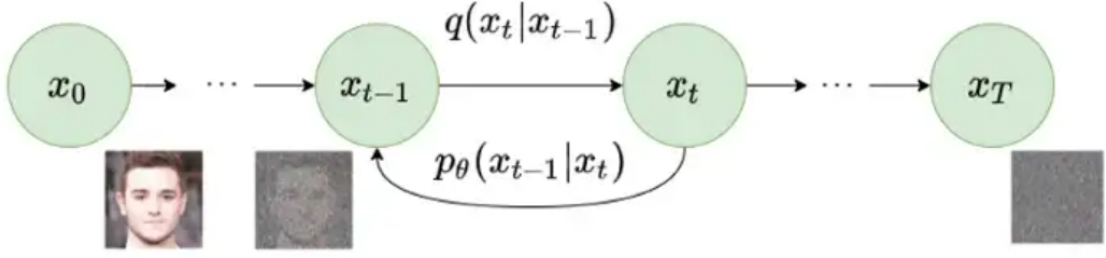


Figure 4.3: Diffusion Model simplified Architecture [35].

and subtracting the noise in the reverse direction. Formally we define the forward process $q(z|x)$ specified in continuous time $0 \leq s < t \leq 1$ as:

$$q(z_t|x) = N(\alpha_t x, \sigma_t^2 I), q(z_t|z_s) = N((\alpha_t/\alpha_s)z_s, \sigma_{t|s}^2 I)$$

where $\alpha_t^2 = 1/(1 + e^{-t})$ and $\sigma_t^2 = 1 - \alpha_t^2$ are the continuous-time noise schedules, $\sigma_{t|s}^2 = (1 - e^{-\lambda_t - \lambda_s})\sigma_t^2$ is the variance term of the s to t transition, and $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$ is the signal-to-noise ratio of the noise schedules that is monotonically decreasing [4]. The forward process can be written in the reverse direction as

$$q(z_s|z_t, x) = N(\tilde{\mu}_{s|t}(z_t, x), (\tilde{\sigma}_{s|t}^2)I)$$

where

$$\tilde{\mu}_{s|t}(z_t, x) = e^{\lambda_t - \lambda_s}(\alpha_s/\alpha_t)z_t + (1 - e^{\lambda_t - \lambda_s})\alpha_s x$$

and

$$\tilde{\sigma}_{s|t}^2 = (1 - e^{\lambda_t - \lambda_s})\sigma_s^2.$$

The reverse process is parameterized by a generative model \hat{x}_θ in the form:

$$p_\theta(z_s|z_t) = N(\tilde{\mu}_{s|t}(z_t, \hat{x}_\theta(z_t, \lambda_t)), \Sigma_{s|t} I)$$

where the variance $\Sigma_{s|t} = (\tilde{\sigma}_{s|t}^2)^{1-\nu}(\sigma_{t|s}^2)^\nu$ is an interpolation between $\tilde{\sigma}_{s|t}^2$ and $\sigma_{t|s}^2$ [4], and ν is the hyperparameter that controls the stochasticity of the sampler. The ancestral sampler is used $\lambda_0 < \dots < \lambda_T = \lambda_1$ for discrete T time steps:

$$z_s = \tilde{\mu}_{s|t}(z_t, \hat{x}_\theta(z_t, \lambda_t)) + q\Sigma_{s|t}^{1/2}\varepsilon$$

where $\varepsilon \sim N(0, I)$ [90].

Whereas in the past the generating capacity of diffusion models was mostly exploited for unconditional production of data, recent attempts have demonstrated conditioned generation by incorporating guided-diffusion models. With their capacity to learn the representation, diffusion models have the potential to be applied in many areas of medicine, being able to produce a wide

variety of medical images, and even expand upon current methods, such as domain-to-domain translation, noise adaptation, noise reduction, super-resolution and data-augmentation.

The diffusion model has its basis on the idea of diffusion maps, being one of the dimensional reduction strategies employed in machine learning literature. It uses Markov Chains, which are widely employed, and was first proposed by Shol-Dickstein [77]. Numerous papers such as the one from Dhariwal, P., & Nichol, A. [23] have proven it can have better results in image synthesis than the widely used Generative Adversarial Networks, being an interesting new approach to this problem [23]. Compared to Generative adversarial networks, they have a large sample diversity and mode coverage, and often outperform them in high sample quality. Additionally, these models have demonstrated that they are resistant to Mode Collapse and can provide a wider variety of images [82]. Due to their advantages over previous methods, diffusion models have already been used in some research papers applied to medicine. T1w MRI images of the brain were produced using latent diffusion models by Walter et al [63]. They created a stack of 100,000 images conditioned on important factors like age, sex, and brain volume using 31,740 brain MRI data from the UK Biobank. In the study "Spot the fake lungs" [4], they used neural diffusion models to produce artificial CT and X-ray pictures of the lungs by using DALLE2 model and the steady diffusion model, having been able to trick even a specialist, proving that data scarcity for medical imaging can be at least partially resolved by diffusion models. The authors demonstrate how pretraining on generated data can enhance downstream performance on a segmentation job (breast MRI segmentation, Dice score 0.95 (+0.04)) in a limited data situation. However, even though these models were trained using resolutions that were significantly lower than those used in clinical practice, up to 32% of the 50 images that were examined by radiologists revealed significant anatomical discrepancies.

Based on the evidence gathered, we have decided to employ diffusion models over GANs in this research work, as they have recently consistently demonstrated superior performance and robustness for synthetic image generation.

4.4.1 Latent Diffusion Models

Currently, there are three popular text-to-image models: Stable diffusion [69], Dalle 2 [66], and Midjourney [52]. These models all utilize a forward process that adds noise to the data and a reverse process that tries to remove the noise in order to reconstruct the original data. However, they differ in their specific formulations and the types of applications for which they are most suitable. These three models are based on latent diffusion models, a variant of diffusion models that operate in a lower-dimensional latent space, rather than on the original high-dimensional input. This approach was introduced by Rombach *et al.* [69] and aims to reduce the computational complexity of training diffusion models by encoding the input into a latent representation using an encoder network. The latent representation is then processed by a standard diffusion model (e.g., a U-Net), which generates new data that is up-sampled back to the original dimensions using a decoder network. This approach has been shown to be effective in generating high-quality synthetic data and has the potential to be applied to a wide range of tasks [69].

Latent diffusion models consist of three parts:

- The autoencoder (VAE)
- A U-Net.
- A text encoder.

The VAE

There are two components that make up the VAE model, the encoder and the decoder. The encoder transforms the image into a low-dimensional latent representation, which will be utilized as the U-Net model's input. The latent representation is converted back into a picture by the decoder. The encoder is used in latent diffusion training to obtain the latent representations (latents) of the pictures for the forward diffusion process, which introduces increasing amounts of noise at each stage. The reverse diffusion process produces denoised latents, which the VAE decoder transforms back into pictures during inference. Latent models only use the VAE decoder [24].

The U-Net

ResNet blocks are used in both the encoder and decoder portions of the U-Net. A higher resolution picture representation that is purportedly less noisy is converted from a higher resolution image representation by the encoder into a lower resolution image representation by the decoder. In more detail, the U-Net output forecasts the residual noise, which can be used to calculate the forecasted representation of a denoised image. In order to prevent the U-Net from losing crucial information when downsampling, short-cut connections are created between the downsampling ResNets of the encoder and the upsampling ResNets of the decoder. The stable diffusion U-Net can also use cross-attention layers to condition its output on text-embeddings [24].

The text-encoder

The text encoder is a transformer-based model that maps an input sequence of tokens to a sequence of latent text embeddings, which can be understood by the U-Net. In the Stable Diffusion approach, the text encoder is not trained during the training process. Instead, it uses a pre-trained text encoder called CLIPTextModel, which was trained by CLIP. This allows for the utilization of the strong language understanding capabilities of CLIPTextModel without the need for additional training [24].

4.5 Stable Diffusion

Stable Diffusion is a deep learning model that generates images based on text descriptions. It was developed by the CompVis group at LMU Munich [79]. In addition to generating images from text, Stable Diffusion can also be used for tasks such as inpainting, outpainting, and image-to-image translations guided by a text prompt. Stable Diffusion was released in 2022 by a collaboration of Stability AI, CompVis LMU, and Runway, with support from EleutherAI and LAION. This model was trained on 512x512 images from the LAION-5B database, which is the largest freely available multimodal dataset [79]. The code and model weights for Stable Diffusion are publicly available and can be run on most consumer hardware with a GPU and at least 8 GB of VRAM. This is in contrast to previous text-to-image models such as DALL-E and Midjourney, which were only accessible via cloud services [79].

Stable diffusion was chosen as the text-to-image model to be explored in this research work, since it is the only open source alternative, which not only has monetary advantages, but also has spawned hundreds of other models and innovations worldwide. In the Stable Diffusion model, input images (x) are transformed into a latent space through the diffusion process and then decoded into output images (\tilde{x}). The decoding process is conditioned on text, images, and other information during the training process. This process is illustrated in the component diagram in figure 4.4.

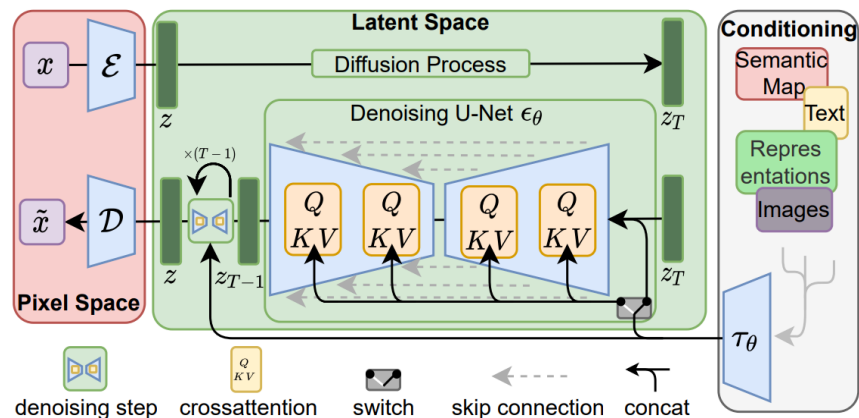


Figure 4.4: Latent diffusion models [69].

4.5.1 Stable Diffusion Versions

In this section, it is presented a comparison of the stable diffusion models, specifically versions 1.4, 1.5, and 2.0.

For starters, the oldest of the three models is the Stable-Diffusion-v1-4 checkpoint [16], which was tuned on 225k steps at resolution 512x512 on "laion-aesthetics v2 5+" and 10% dropping of the text-conditioning to improve classifier-free guiding sampling after being initialized with the weights of the Stable-Diffusion-v1-2 checkpoint [16]. All versions use a latent diffusion model which combines an autoencoder with a diffusion model and involve encoding both images and

text prompts. However, there are some differences in the specifics of the training procedures. For example, version 1.5 [17] was tuned on a larger number of steps at a higher resolution than version 1.4, with 595k steps. Both of these versions use the OpenAI's CLIP, an open-source model that learns how well a caption describes an image. [17]

Stable diffusion 2.0 however makes use of OpenCLIP, an open-source variant of CLIP that was trained using a well-known dataset, a subset of LAION-5B that excludes NSFW images for aesthetic purposes [7]. According to Stability AI, OpenCLIP really outperforms an unreleased version of CLIP on measures and "greatly enhances the quality" of created images. However, version 2.0 has removed many artists and styles due to legal issues, which has resulted in lower quality images in some cases [7]. This model includes many additional features such as the Super-resolution Upscaler Diffusion Models which enhances the resolution of images by a factor of 4, the Depth-to-Image Diffusion Model which infers the depth of an input image, and then generates new images using both the text and depth information and an updated inpainting diffusion model [1]. Since the version 2.0 was only released to the public, in the end of this dissertation, the open source community still had not adapted the various tools which were already created for the other models, compromising the utility of this model for this work

In conclusion, each of the stable diffusion models has its own strengths and weaknesses. But despite all the new functionalities of the version 2.0, none of them are well-suited for the current study, which makes the version 1.5 model the most effective option in this study due to its ability to produce the highest quality images overall and for the tools developed by the open-source community.

4.5.2 Textual Inversion

Text-to-image models provide a unique opportunity to generate images based on natural language descriptions. However, these models have less flexibility when generating images of unique concepts, modifying their appearance, or combining them in new roles and novel scenes. In other words, the question remains of how to use language-guided models to create images from the user input.

In the paper from Gal, Rinon [32], the authors propose a simple approach that allows for this type of creative freedom. By using only 3-5 images of a user-provided concept, such as an object or style, the authors demonstrate how to learn a representation of the concept through new "words" in the embedding space of a frozen text-to-image model. These "words" can be easily composed into natural language sentences to intuitively guide personalized creation. The authors also find evidence that a single word embedding is sufficient for capturing unique and varied concepts [32].

The proposed approach is compared to a range of baselines and is shown to be more effective at accurately portraying the concepts across a variety of applications and tasks. The authors refer to this approach as "Textual Inversion" [32].

4.5.3 Dreambooth

The goal of Dreambooth is the same as that of textual inversion, which is to employ a few images of a subject to identify it uniquely in the output domain of a model so that it may be synthesized. Dreambooth proposes using an uncommon token identifier to describe the subject and to optimize a pre-trained text-to-picture framework that works in two processes, first creating a low-resolution image from text, then using super-resolution diffusion models. A special identifier followed by the subject's class name is included in the input photos and text prompts of the low-resolution text-to-image model (e.g., "A [V] cell") [72].

To avoid overfitting and language drift that could allow the model to associate the class name (e.g., "cell") with a particular instance, an autogenous, class-specific prior preservation loss is suggested. This loss supports the production of varied instances of the same class as the subject by using the semantic prior on the class that is encoded in the model. These are called regularization images [72], and are particularly important in cases where the model is being used for inversion, as it can help to ensure that the resulting images are both realistic and editable. Regularization can also help to prevent overfitting by introducing additional diversity into the training set, which can help to reduce the risk of the model becoming overly reliant on specific characteristics of the training data. While current implementations of techniques such as Dreambooth may still exhibit some drifting, the use of regularization can still be an effective means of improving the overall performance and reliability of the model [72].

The second phase allows the model to maintain high fidelity to the subject's tiny details by fine-tuning the super-resolution component using pairs of low-resolution and high-resolution copies of the input photos. The original paper from Ruiz, Nataniel *et al.* [72] employs the Imagen model as its foundation, although the strategy is not restricted to any one text-to-image diffusion model. Dreambooth does not place any limitations on the input image capture settings, and the context of the subject image can vary. Some factors that can be changed are the location of the subject, the subject's species, color, or shape, as well as the subject's stance, expression, material, and other semantic adjustments. The Dreambooth training procedure is shown in figure 4.5

Dreambooth fine-tunes the entire text-to-image model so that it learns to link a unique identifier with a particular concept, in contrast to textual inversion methods that just train the embedding without altering the basic model (object or style). As a result, compared to textual inversion, the resulting visuals are better tailored to the object or style. In figure 4.6, it is shown a comparison between the two models. Due to its superior results in image synthesis when compared to Textual Inversion, Dreambooth was the chosen technology to generate the synthetic single cells.

Model Selection

The open-source nature of Dreambooth has led to the development of various versions with different features. One such version, created by Joe Penna [41], is specifically designed for digital artists to train their own likenesses, characters, and styles into a Stable Diffusion model. This version is possible to run with a GPUs with 24GB of VRAM, although it should be the only program

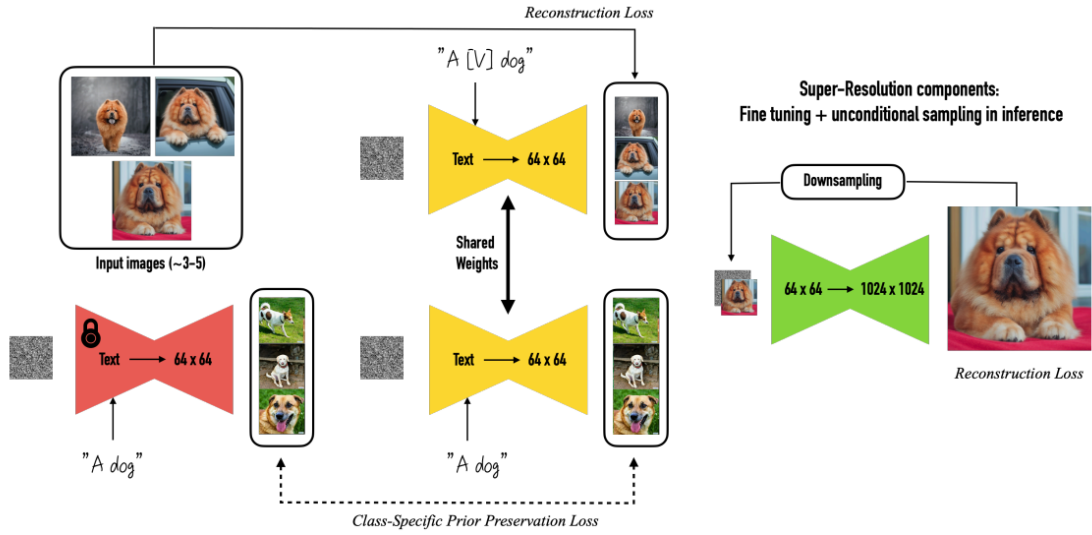


Figure 4.5: Dreambooth training procedure [72].

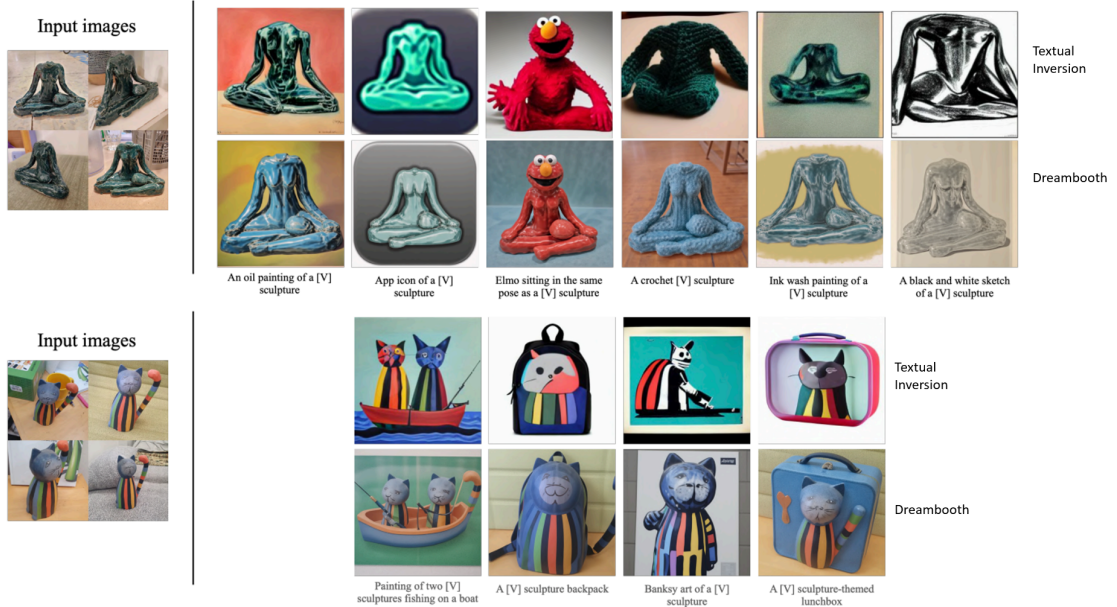


Figure 4.6: Comparison between Textual Inversion and Dreambooth [72].

running and training might be very slow. However, this implementation does not fully implement Google's ideas on preserving the latent space, which can result in shifted images that are similar to the training data. In order to avoid this, it may be necessary to train for fewer steps, use better training images, or provide additional prompting. The prompt should consist of the token the user trained, followed by the class which it belongs, such as "ASCUS cell".

In the XavierXiau approach [87], the default learning rate is set to $1.0e-6$, as the $1.0e-5$ suggested in the Dreambooth paper leads to poor editability. The parameter `reg_weight` corresponds to the weight of regularization in the Dreambooth paper, and the default is set to 1.0. This model also recommends using values of 100 or 200 regularization images to better align with the original paper. This version of Dreambooth requires a placeholder word [V], called an identifier, similar to the rare word used in the T5-XXL tokenizer in the original paper [72]. The author uses a random word, "sks", and hard codes it. The default training is 800 steps, but the model typically performs well enough with 500 steps. Using two A6000 GPUs, this model takes approximately 15 minutes to generate excellent images in 400 steps [87].

The ShivamShirao model [74] is able to generate high quality images in as little as 400 steps and the author suggests using 30 training images for best results. This model has the capability of training multiple tokens into it, thus it allow us in our research work to train the different cell classes in a single model, a characteristic which no other model had. It is also possible to run locally, in contrast with the TheLastBen repository. To train the model the user has to choose 30 images of each class, and rename the image to a token which should not exist in the original stable diffusion model [74]. If the user is not satisfied with the outputs of a model because of the low amount of steps, it is also possible to resume the training at any time. The recommended total number of steps is calculated as the number of instance images multiplied by 100, so for 30 images, the total number of steps should be 3000. For our 4 classes, each with 30 images, the total number of steps should be 12000. This model uses mixed precision fp16, a learning rate of $1e-6$, and a default of 800 steps. It also includes a concept list json where the names of tokens, classes, and regularization can be controlled [74].

4.5.4 Inpainting

Image inpainting refers is the task of filling-in missing data in a designated region of the visual input [9]. It can be used for a variety of purposes, such as restoration of damaged artwork, removal of blemishes from photographs, and completion of occluded regions in videos [92]. This can be done using a mask, represented by a binary image, to specify the region of the image that requires replacement with fresh content, which allows for greater precision and can enhance the quality of the results. There are several approaches to inpainting, including exemplar-based methods, which copy information from similar regions in the image, and texture synthesis methods, which generate new pixels based on the patterns in the surrounding image [92].

The topic has been explored since the pre-deep learning era [10] [20], and recent advances have been made thanks to the use of adversarial learning [60] [37] and deep and wide neural networks [48]. These algorithms are usually trained on large automatically generated datasets,

produced by randomly masking real images, and it is common to use complex two-stage models with intermediate predictions, such as segmentation maps [78], edges [56] and smoothed pictures [49].

In the work of Suvorov, Roman, et al [80] it was achieved state-of-the-art results with a straightforward single-stage network. This method takes into account that the large structure of natural images must be "understood" in order to solve the problem of realistically filling in missing areas of images, as well as image synthesis. Additionally, with a large mask, even a broad but constrained receptive field may not be sufficient to access the data required to produce an accurate inpainting [80]. Typical convolutional designs could not have an effective receptive field that is big enough, so to solve the issue and maximize the effectiveness of the one-stage solution, each system component was intervened. They suggested an inpainting network based on fast Fourier convolutions (FFCs), which enables a receptive field that completely encloses the picture even in the early layers of the network [80]. This method enhanced not only the network's parameter efficiency but also perceptual quality. It was also used perceptual loss based on a semantic segmentation network with a high receptive field. Finally, in order to force the network to fully leverage the model's and loss functions high receptive field, they adopted an aggressive training mask creation technique, which generated brand and huge masks. They called this model Large mask inpainting (LaMa), which will be used in this dissertation to remove cells from the optic disk [80]. The summarized architecture can be seen in figure 4.7

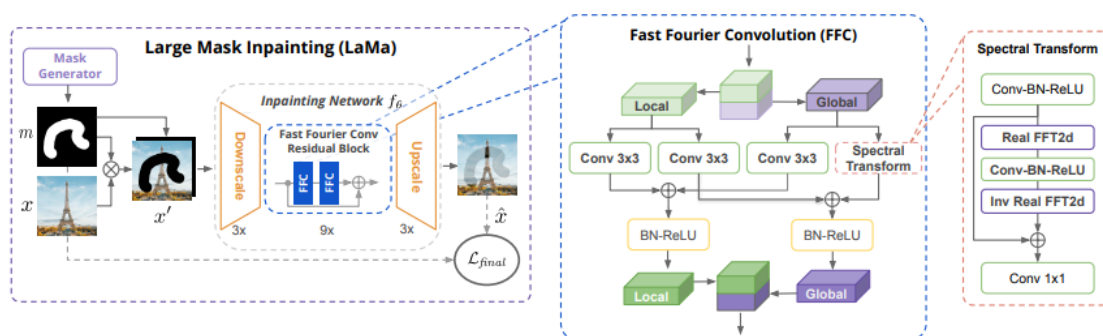


Figure 4.7: The scheme of the proposed method for large-mask inpainting (LaMa) [80].

The work of Rombach *et al.* [69], tested the applicability of latent diffusion models to this area, having state-of-the-art results. In this work, they compared their general method with more specialized approaches, which were more specific, still having better ones. Their evaluation uses an architecture inspired by the LaMa protocol, and trained a diffusion model in the latent space of the first VQregularized stage without attention [69]. The U-Net of this diffusion model includes 387M parameters, the BigGan residual block for up- and downsampling, and attention layers on three levels of its feature hierarchy. Training it at a resolution of 512x512, established a new state-of-the-art Fréchet Inception Distance (FID) for picture inpainting [69]. This inpainting model will be used in this dissertation to inpaint singular cells, in multiple cell images.

Chapter 5

Methodology

The primary objective of this research is to generate synthetic images of cytological patches in order to improve the performance of the detection algorithms developed by Fraunhofer AICOS. More specifically, the focus is to expand the overall volume of training data, increase the number of instances of underrepresented classes, and balance the data volume between the different classes.

The validity of the generated images will be evaluated by two cytopathologists based on realism and accuracy. These images will then be used as training data for previously developed cervical screening detection models, namely a Region-based [73] and a Nuclei-based [53] model, to evaluate potential improvement in performance.

The first step in the process is to generate a fine-tuned stable-diffusion model, which can generate single cells using the Dreambooth tool presented in the section 4.5.3. Once the model is successfully generating single cells for all abnormality classes, the model will be used to inpaint them into multi-cellular patches. In order to validate the generated images, the detection models will be trained using them to evaluate the improvements they provided. Additionally, cytopathologists will be asked to review the generated images for their degree of realism and abnormality class. The complete pipeline of this work can be seen in Figure 5.1.

Overall, the goal of this methodology is to utilize generative modeling and other AI-based image generation techniques to enhance cytological imaging collections, with the ultimate aim of improving cervical cancer screening. The results of this research will be used to improve the already developed decision support system developed in the TAMI project [31] and will be examined in terms of realism and relevance by medical experts for each of the respective classes.

5.1 Hardware

While central processing units (CPUs) can execute many common jobs quickly and sequentially, graphics processing units (GPUs) use parallel computing to divide massively complex problems into numerous smaller simultaneous calculations. This makes GPUs excellent at managing the massively distributed computational operations needed for diffusion models [50].

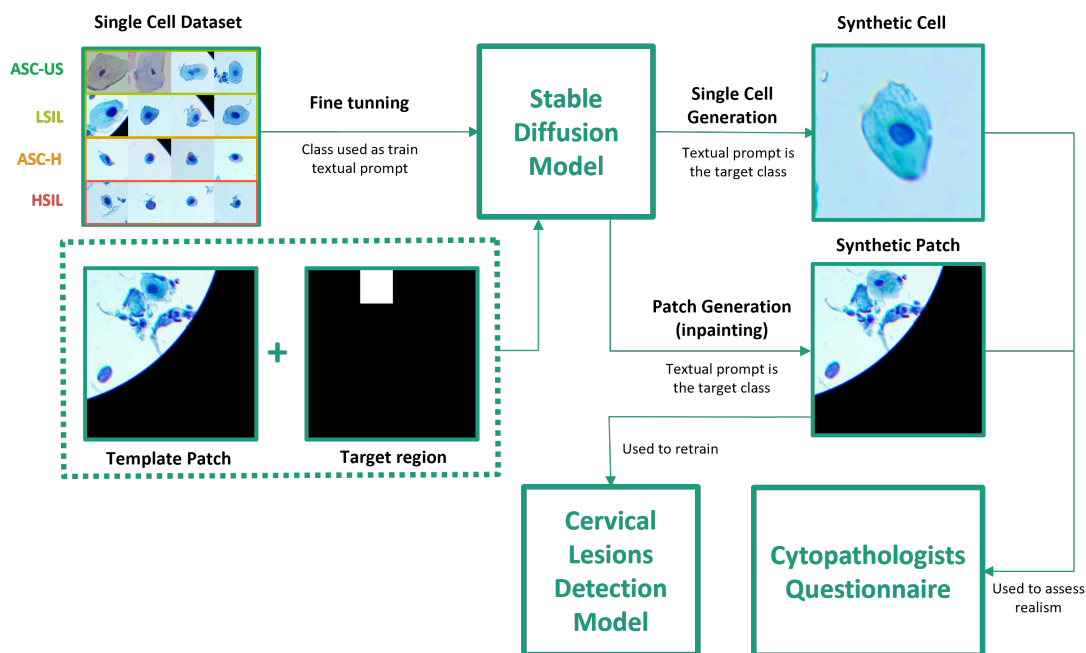


Figure 5.1: Overall pipeline of the proposed methodology.

The use of advanced frameworks and technologies, as well as specialized software, was facilitated by Fraunhofer AICOS [30] and played a crucial role in enabling the training and inference of our models. The training and inferring was conducted using Fraunhofer’s High Power computing tool, which utilizes multiple machines with different resources, as shown in Table 5.1.

Name	CPU	Cores/Threads	Momory	GPU
hpc01	2x Xeon 4114	20/40	128G	1xV100 16G
hpc02	2x Xeon 4114	20/40	64G	1x V100 16G
hpc03	2x Xeon 4214	24/48	128G	1x V100 32G
hpc04	2x EPYC 7502	64/128	128G	1x V100S 32G

Table 5.1: High power computing hardware provided by Fraunhofer.

5.2 Fine tuning stable diffusion model

This section aims to explain the methodology of creating a fine-tuned version of the stable diffusion model for cervical cancer screening by utilizing the Dreambooth technology. As presented in Section 4.5.3, Dreambooth is able to learn about a subject by using a few images and then synthesize realistic generated images associated with an uncommon token identifier.

In the first sub-sections, it is presented the model selection approach and the process of preparing the dataset. The dataset was modified by merging and removing certain classes to improve the

balance of images and adapted for the detection models. Additionally, some regularization techniques were also employed to address issues of overfitting and class preservation. Then, a diverse range of images was carefully selected for training, including cells with a wide range of characteristics such as color and shape. The optimal number of images for training with Dreambooth was determined to be between 20 and 30 through experimentation. This allows to create a suitable dataset for training the model and improve the performance and realism of the synthetic images.

In the hyperparameter tuning sub-section, a series of experiments were conducted to determine the optimal settings for the Dreambooth model. The goal of this process was to find the combination of hyperparameters that would produce the best results in terms of the performance of the generated synthetic images.

Finally, the last sub-section presents how the performance of the generated synthetic images was validated, with the ultimate goal of selecting the best model for the inpainting task.

5.2.1 Model selection

In this section, various Dreambooth versions were evaluated for their performance in fine-tuning a stable diffusion model. To ensure a fair comparison, the default settings provided in the respective Dreambooth repositories were used for each model, as they were determined through experimentation by the original authors to be optimal for the given implementation.

The first comparison was between the JoePenna [41] and the XavierXiao [87] model. Both models were trained with their default settings, however, the XavierXiao model had a default number of 800 steps, while the JoePenna model had a default number of 2000 steps. This difference in the number of steps may lead to variations in the resulting images. To further investigate the performance of the JoePenna implementation, the single cell nuclei dataset was trained using this model with different numbers of steps: 2000, 5000, and 10000. This way, it is possible to gain insight into the model's performance under varying conditions, in contrast to just using the default parameters provided in the respective repository.

The ShivamShrirao [74] model was also evaluated. Unlike the other models, it has the advantage of being able to train multiple classes of abnormalities simultaneously. The first test was trained using the default settings provided in the repository, with a total of 10000 steps, due to the increased number of training images as explained in section 5.2.3. All of the models were trained using 30 images per class, except for the carcinoma class, which was trained with only 13 images due to limited availability. This is below the recommended value of 20-30 images per class.

In addition to the Dreambooth model, it was also necessary to choose which Stable Diffusion version to fine-tune, given that all the Dreambooth implementations supported both model Stable-Diffusion-v1-4 [16] and Stable-Diffusion-v1-5 [17]. The ShivamShrirao model was trained twice using the exact same parameters, with the only difference being the Stable Diffusion version used for fine-tuning.

5.2.2 Dataset preparation

In the process of preparing the dataset for this study, several modifications were made. First, the HSIL and SCC classes were joined together in order to create a larger dataset of images with abnormal cells, due to the big lack of carcinoma instances in the single cell dataset. This was done in order to contrary the big unbalancing of carcinoma images when compared to the other classes, and especially because both Region-based and Nuclei-based cervical lesions detection models [53, 73] used in this work also joined this two classes. The classes of "glandular cell", "pavimentosacell", and "transformation zone" were also removed from this dataset since they were not directly relevant to the aims of this study. The dataset distribution after the changes is presented in Table 5.2

Cell Class	#Instances
ASC-US	352
ASC-H	79
LSIL	58
HSIL+SCC	216

Table 5.2: Region-based Dataset distribution after data preparation procedures.

For the single cell dataset nuclei instances, the nuclei were extracted from the original patches and divided according to the class of abnormality. The classes used in this dataset were the same as those used in the previous dataset, which included the merged HSIL and SCC classes. It was chosen a padding of 0% in order to be more suitable for the further inpainting step.

To ensure compatibility with Dreambooth, the images were resized from their original size of 256x256 in the single cell regions dataset, and from various sizes in the nuclei one, to 512x512 using the Bulk Image Resizer (BIRME) [11] online tool. This was necessary as Dreambooth requires input images to be of this specific size, and any images that do not meet this requirement are automatically cropped, potentially losing important features of the cells [72].

5.2.3 Model training

Regarding the training of the Dreambooth model, a series of experiments were conducted to evaluate the parameters which influenced its performance. One of the most impactful parameters in Dreambooth is the ratio between the learning rate and the number of training images [25]. As previously discussed in section 4.5.3, the total number of steps should be calculated by the number of training images multiplied by 100. The region and nuclei instances of the single cell dataset consist of four different classes (ASC-H, ASC-US, LSIL, HSIL+SCC). For these classes, a learning rate of 1.0e-6 and a total of 10000 steps, in comparison to the 12000 steps, produced the most satisfactory results. These findings align with the recommendations of Hugging Face [25], which advises

that the Dreambooth model tends to overfit quickly, making it important to gradually increase the number of training steps for optimal performance.

In this work, the token for the text-to-image model includes a suffix consisting of an '@' symbol appended with the class of the cell in lowercase and without hyphens, followed by the word "cell". Since the HSIL and the SCC class were joined together, its textual prompt includes the two classes. The following Table 5.3 illustrates the correspondence between each class and the corresponding prompt:

Cell Class	Text Prompt
ASC-H	"@asch cell"
ASC-US	"@ascus cell"
LSIL	"@lsil cell"
HSIL	"@hsilScC cell"

Table 5.3: Cell class and corresponding textual prompt for the Text to Image model.

5.2.3.1 Regularization images

In order to address the issues of overfitting and class preservation, regularization can be used during the training of a model [72]. By introducing a "class" of reference images through regularization, the model can be prevented from drifting into unrelated classes during training. As a relatively recent technology, the optimal selection of regularization images for use in training models is still an area of active research and experimentation [72]. In this work, it is conducted a series of tests to determine the impact of different types and quantities of regularization images on the performance of the model.

In the first approach, all of the cells from the single cell dataset regions instances were used as regularization images. The aim of this experiment was to utilize the maximum number of available images representing the class that was going to be trained. In the second approach, all of the classes from the dataset were also used except for the one being trained, with the objective of not including the training class in the regularization images. For example, using all cell classes except for ASC-US when training the model on that class. In the third approach, it was only used the classes that were most closely related to the one being trained, such as using SCC and ASC-H as regularization images when training on the HSIL class. Finally, it was used regularization images that were completely unrelated to the dataset, following the suggestion of JoePenna, that using the class "dog" as regularization images had produced the best results when training on himself [41]. The chosen class for this last approach, was "person", being the most common one in Dreambooth models. The objective of this test was to examine the use of regularization images that were entirely unrelated to the dataset.

5.2.3.2 Training images

In order to determine the optimal number of images for training with the Dreambooth technology, a series of experiments were conducted in which the number of images used for training was modified. In each experiment, the performance of the generated synthetic images was evaluated using visual inspection, allowing to identify the range of image instances that produced the best results. Although the original paper where Dreambooth was introduced affirmed the model had good results when trained with 3 to 5 images [72], it did not mention the ideal number. Relying only on experimentation, the consensual optimal number of images for training with Dreambooth is between 20 and 30 images, according to the most popular Dreambooth repository owners such as Joepenna [41] and XavierXau [87].

The first two experiments utilized regularization images comprising all cells except for the class under examination. In the first one, all the cells from the class under training were used as training images, while in the second one, only a subset of 30 images was picked. The 30 images in this latter experiment were selected randomly to accurately assess the impact of image quantity on the results. To further evaluate the effectiveness of training images on model performance, three additional experiments were conducted utilizing people as regularization images. The third experiment replicated the approach from the previous test by utilizing all images from the dataset. In the fourth and fifth experiments, a subset of 30 images was hand-selected. The criteria used in the fourth experiment was to choose cells that were completely isolated from other cells, discarding images that contained regions. Figure 5.2 illustrates some examples of the division made in the fourth and fifth experiments, using these criteria.

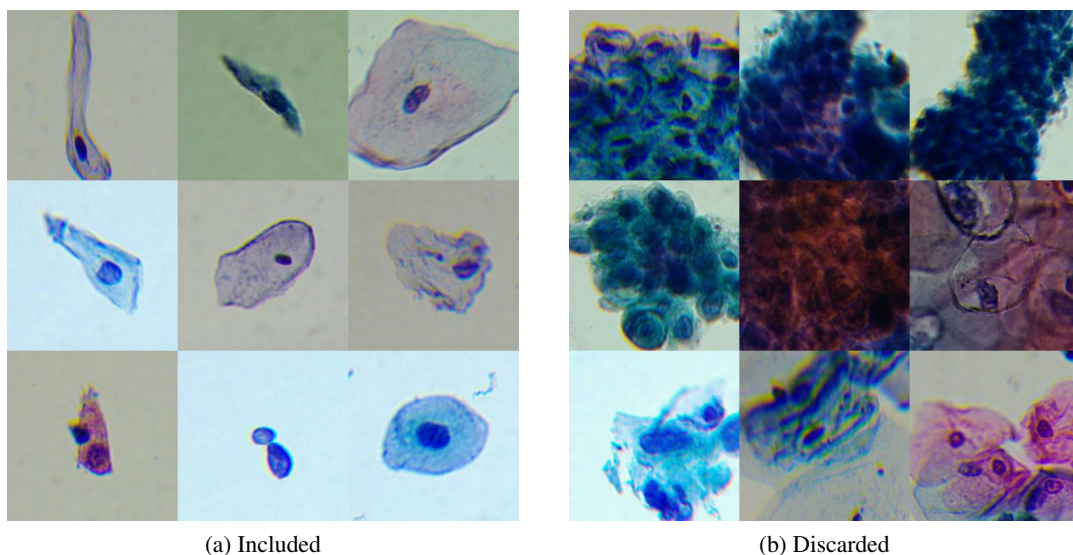


Figure 5.2: Comparison between images which were (a) included and (b) discarded from the training set on the experiments with 30 hand-selected images.

In the fifth experiment, to evaluate the model's ability to learn regions, the opposite approach was taken, selecting 30 images that contained regions in contrast to single cells. Regarding the

experiments where only a subset of training images was selected, it was important to include cells with a wide range of characteristics such as color and shape. By training the generation model on a diverse set of images, the realism and relevance of the synthetic images is expected to increase.

5.2.4 Validation

In order to determine the optimal parameters for training the model, a series of experiments was conducted to evaluate the performance of different configurations. The ultimate goal was to select the model that would produce the best results for image inpainting. However, as a preliminary step, a text-to-image model was used to generate individual cells and assess their quality rather than attempting to inpaint them into a multicellular image. This allowed to isolate the performance of the model itself and identify any issues with the generated cells, rather than attributing any problems to the inpainting process.

During the evaluation of the model's outputs, two main factors were considered: the realism of the generated images and the ability of the model to accurately distinguish and reproduce the specific characteristics of the different classes of cancer cells. The realism of the images was an important consideration, as it is crucial that the model is able to generate visually plausible images that are similar to real cells. At the same time, the ability to accurately differentiate between the various classes of cancer cells and accurately reproduce their specific characteristics is critical for the model to be useful in a clinical setting. Additionally, using the text-to-image method was a more efficient method for generating the cells for evaluation.

For each experiment, 40 images of each class of cell were generated in order to obtain a comprehensive understanding of the quality of the generated instances. This enabled to thoroughly assess the performance of the model and make informed decisions about the appropriate parameters for training. Overall, the results of these experiments provided valuable insights into the characteristics of the model and allowed to optimize its performance for image inpainting tasks.

To further validate the results of the experiments, a questionnaire containing 100 cell images was created, half of which were real and half of which were generated by the model. This questionnaire was then given to two cytopathologists of the Portuguese Oncology Institute of Porto, who were asked to differentiate the real images from the generated ones. The initial objective was to additionally ask for the abnormality class of each cell; however, despite the specialists' ability to recognize many of the characteristics of each class of cell in the generated images, they noted that it would be more accurate to make a precise evaluation of the cell class if a multicellular image was provided, as the assessment of cell classes often relies on comparing cells within the same image. Given that the evaluation of individual cells was already a challenge for the specialists, it was decided not to create a validation questionnaire for the nuclei images. This was due to the fact that nuclei are often very homogeneous and of low quality when examined individually, making it difficult to accurately evaluate them. This feedback highlighted the importance of considering the context in which cells are evaluated and reinforced the need for the model to be able to generate high-quality multicellular images for clinical use. Screenshots of the single cell questionnaire can be seen in section [B.1](#) in the Annex.

5.3 Multiple Cell inpainting

In this section, the methodology for inpainting multiple cell cytological images is presented. The objective is to use inpainting to modify single cells from multicellular patches in order to change the abnormality class of the target cell.

In the following sub-sections, the dataset preparation process is described, including the isolation of patches with abnormal cells, the creation of masks, and the consideration of cell size and characteristics. The hyperparameter tuning of the inpainting process is also discussed, including the number of steps, denoising strength, CFG scale, and full resolution.

Finally, the validation process of the method is explained, referring to the different forms that the model is going to be evaluated. Firstly, the images will be examined by two specialists, who will evaluate the realism of the patches and identify the class of abnormality of the generated cells. The second validation method consists of incorporating synthetic images into the original Region-based and Nuclei-based datasets, in order to evaluate the performance of the corresponding detection models when trained with the generated images.

For each detection model, there will be four tests to evaluate its performance:

1. Increasing the dataset volume using fixed-size synthetic image generation.
2. Increasing the dataset volume while controlling the size of the generated synthetic images.
3. Comparing the impact of data augmentation using fixed-size synthetic image generation versus basic image manipulation.
4. Comparing the impact of data augmentation using synthetic image generation versus basic image manipulation, while controlling the size of the generated images.

The methodology and results of each test will be thoroughly discussed in the subsection entitled [AI-based Cervical Lesions Detection Algorithms](#).

5.3.1 Dataset preparation

The dataset preparation process for the inpainting was a crucial step in the research, as the quality and accuracy of the data used can significantly impact the results of the study.

The first step in the process was to filter the most suitable training patches for the Region-based and the Nuclei-based dataset. For the Region-based dataset, it was necessary to identify and select patches that contained any squamous cell abnormality classes. This was done using the information provided in the dataset metadata, which detailed the location and characteristics of the abnormal cells. Regarding the Nuclei-based model, two different approaches were followed. The dataset used for the first and third experiments (mentioned in section 5.3) were composed of images that had any squamous cell abnormality classes, just like the Region-based model. However, for tests 2 and 4, in which the size of the generated images was controlled, the dataset had to be constituted of patches that had squamous nuclei, cells without any abnormality class. This is due

to the fact that all abnormality classes have average nuclei sizes bigger than the normal squamous nuclei, which avoids inpainting incongruities (i.e. the target bounding boxes with adapted size for inpainting abnormal cells will always be larger than the original template normal bounding box).

In order to ensure the integrity of the training process for the detection models, only patches that belonged to the training sets of both Region-based and Nuclei-based datasets were considered. This was necessary to avoid any potential biases in the evaluation of the model's performance. In the Region-based dataset, the training instances made up approximately 73% of the original dataset. This distribution is consistent with the standard train/test split of 80% for training and 20% for testing, which was used in the development of the Region-based detection model [73].

Once the patches with abnormal cells were identified, a mask was created for each patch. The mask consists of a white square that covers the portion of the image with abnormal cells and was created to ensure that the region of interest (i.e., the abnormal cells) was clearly defined and could be easily analyzed in our study. This process was necessary to accurately isolate and study the abnormal cells. In Figure 5.3 is presented on the left, the patch that is going to be inpainted with a synthetic abnormal cell of a specific target class, and on the right, the corresponding mask indicating that two cells in that specific position will be inpainted.

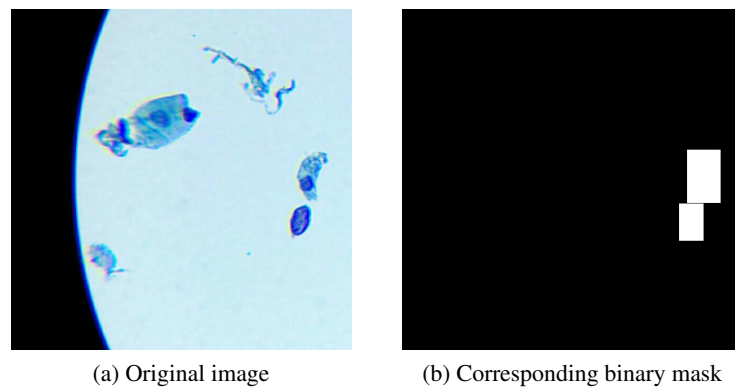
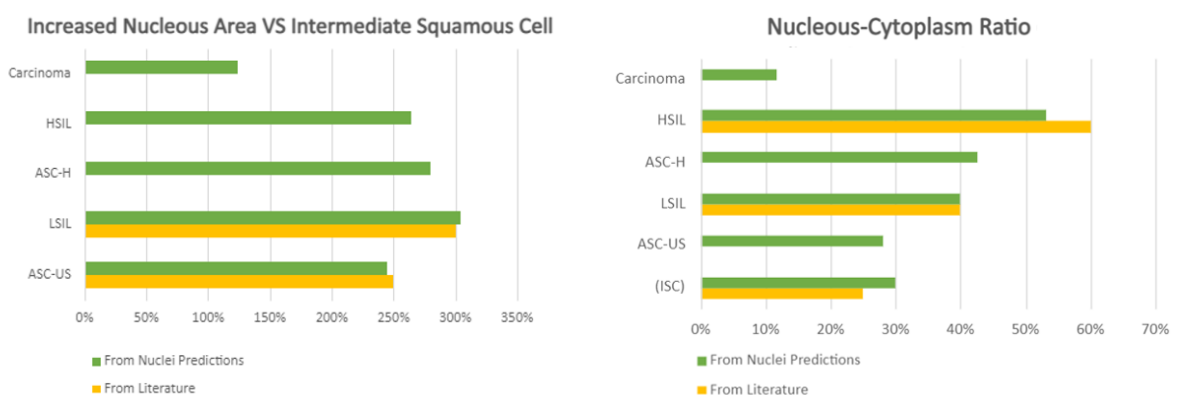


Figure 5.3: Example of a patch (a) and corresponding inpainting mask (b).

For the test in which the size of the mask was varied, it was necessary to consider the class of the original cell and the target abnormality class, to change the size according to the domain knowledge (e.g. average area of the bounding box for the target class). However, an issue arose when transforming larger cells into smaller cells. If the abnormal cells were located in an area of the image with aggregates of cells, the transformation would cause obvious incongruities in the limits of the generated cell, and consequently would be easy to distinguish a real patch from a generated one. To address this issue, patches with relatively isolated abnormal cells were preselected (i.e. annotations within cell aggregates were discarded) to avoid such incongruities. From the 613 patches that existed in the original training dataset, only 435 were preselected, with a total of 29% of the patches discarded.

Squamous lesion cells can vary significantly in size and characteristics, and there are currently no concrete values defined in the literature on the size for each abnormality class or how they

compare from one class to another [5]. While there exists information on the size of some classes of abnormal squamous cells, it was not possible to determine from the literature the average size for all classes. To address this issue, the average size of each class of abnormal squamous cells was calculated by making an average of the isolated cells dataset and the Nuclei-based dataset. While this method is not perfect, the values obtained through this method showed a strong resemblance to the values defined in the literature. This can be seen in Figure 5.4a, which depicts the average nuclei size increase for each class, when compared with the average nuclei size of normal intermediate squamous cells, with a maximum of a 5% deviation for the LSIL class. The Nucleous-Cytoplasm Ratio also presented a strong resemblance between the obtained values and the literature, having a maximum of 7% difference for HSIL 5.4b.



(a) Per class average nuclei size increase versus average nuclei of normal intermediate squamous nuclei, for both the Nuclei-based dataset and the literature.

(b) Per class average nucleous-cytoplasm ratio, for both the Nuclei-based dataset and the literature.

Figure 5.4: Comparison between Nuclei-based dataset average values and the literature.

Table 5.4 presents the relative average size comparison between different cervical lesions classes using the isolated regions dataset. Each value in the Table represents the percentage of the size of the cell class in the row compared to the size of the cell class in the column. For example, the value in the row "ASC-US" and column "ASC-H" is 45.5, which means that when transforming from ASC-US to ASC-H, the ASC-H cells will be approximately 45.5% the size of the original ASC-US cells.

Table 5.5 presents the relative average size comparison of different cervical cell nuclei lesions to an average intermediate squamous cell nucleus. From this Table it is possible to determine that LSIL nuclei are on average 305% bigger than intermediate squamous cell nuclei.

In order to compare the impact of changing the data augmentation strategy explored in the original works [53, 73] based on basic image manipulations versus the usage of the synthetic image generation approach proposed in this work, it is essential to establish a data distribution that is comparable to the one utilized in previous experiments. This will allow to gauge the extent to which the generated images improve model performance under the same conditions as the conventional data augmentations (Conventional DA). Specifically, for the Region-based detection

Class	ASC-US	ASC-H	LSIL	HSIL	SCC
ASC-US	100	45	64	36	54
ASC-H	220	100	140	80	120
LSIL	157	71	100	57	86
HSIL	275	125	175	100	150
SCC	183	83	117	67	100

Table 5.4: Relative average size comparison between different cervical lesions classes (in percentage).

Class	ASC-US	ASC-H	LSIL	HSIL
Relative size(%)	245	280	305	265

Table 5.5: Relative average size of cervical cell nucleus lesions compared to an Intermediate Squamous nucleus (in percentage).

model, the number of patches produced through data augmentation is outlined in Table 5.6.

Class	#Data augmentation patches
ASC-US	134
ASC-H	357
LSIL	391
HSIL	341

Table 5.6: Number of patches created for data augmentation for the Region-based detection model in the original paper [73].

The patches selected for the test in which the size of the generated cell is not controlled were chosen randomly from the training set. However, in the test in which the cell size is controlled, the patches were drawn from the isolated cells dataset.

Similarly, for the Nuclei-based detection model, the same process was followed. However, instead of dividing the dataset by the number of patches, it was split by the number of annotations per class. The distribution is seen in Table 5.7.

Class	#Data augmentation annotations
ASC-US	79
ASC-H	325
LSIL	263
HSIL	226

Table 5.7: Number of annotations created for data augmentation for the Nuclei-based detection model in the original work[53].

In addition to the images themselves, the detection models also take as input a *JSON* file that contains various information about the images, such as their id, file path, and the location of any abnormal cells present within the image. For each test, it was necessary to make modifications to this *JSON* file, such as adding information about the newly generated images or removing the data augmentation instances. For the resized images, it was also necessary to recalculate the coordinates of the cells. This was achieved by determining the new width and height, and adjusting the x and y values accordingly. In instances where the mask would extend beyond the boundaries of the image, the x minimum was calculated by subtracting the new width from the maximum image size.

In the case where the masks were resized, if the class to transform to was smaller than the original one, the region between the bigger square and the outside of the smaller square is not inpainted, staying equal to the original patch. To solve this problem, there will be an additional approach in which the initial cell is removed using the Lamma cleaner previously presented in the 4.5.4 section. Two resized masks of different classes can be seen in Figure 5.5.

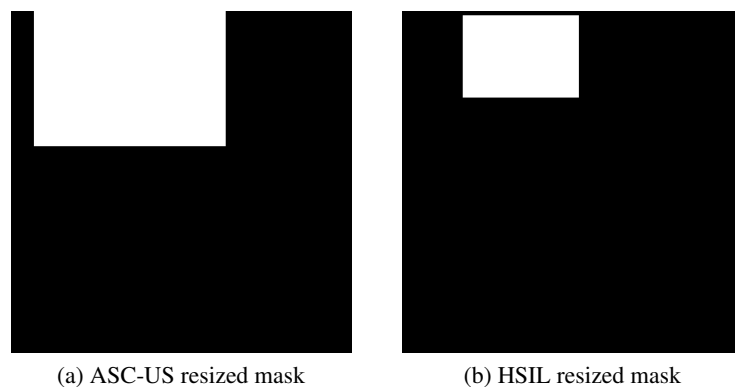


Figure 5.5: Comparison between two resized masks of different classes, for the same patch.

5.3.2 Hyperparameter tuning

Hyperparameter tuning is a crucial step in machine learning, as it allows for fine-tuning the performance of a model. When it comes to inpainting images, several key parameters must be considered, the most impactful being: the number of steps, the denoising strength, the classifier-free guidance scale (CFG Scale), and the full resolution [6].

Full resolution is especially important when inpainting small areas with fine details, which is often the case in cytological images. When inpainting at full resolution, several steps are taken. First, the smallest rectangular area that includes all pixels of the mask is determined, and then it is expanded in each direction by a specified number of pixels [6]. Its expansion continues until the rectangular area reaches its maximum size, determined by the values set for width and height. Once the image is scaled, it is then processed by the Stable Diffusion model, scaled back down, and inserted back into the original image. The padding helps the model gather information from

outside the image, which helps it understand what should be changed within the image. This parameter can be set to either True or False [6].

Denoising strength controls how much the image will change compared to the original. When set to 0, there will be no changes to the image, but when set to 1, the inpainted image will be completely different from the original [6]. In the case of cytological images, the denoising strength should generally have high values, due to the high level of detail present in the cells and the small distinctions between various classes of cells. This requires a high level of precision, which is defined by this parameter. However, if the value is set too high, the inpainted image will not have the same colors as the rest of the patch, making it easy to identify the limits of the inpainted rectangle. The default value for the denoising strength is 0.2, but it should be increased to higher if the image is too similar to the original. Due to the limitations of the hardware, the tests were based on the output of the previous test. The first experiment was done with 0.5, followed by 0.75 and 0.9.

The CFG scale defines how closely the inpainting should follow the prompt. This parameter can vary between different values, with lower numbers (0–3) giving the inpainting more freedom, while higher values (7–10+) instruct the model to closely follow the prompt with minimal deviation [6]. While it is important for the cells to closely follow the prompt, the inpainting function should also take into consideration the cell on which it is inpainting, in order to better transform it and interpolate the values. The experimental values for this parameter were 4, 7, 10 and 20.

Finally, for the number of steps in the inference, the training values were 20, 50, and 100, based on previous experiments with the Text-to-Image model.

5.3.3 Validation

5.3.3.1 AI-based Cervical Lesions Detection Algorithms

The proposed approach for assessing the potential of the synthetically generated inpainted images to improve models' performance involves incorporating them into the original training set for both the Region-based and Nuclei-based detection models. Four tests will be conducted to evaluate the performance of the models with the synthetic images:

1. Increasing the dataset volume using fixed-size synthetic image generation.
2. Increasing the dataset volume while controlling the size of the generated synthetic images.
3. Comparing the impact of data augmentation using fixed-size synthetic image generation versus basic image manipulation.
4. Comparing the impact of data augmentation using synthetic image generation versus basic image manipulation, while controlling the size of the generated images.

The purpose of generating synthetic cells is to address the limitation of the small initial dataset, which is believed to have contributed to the poor performance of the detection models. This is

evident in the fact that the best classification performances were observed in the more highly represented classes, specifically ASC-US [73]. In order to assess the improvement in performance achieved through the use of generated images, four corresponding inpainting images were created for each original image. In each of these inpainting images, all the cells within the original patch were manipulated to simulate a specific class of abnormality, resulting in a dataset that was five times larger.

Although there should exist more tests, to verify what would be the perfect relation between the number of generated images, and original images which would give the best performance, this work was restrained by both time and hardware resources, being possible to only realize one of these tests. The model was trained with both the dataset and hyperparameters which yielded the best result, including the images generated with basic image manipulation, so it would be possible to achieve the best performance of the model possible. Since the Nuclei-based approach demonstrated worse results when trained with basic image manipulations, the baseline for the first test in this approach will be without data augmentations.

The second test in the study builds upon the methodology of the first test, with some important distinctions. One key difference is that the second test involves controlling the size of the generated images based on the class which is being generated. Although this method is expected to improve performance, it may also have introduced unnecessary complexity, as the model effectively adjusts the size of the images by itself without the need for this step. For example, here is an image demonstrating two generated cells, one from ASC-US and other from HSIL. From image 5.6, it is possible to verify that the generated HSIL cell is smaller than the ASC-US one, without the resized mask.

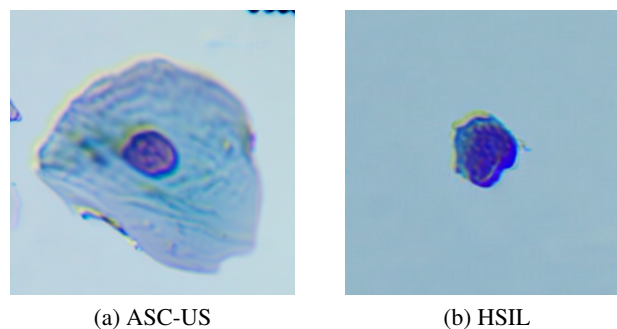


Figure 5.6: Comparison between two generated cells. The ASC-US image (a) is clearly bigger than the HSIL image (b).

To account for these nuances, both the first and second tests were conducted in parallel. Another key difference between the two tests is the use of a different training dataset for the second test. Specifically, the second test was trained using a dataset of isolated images, which comprises only 70% of the original dataset used in the first test, as stated in section 5.3.1. As a result, it is not possible to make direct comparisons between the results of the two tests, as the difference in the amount of training instances may affect the model's performance.

In the third test, the performance of synthetic image generation is compared to that of Conventional Data Augmentation (CDA). Basic image manipulations are a widely used technique to increase the size and diversity of the training datasets and have been shown to improve the performance of various machine learning models. However, image generation is a more advanced form of data augmentation that involves creating new images from an existing dataset. The main benefit of image generation is that it can produce highly diverse and realistic images that are difficult to obtain through conventional data augmentation methods. This test has the objective of specifically evaluating the performance of the generated images with that of CDA and its ability to generalize to unseen data.

The fourth experiment maintains a similar design as the previous test, but, similar to the second experiment, it manipulates the dimensions of the generated image based on the classification of the abnormality. By using the same dataset as the third experiment, we can effectively compare the performance of both approaches and determine the impact of controlling the size of the generated image on the overall accuracy of the model.

All of the tests for both approaches will be evaluated using the mean Average Precision (mAP) at 50% Intersection over Union (IOU) for all the classes. For the Region-based approach, it will also be examined the mean average recall across all images with at most 10 detections (AR10) and for the Nuclei-based approach, it will be used the AR100. This difference is due to the fact the Nuclei-based dataset might have more than 10 detections per patch, making a limit of 10 too restrictive.

Table 5.8 presents the number of patches that are going to be used for each test. TR0 corresponds to the baseline results of the original model trained without generated images [73], while TR1, TR2, TR3, TR4 correspond to the test 1, 2, 3 and 4 respectively.

Test	TR0 [73]	TR1	TR2	TR3	TR4
Number of Patches	686	3430	2426	1931	1931

Table 5.8: Total number of patches for the different Region-based tests.

Table 5.9, presents the number of patches for each test for the Nuclei-based approach. TN0 represents the baseline test used in the Nuclei-based detection model [53], without the synthetic images. TN1, TN2, TN3, TN4 represent the test 1, 2, 3 and 4 respectively.

Test	TN0 [53]	TN1	TN2	TN3	TN4
Number of Patches	1763	3875	2691	2131	2128

Table 5.9: Total number of training patches for the different tests of the Nuclei-based approach.

5.3.3.2 Cytopathologist Questionnaire

Similar to the single cell questionnaire described in section 5.2.4, a multiple cell questionnaire was created to evaluate the realism of the generated images by two experienced cytopathologists at the Portuguese Oncology Institute of Porto. This questionnaire consisted of 100 images, half of which were real and half of which were generated. While the other questionnaire evaluated single cells, this form pretends to evaluate the performance of the inpainting, showing the whole patch. Figure 5.7, presents how each image was presented in the questionnaire. Section B.2 presents some additional screenshots of the questionnaire.

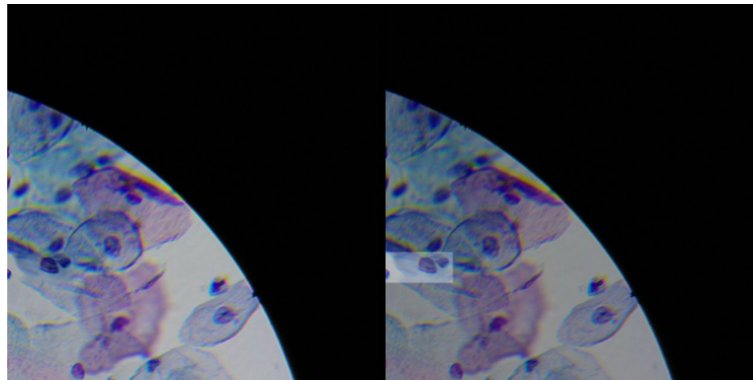


Figure 5.7: Synthetic multi-cellular image presented in the questionnaire to the cytopathologists.

In Figure 5.7, two patches are presented side by side. In fact, both of these patches are exactly the same image, the second one simply indicating that the section of the image identified by the brighter rectangle might have been inpainted. Each image in the questionnaire always has two patches, which are either both real or both generated. It is important to note that when the patch is labeled as "generated," this means that the entire image is unchanged, apart from the rectangles represented by the brighter area. When the image is real, the rectangle simply identifies the zones of the image that contain abnormal squamous cells. The image on the right is obtained by overlaying the corresponding mask with the image on the left with an opacity of 50%.

For each image, the cytopathologists were asked to respond to two questions. The first question asks whether the image is real or generated, in order to evaluate the realism of the image. The second question asks the experts to identify the class of the cell that is being identified by the brighter square. In cases where there is more than one abnormal cell in the image, the experts were asked to insert multiple answers corresponding to the class of each cell. To account for the difficulty of identifying the class of cells based on only the highlighted patch, experts were also allowed to respond with more than one class. In this case, the answer was considered correct if one of them was right.

From this questionnaire, it is possible to extract many valuable data, such as the comparison between the percentage of synthetic images which were considered real, versus the percentages of true positives for real images, and the agreement between the expert's opinion and the dataset annotations of the neoplastic changes according to The Bethesda System.

Chapter 6

Results and Discussion

This chapter presents the findings of the experimental studies of this work. The study focused on two approaches: a Region-based approach and a Nuclei-based approach, using two different datasets. The chapter presents a detailed analysis of the results obtained from the experiments, including a comparison of the generated images with real images, an evaluation by two cytopathologists, and the results obtained by training detection models with the generated images. The chapter also discusses the implications of these findings for the use of synthetic images in the field of cervical cancer diagnosis and research.

6.1 Fine tuning stable diffusion model

This section presents the results of fine-tuning the Stable Diffusion model using various configurations of regularization images, training images, and step counts, in order to optimize the performance of the model for image inpainting tasks in the context of cervical cancer cell detection.

6.1.1 Model selection

In Figure 6.1, it is illustrated a comparison between 12 images generated by the XavierXiau model (a) and 12 cells generated by the JoePenna model (b) for the regions instances of the single cell dataset.

The results of the initial experiments using the XavierXiau and JoePenna Dreambooth models with default settings were not very promising. As seen in Figure 6.1, the generated images were not very realistic, with poor color representation and a lack of detail. The XavierXiau model, in particular, struggled to capture the unique characteristics of the carcinoma cells, such as the large nuclei and lack of cytoplasm, as well as the original colors of the images.

In contrast, the JoePenna model was able to produce images that were more faithful to the original images, particularly in terms of capturing the size and shape of the nuclei. However, even in the JoePenna model, the color representation was not fully accurate. Additionally, both models were able to generate images for only one class at a time, which limited their applicability.

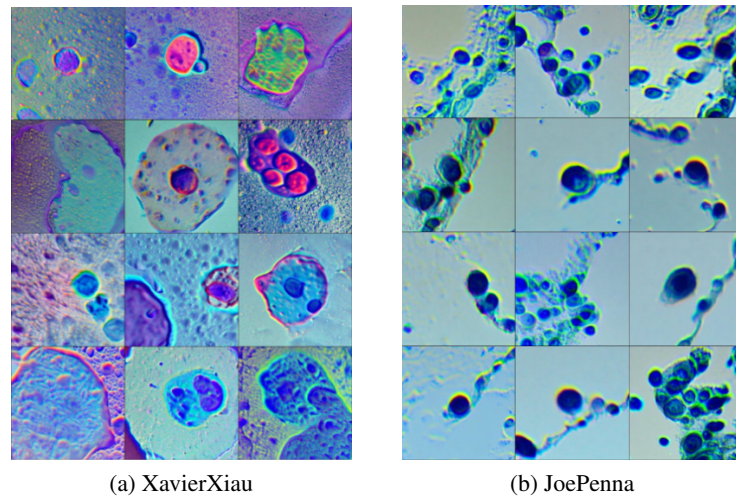


Figure 6.1: Comparison between (a) XavierXiau Dreambooth model and (b) JoePenna.

For the nuclei model, the performance of the JoePenna model was also evaluated by training it with different numbers of steps, 2000, 5000 and 10000. The results of these experiments are illustrated in Figure 6.2.

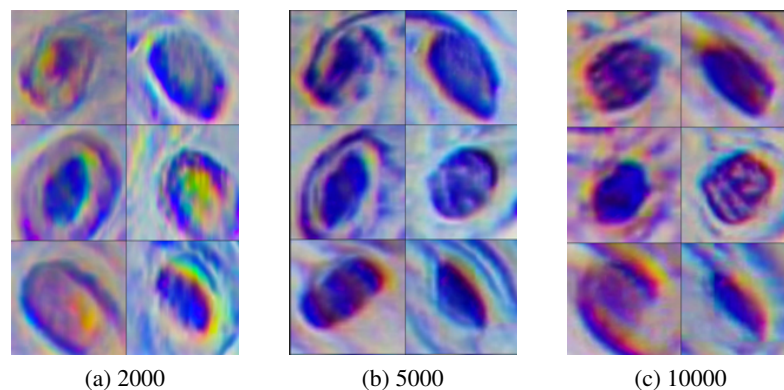


Figure 6.2: JoePenna model results for the Nuclei dataset. Trained with 2000 steps (a) 5000 (b) and 10000 (c).

Visual analysis of the generated images revealed that the model trained with 2000 steps exhibited a significant degree of color aberration, while the models trained with 5000 and 10000 steps produced images with superior quality. There is little difference between the images produced by the 5000 and 10000 step models, despite the latter using twice the resources to train. However, even higher step counts resulted in low-quality images and a poorly represented background. The low quality is due to the fact that the quality of the generated images is directly related to the quality of the original dataset. This finding suggests that the number of steps required for the JoePenna model to converge is dataset-dependent. Both the realism and the abnormality class of the nuclei are harder to evaluate by visual inspection than the cells, due to their quality, isolation, and lack of expertise.

After these experiments, the ShivamShrirao model was evaluated. The initial results for the carcinoma class using the ShivamShrirao model are shown in Figure 6.3, where a comparison is made between the generated images for the carcinoma class (a) and the original training images (b).

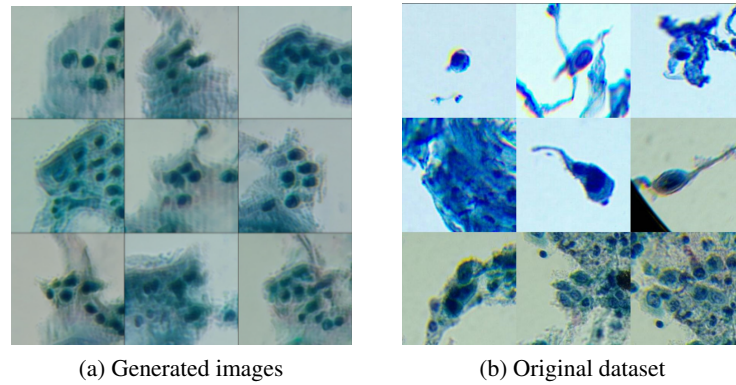


Figure 6.3: Comparison between generated images (a) with the model and the original images from the training dataset (b).

The generated images show similarities to the training images, both in terms of color and the distribution and format of the cells. The carcinoma class is characterized for having a high percentage of regions, in contrast to the single cells that the model aims to generate, which greatly impacted the results. It is possible to observe that the obtained images are still blurred, also due to the high variability of the training images, as well as of the regions that they contain.

When trained on the nuclei dataset, the ShivamShrirao model produced images as shown in Figure 6.4. The Figure compares the generated ASC-US images using the ShivamShrirao model (a) and the original ASC-US dataset (b).

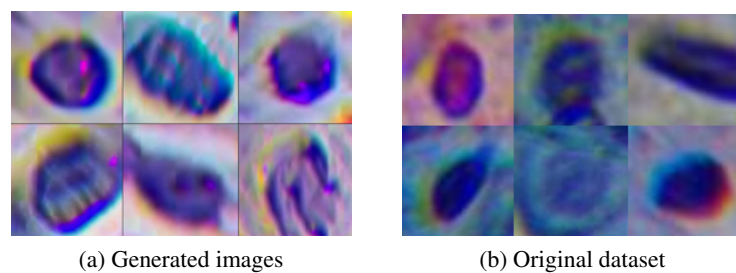


Figure 6.4: Comparison between generated nuclei with the ShivamShrirao model (a) and original images from the nuclei dataset (b).

The results indicate that the ShivamShrirao model, like the JoePenna model, generated images of low quality when trained on the nuclei dataset. This is likely due to the limitations of the original dataset, which is of low quality and less diverse than the regions' instances of the single cell dataset.

Upon visual analysis of the generated images, it was observed that the ShivamShrirao model produced images that resemble the original images, although it was found to be challenging to evaluate the nuclei images with precision. This could be due to the complex nature of nuclei structures, which can be difficult to capture and replicate in generated images. However, it was also noted that the images generated by the ShivamShrirao model appeared to be more realistic than those generated by the JoePenna model, as they featured a more homogeneous background. This suggests that the ShivamShrirao model may be better equipped to handle and replicate the background of the images, which is also an important aspect of image realism.

After carefully considering the various versions of Dreambooth available, it was decided to use the model developed by ShivamShrirao. This model has several advantages: Firstly, it has the ability to generate excellent images in a relatively short period of time, making it efficient for the research purposes of this study. Additionally, it is capable of training multiple tokens, allowing to train all the different classes of cancer cells in a single model. This saves time and resources compared to using multiple models for each cell class. Furthermore, this model offers the option to continue training from a previous model, which is a useful feature. Overall, the ShivamShrirao model is the most suitable for this work needs, and these initial images already show a glimpse of what this model could achieve.

Finally, it was necessary to choose the Stable Diffusion version that would be fine-tuned. The results can be seen in Figure 6.5.

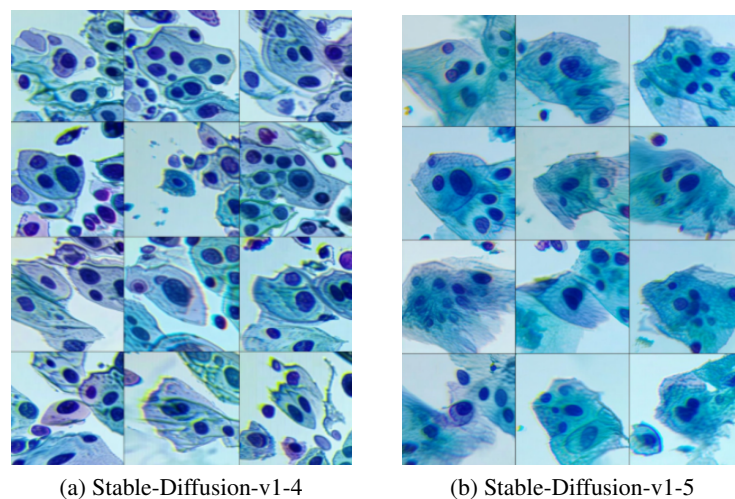


Figure 6.5: Comparison between fine-tuning the model over the sd1.4 model (a) or the sd1.5 (b).

Upon examination of the images, it was observed that the 1.5 model produced better results overall when compared to the 1.4 version. While both models produced a relatively similar output, regarding the shape of the cells, number of nuclei, and color of the generated cells, the Stable-Diffusion-v1-4 images were characterized by sharper contours around the cells, which reduced the realism of the image. Therefore, it was concluded that version 1.5 was the best option for fine-tuning with the Dreambooth model.

6.1.2 Hyperparameter tuning

The regularization images play a crucial role in determining the output of the Dreambooth model. These images are a class of reference images that help address the issues of overfitting and class preservation. Four experiments were conducted to investigate the impact of regularization images on the Region-based Dreambooth model.

6.1.2.1 Regularization images

In the first experiment, all cells from the database were used as regularization images. The results of this experiment can be seen in Figure 6.6, which shows 12 generated images of ASC-US (a) and 12 images of HSIL (b).

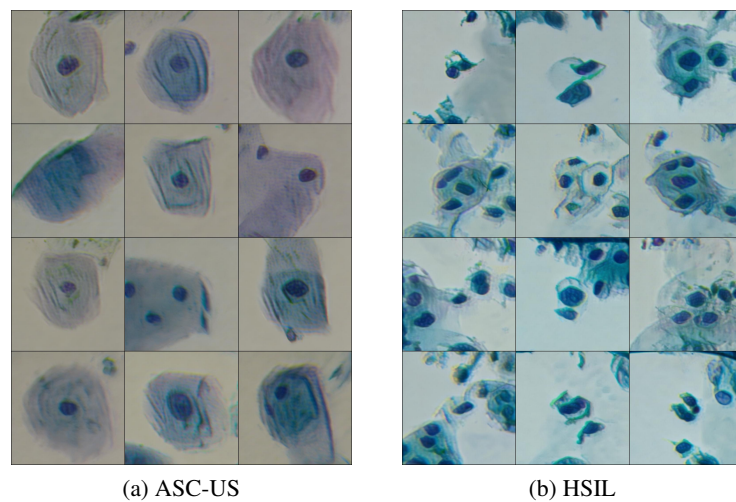


Figure 6.6: Generated images using all the regions instances of the single cell database as regularization images for the Region-based approach.

The findings indicate that the model demonstrated an understanding of ASC-US images when generating single cells; however, the synthetic cells exhibit a lack of clarity and well-defined boundaries, as well as an unrealistic distribution of nuclei. Additionally, the model demonstrates greater difficulty in understanding HSIL cells, which can be attributed to the tendency of these cells to aggregate, resulting in a dataset that is primarily composed of regions rather than single cells, as explained in section 2.3.2. Despite these challenges, it is evident that the model is able to differentiate between the two classes, which possess vastly different characteristics.

Regarding the Nuclei dataset, the outputs obtained from using all images from the dataset as regularization images can be seen in Figure 6.7

The resulting images, as depicted in Figure 6.7, are representative samples of the ASC-US and HSIL outputs generated by the model. Upon visual inspection, the generated images appear to be similar to the input images. However, a preliminary observation suggests that the model may struggle to differentiate between the ASC-US and HSIL classes, as the generated images for both classes appear to be similar to the untrained eye. Further investigation, such as quantitative

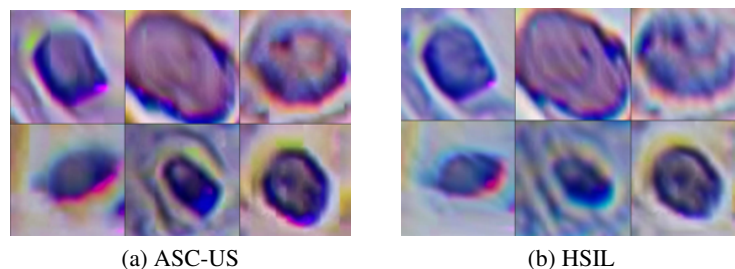


Figure 6.7: Generated images using all the nuclei instances of the single cell dataset as regularization images.

analysis and comparison with other models, is necessary to gain a deeper understanding of the model's performance and ability to differentiate between these classes.

In the second experiment, all cells from the datasets were used, excluding the class that was being trained. The outputs of the ASC-US and HSIL images can be seen in Figure 6.8.

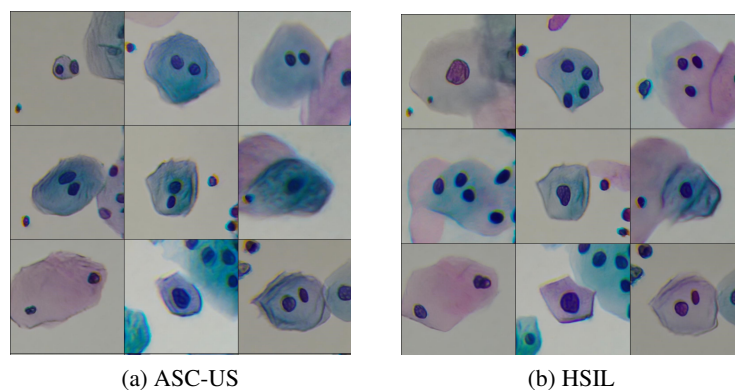


Figure 6.8: Generated images using all the regions instances of the single cell dataset, apart from the class that is being trained as regularization images.

The results show that removing the training class improved the definition of the output, resulting in more precise contours for both classes. However, it also made the classes more general, reducing the differences between the four classes of the model. The results of the second experiment, as depicted in Figure 6.8, demonstrate that removing the training class from the regularization images resulted in both the ASC-US and HSIL images having a similar nucleus-cytoplasm ratio. This is noteworthy, as the main characteristic that distinguishes these classes is precisely this ratio.

In the third test, to counteract this problem, only images from the adjacent class that was being trained were used as regularization images. For example, for the training of HSIL, Carcinoma, and ASC-H images were used as regularization. This approach did not bring any substantial differences, producing a similar output to the previous test.

Finally, the default regularization images from the repository, which were images of people, were used as regularization images. This experiment builds on previous research, as described

in Section 5.2.2, which revealed that using unrelated classes as regularization images may yield improved results. The outcome of this experiment is presented in Figure 6.9.

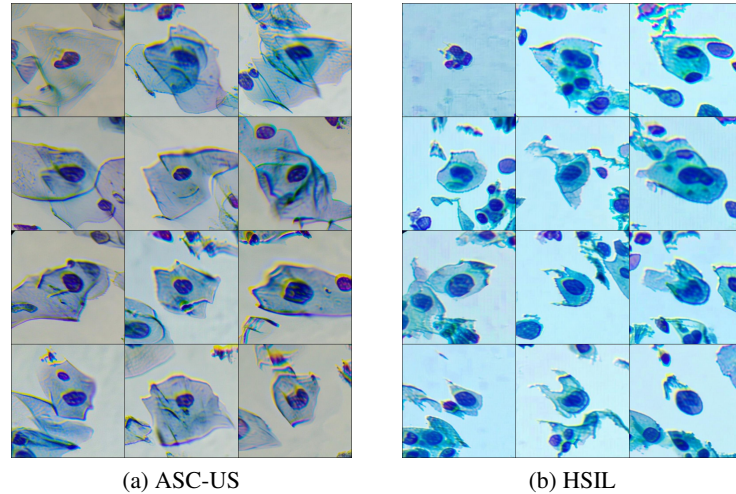


Figure 6.9: Generated images using the unrelated class as regularization images.

The results of the fourth experiment for the regions' instances of the single cell dataset, as depicted in Figure 6.9, demonstrate that using regularization images that are completely unrelated to the training class led to more discernible distinctions between the ASC-US and HSIL classes, as it can be seen for example on the nucleus-cytoplasm ratio. It is worth mentioning that the presence of multiple cells in each image is not a result of the regularization images, but rather a consequence of the selected training images, which also featured regions. Furthermore, the cells exhibited more distinct and well-defined contours compared to the results of the initial experiment. These findings provide strong evidence for the effectiveness of using non-related regularization images in subsequent experiments.

The results of the Nuclei-based approach using people as regularization images were not as successful as the Region-based approach, as evidenced by the unrealistic appearance of the generated images, as shown in Figure 6.10. The generated ASC-US instances (a) and HSIL instances (b) clearly illustrate this point.

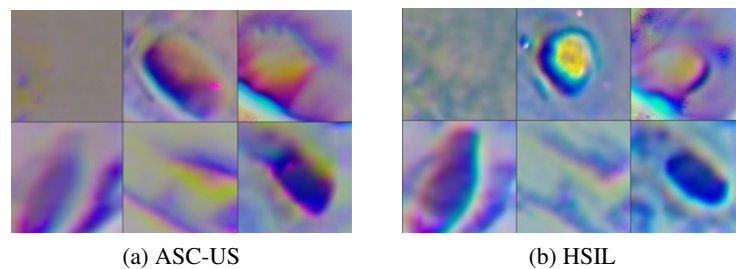


Figure 6.10: Generated images using the unrelated class as regularization images for the Nuclei dataset.

The model struggled to properly learn the shape of the nuclei, with some generated images being blank and lacking any cell structure. Furthermore, the images that were generated were often blurry and poorly defined, and featured multiple unrealistic colors. This experiment highlights the importance of selecting appropriate regularization images for training models and the need for further investigation to improve the model's performance. Compared to the other regularization images, the ones that exhibited more realism were the ones that used all cells for the Nuclei-based approach.

6.1.2.2 Training images

Regarding the training images, a multi-step process was employed to determine the most effective set of images. Figure 6.11 illustrates a comparison of generated ASC-US images, between the approach where all the cells were used as training images, and a random sample of 30 cells.

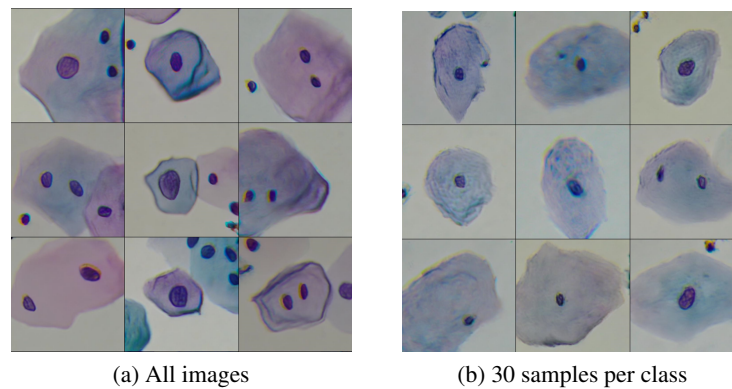


Figure 6.11: ASC-US generated images using all the regions instances of the single cell(a) and only 30 (b) for training.

An examination of the images from the second test revealed that they were more representative of the true nature of the cells being studied. The images produced in the first test, however, showed that the model had learned to recognize cells as comprising multiple dispersed nuclei and overlapping cells. In contrast, the images produced in the second test depicted individual cells with a single, distinct nucleus, which is a more accurate representation of the cells being studied.

The results of the three subsequent experiments made with people as regularization images are presented in Figure 6.12.

Even though the same training parameters were used, the output varied greatly depending on the training images used, which reveals the significant impact training images have on the model's output. When using all the images, the model produced many overlapped cells, many of which were not clearly defined and had a very low opacity in addition to containing multiple nuclei and lacking sharp contours. The region's experiment produced the most unrealistic results. Although these images were supposed to depict many cells, the produced images did not resemble real regions but instead resembled a collection of clearly defined single cells. The images also had stronger contours than the original ones. On the other hand, the experiment using images of single

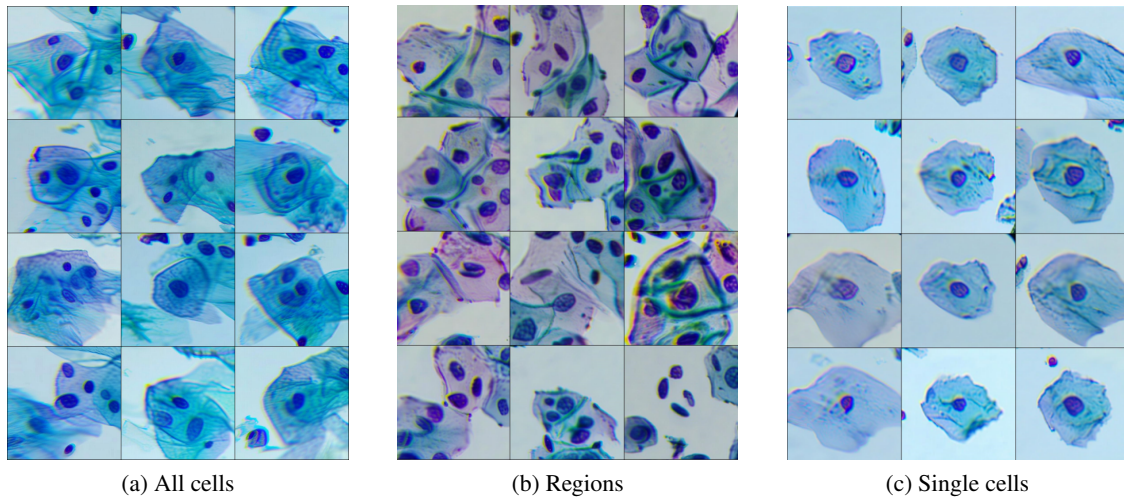


Figure 6.12: Generated images using all the region instances of the single cell (a), only regions (b), and single cells (c).

cells produced the most realistic results, as the model was able to learn to represent each class in a realistic and isolated manner.

This study demonstrates the importance of carefully selecting training images in order to improve the performance of the model. The results of the initial experiments using as regularization images all cells except for the class under examination showed that using a subset of 30 images selected randomly produced more realistic and accurate results than using all images. Additionally, the results of the follow-up experiments using people as regularization images further reinforced the importance of careful image selection. The experiment using images of single cells produced the most realistic results, which highlighted the need to use a diverse set of images for training, covering different variations of the cells, such as different types of cells, lighting conditions, and noise levels, in order to improve the model's robustness and generalization ability. The nuclei dataset utilized a similar approach for the selection of training images as the Region-based approach, however, this decision was made due to the constraints of time and limited resources available.

6.1.2.3 Number of steps

The number of steps is an important parameter that significantly influences the generated images. In the above experiments, all the results were obtained using 10,000 steps; however, the default setting for other Dreambooth models is a significantly shorter number of steps for each class, making it necessary to experiment with different values. To determine the ideal number of steps for the Region-based model, two experiments were conducted. Figure 6.13 illustrates the results of the ASC-H class trained with 5000 steps (a) and with 10000 steps (b).

As shown in Figure 6.13, the output of the 10000 steps is much more defined than the output of the 5000 steps. In the 5000 steps experiment, the cells are not clearly defined, showing overlapped

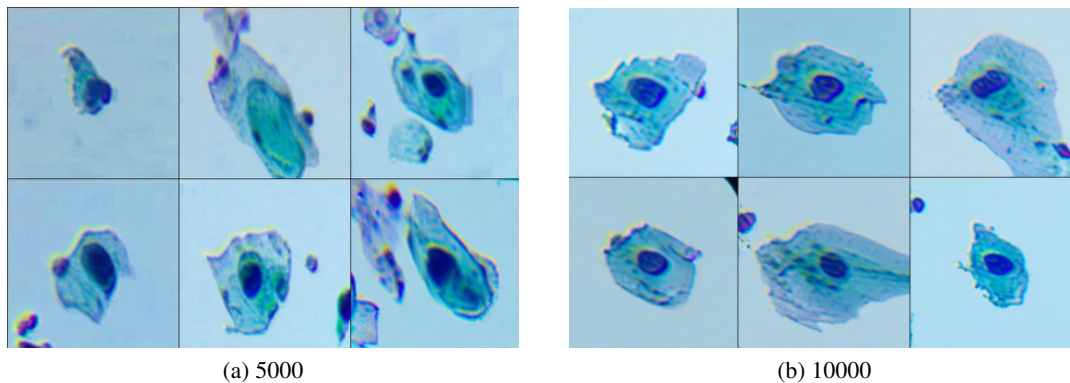


Figure 6.13: Generated ASC-H cells for the Region-based approach using 5000 steps (a) and 10000 steps (b).

cells and even lacking a clearly visible nucleus. As the number of steps increases, the model is able to learn more features from the data, which leads to more defined and realistic images. However, a high number of steps also increases the training time and computational cost. Therefore, it is important to find a balance between the number of steps and the computational resources available. Additionally, it is worth noting that the selection of the number of steps is problem-dependent, and it should be carefully selected to fit the specific task and dataset.

The Nuclei-based dataset experiment was designed to evaluate the effect of incremental step counts on the quality of generated images using all cells as regularization images. The experiment began with 2000 steps and gradually increased the step count by 2000 at a time until reaching a final step count of 8000. The results of the experiment are illustrated in Figure 6.14, which shows a representation of the ASC-US class of images generated with 2000, 4000, 6000 and 8000 steps respectively.

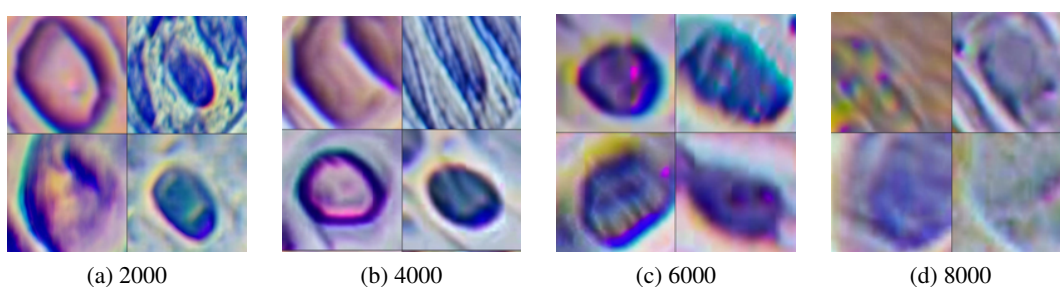


Figure 6.14: Generated nuclei using different step counts.

The results of the experiment indicate that there is a correlation between the number of steps and the quality of the generated images. The images generated with lower step counts, such as 2000, appear to have lower quality and exhibit strange colors, not appearing realistic. As the step count increases, the quality of the generated images also increases, with the best result being observed at the 6000 step count. However, the experiment also shows that there is a point of

diminishing returns, as the 8000 step count experiment resulted in a decrease in image quality, with the model failing to accurately capture the shape of the nuclei.

6.1.2.4 Final parameters

In conclusion, this study demonstrated the importance of regularization images, training images, and number of steps in the Dreambooth model. A multistep approach was employed to determine the most effective set of parameters for the Region-based approach. The results indicated that using an unrelated dataset as regularization images, a subset of 30 hand-picked images for each class as training images and 10000 steps as the number of steps yielded the most realistic and accurate results for the Region-based approach. Illustrative synthetically generated images obtained after parameters tuning for the four classes, ASC-US, ASC-H, LSIL and HSIL respectively can be seen in Figure 6.15.

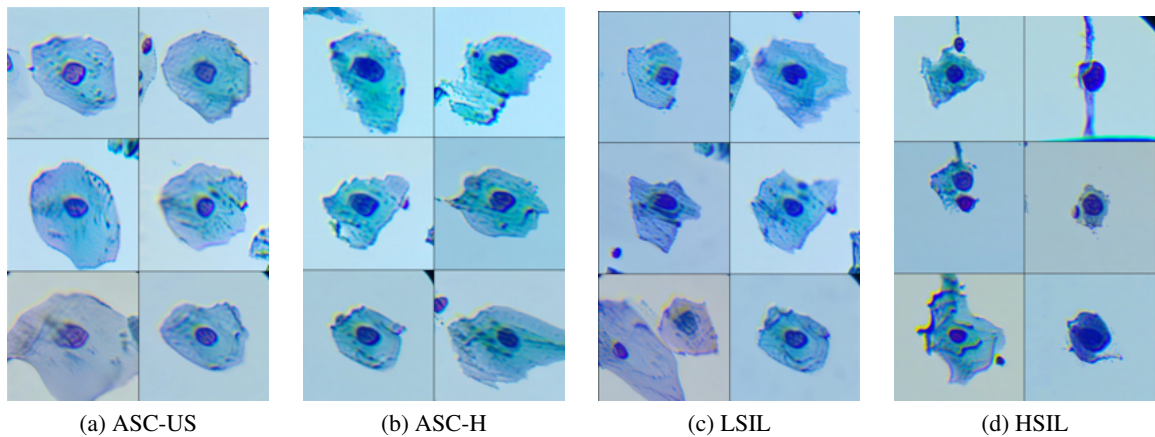


Figure 6.15: Illustrative synthetically generated images obtained after parameters tuning for the Region-based Dreambooth model.

Based on the results of the experiments, the final parameters chosen for the nuclei dataset were all nuclei instances from the single cell dataset as regularization images, a step count of 6000, and 30 handpicked cells as training images. These parameters were selected as they resulted in the highest quality images, with the model accurately capturing the shape of the nuclei and displaying realistic colors. The final outputs for each class can be seen in Figure 6.16

6.1.3 Cytopathologist validation

Regarding the Cytopathologist questionnaire for single cell images, the results considering the realism of the images are illustrated in Figure 6.17.

The questionnaire was completed by two cytopathologists, with a total of 161 questions answered, half of which were based on real images and half of which were based on synthetic images. All of the real images were region instances from the single cell dataset, and all the synthetic images were generated by the developed Region-based Dreambooth model. The results indicate that

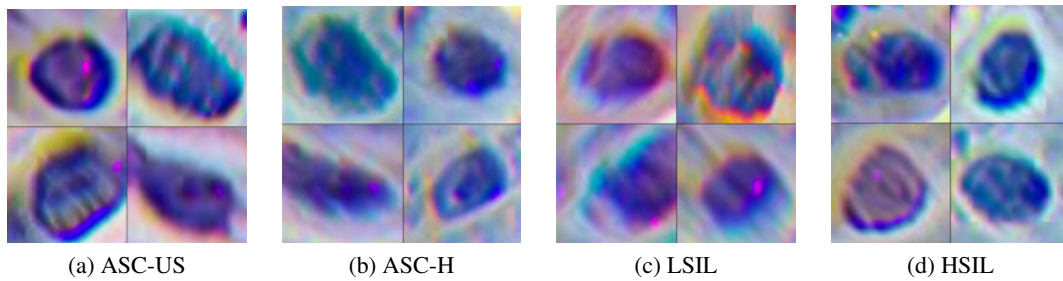


Figure 6.16: Illustrative synthetically generated images obtained after parameters tuning for the Nuclei-based Dreambooth model.

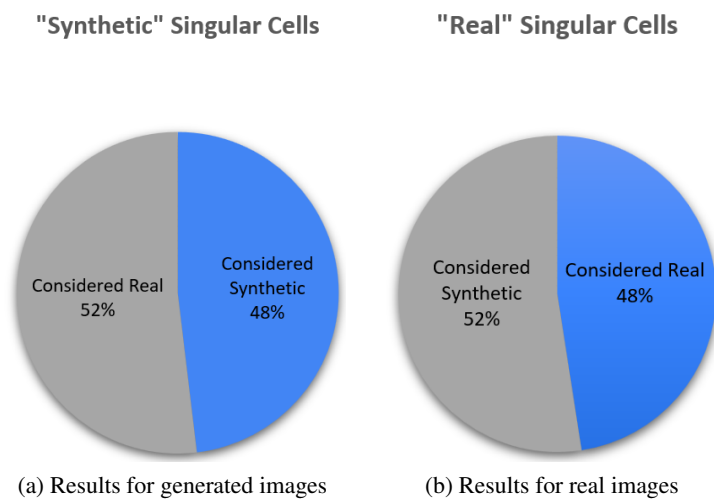


Figure 6.17: Questionnaire results regarding the realism of single cell images.

52% of the synthetic images were considered realistic by the cytopathologists, which is an impressive level of realism. In addition, when examining the real images, 52% were also incorrectly identified as synthetic, the exact same percentage of misclassified synthetic images. Results by specialists are illustrated in section B.1.2.

The high percentage of real images that were incorrectly identified as synthetic may be due to the high degree of realism achieved by the synthetic images. Additionally, the way in which the cells were presented in the images may have further contributed to the cytopathologists' difficulty in distinguishing between real and synthetic images. In particular, the cells were being examined in isolation and at an expanded scale, which may have affected the cytopathologists' ability to accurately identify the images as real. The low quality of the original images of the database used in this study can also be attributed to the equipment used to capture them, specifically the μ Smartscope [71]. The μ Smartscope is a device that utilizes a smartphone camera to capture cytological images. While this type of equipment is portable and cost-effective, it does not have the same level of resolution and image quality as specialized cytological imaging equipment, which can affect the realism of the images generated by the computer algorithm.

In conclusion, the present study demonstrates that synthetic single cell images generated by the Dreambooth model can have a high degree of realism. However, the results also highlight the need for further improvements in image quality when examining cells in isolation, as well as the need to consider the potential challenges that may arise when using synthetic images in practice.

6.2 Multiple Cell Inpainting

This section presents the performance of the multiple cell inpainting method for cytological images, using the Region-based and the Nuclei-based dataset. The dataset preparation process, as described in Section 5.3.1, included the isolation of patches with abnormal cells, the creation of masks, and the consideration of cell size and characteristics. The hyperparameter tuning process, described in Section 5.3.2 was used to optimize the inpainting process to achieve the best generative performance.

After conducting a series of experiments, it was determined that the best results were achieved using the following hyperparameters: full resolution, CFG 4, 100 steps, denoising strength of 0.75, and mask blur of 4. These settings provided the best balance between accuracy and realism in the inpainting process and were used for all subsequent tests in the study.

The initial approach towards multiple cell inpainting involved generating four modified versions of each patch, each representing a different abnormality class for the target cell. A corresponding mask was created for each patch, which was used to identify the regions of the images that would be inpainted for all abnormality classes. Figure 6.20 illustrates this approach, with example (a) showcasing the inpainting of a region and example (b) showcasing the inpainting of a single cell. Each image is constituted of the original patch, the corresponding mask, and the four inpainted images, one for each class.

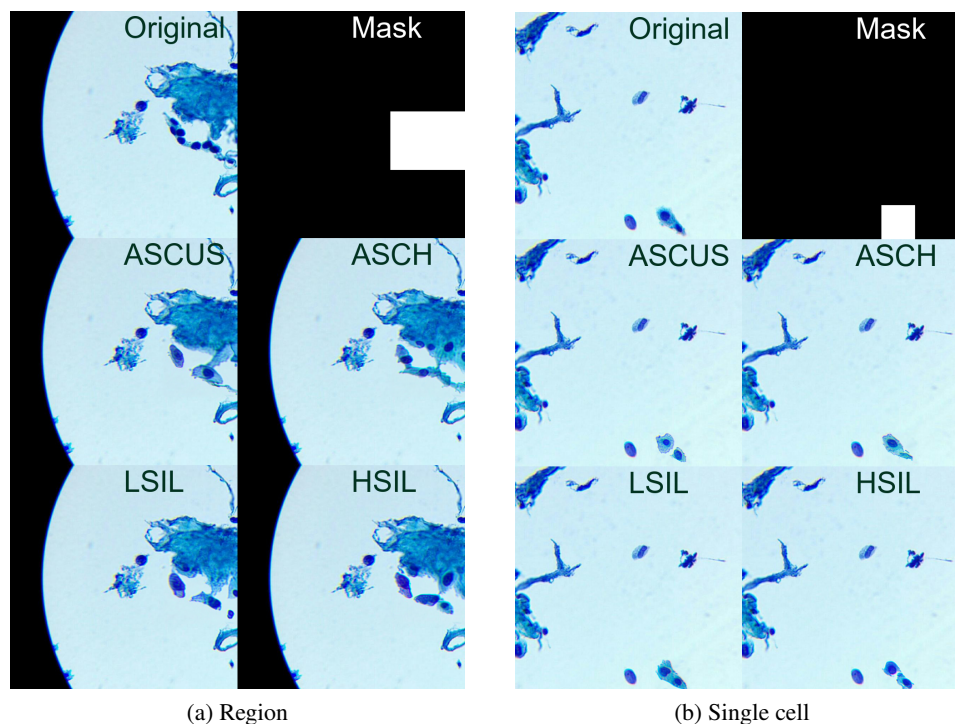


Figure 6.18: Inpainted Region-based patches for each abnormality class with fixed inpainting area size.

The results of this approach were highly realistic, as demonstrated by the inpainted cells in both images. In the inpainting of a region (a), it is evident that the generated images closely resemble the cells that were present in the original patch, indicating that the model effectively considered the existing cells. Furthermore, the model also appears to have taken into account the class of the abnormality being inpainted, as demonstrated by the significantly lower nuclei-cytoplasm ratio of the HSIL generated cells in both (a) and (b) when compared to other classes such as ASC-US.

The same approach was taken for the nuclei dataset, where all cells with any abnormality class in the original patch were transformed into four generated patches, one for each class. In Figure 6.19, some generated batches are illustrated for two original patches, each with a corresponding mask.

It was observed that the generated nuclei images were not as realistic as the previous approach, often appearing blurry. While further evaluation by specialists is needed, it appears to the untrained eye that the generated nuclei were very similar to one another, not being possible to clearly distinguish between the various classes. However, it is important to note that this similarity is also present in the training images. The most easily observable difference between the different abnormalities classes is the size of the nuclei, which is constrained by the use of a fixed-size mask. Therefore, it is expected that the use of resized masks in the experiment will yield better results.

In an effort to impose limitations on the model's size, a series of experiments were performed in which the area of the patch being inpainted was varied in accordance with the abnormality class.

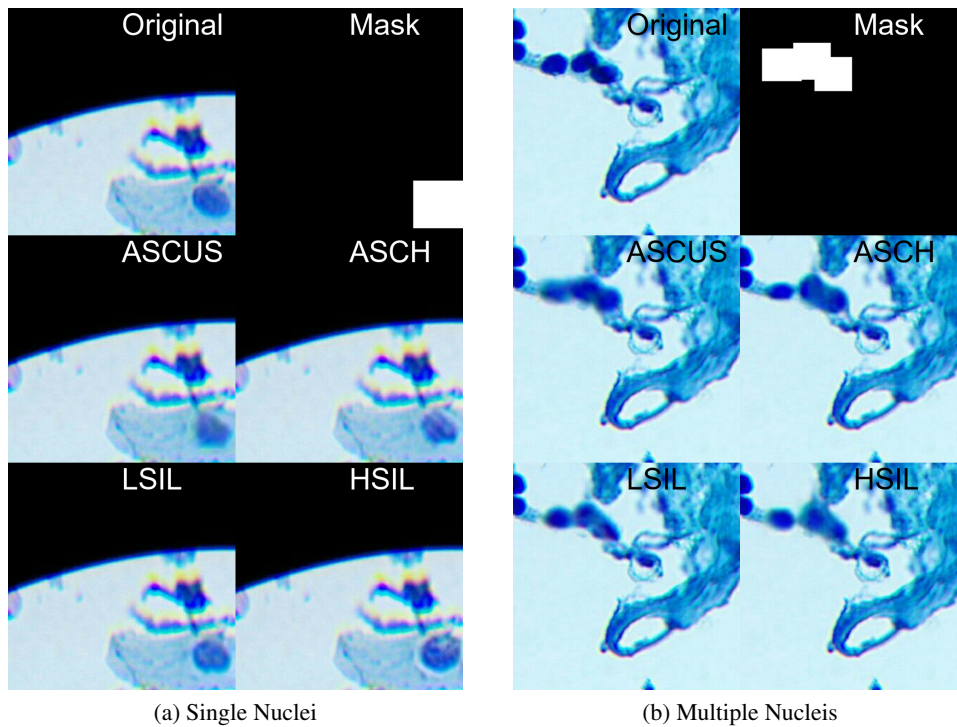


Figure 6.19: Inpainted Nuclei-based Patches for each abnormality class with fixed inpainting area size.

In the initial experiment, the Lamma tool (as detailed in section 4.5.4) was utilized to eliminate the original cell from the starting patch, enabling the inpainting of a smaller cell without it appearing artificially generated. However, due to the fact that the inpainting process takes into account the original area of the image being inpainted, a significant number of the generated cells appeared to be blank, resulting in a dataset that was too small to effectively support the distribution of data augmentation present in the original work [73]. As a result, a second strategy was implemented in which the generated cells were directly inpainted over the original patch, even if the new cell was smaller in size than the original. The outcomes can be observed in Figure 6.20.

From the analysis of Figure 6.20, it can be observed that the ASC-US image exhibits a slightly larger dimension compared to the HSIL image. Additionally, it is apparent that the left and right regions of the HSIL cell remain consistent with the original image, as only the central portion underwent inpainting. For the Nuclei-based approach, it was also used the resizing approach, however, since the squamous nuclei were not significant for this work and were the smallest class, it was possible to only modify these cells. This had the advantage that the Lama tool was no longer needed for this approach, as the inpainted nuclei would always be bigger than the original one. This approach resulted in more distinguishable nuclei between classes, which is expected to provide better results in the detection models.

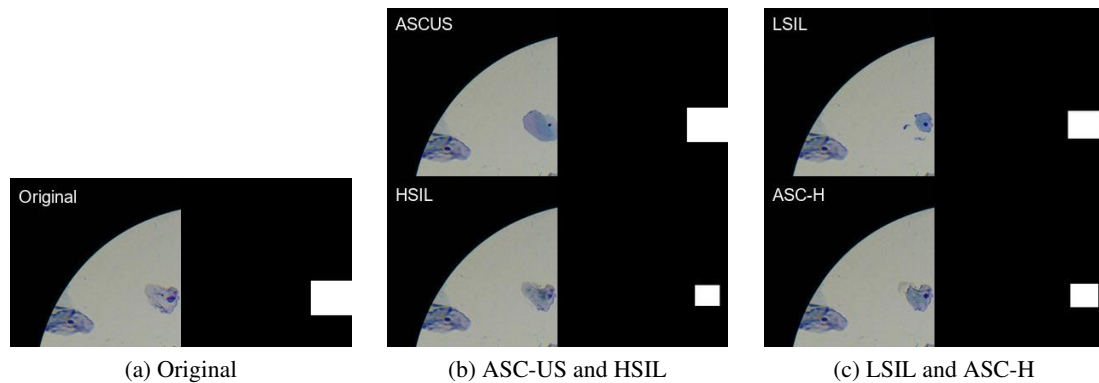


Figure 6.20: Inpainted Region-based patches for all the abnormality classes with variable inpainting area size and respective masks.

6.2.1 Cythopathologist validation

In this section, it is presented the results of the Cythopathologist questionnaire for multiple cells. The results regarding the realism of the patches are presented in Figure 6.21.

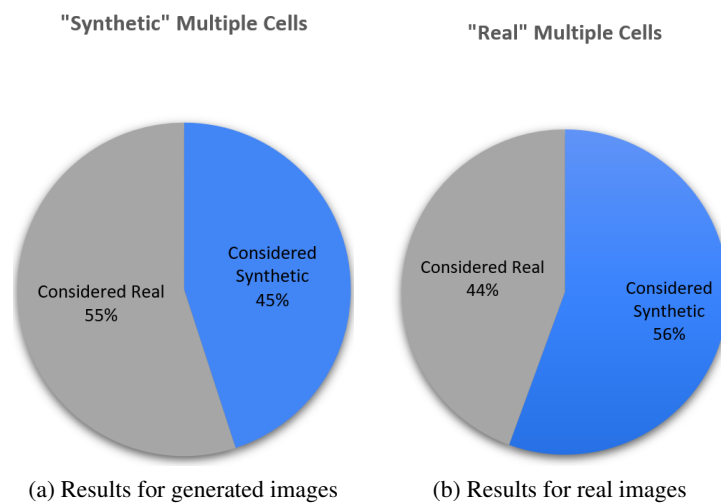


Figure 6.21: Questionnaire results regarding the realism of multiple cell images.

The results obtained from the synthetic images were found to be highly similar to those of the single cell questionnaire, with 55% of the synthetic images labeled real. Since this questionnaire had images of the whole patch, as opposed to expanded low-resolution single-cell images, the specialists were able to make more accurate decisions through the use of reference cells. Nonetheless, due to the realism of the synthetic patches, the specialists could not distinguish between the real and generated cells. As for the real images, there was an 18% difference in comparison to the single cells' questionnaire, with specialists labeling them correctly 56% of the time. This may be attributed to the superior quality of the real images and the ability to evaluate the entire patch as opposed to just one cell. Overall, it can be concluded that the generated images were highly

realistic and indistinguishable from real images, even for specialists. Results by specialists are presented in section B.2.2.

Another important characteristic of the model is that it should distinguish between the different classes of abnormalities, reproducing the specific characteristics of each one. Although all of the images had an abnormality class, the specialists opted to also have an extra option in which they would consider that the cell did not have any abnormality class. Although this option would always be incorrect in the questionnaire, since the specialists knew all the instances had been previously annotated as having an abnormality class, they chose to include this option in case they doubted the previous annotation. There were a total of 100 answers regarding the abnormality class, only from one specialist. The final results regarding the neoplastic changes can be seen in Figure 6.22.

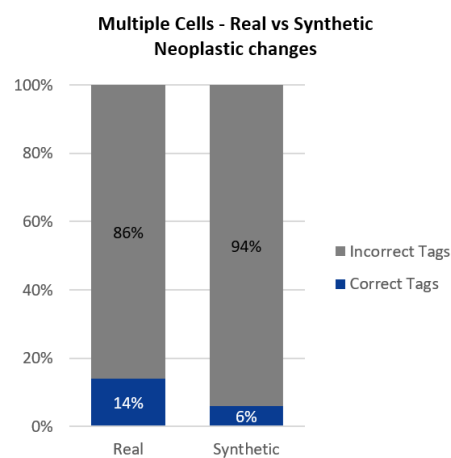


Figure 6.22: Agreement between ground truth and cytopathologist in evaluating abnormality class of multiple cells.

The data illustrated in this graph reveals a high number of misattributions. The real images had a disagreement rate of 86% between the cytopathologist evaluation and the ground truth, with a total of 78% being classified as not having an anomaly. The synthetic images had a disagreement rate of 94%, with a total of 70% being classified as not having an anomaly.

These results suggest that the model was not fully able to differentiate between the different abnormality classes, resulting in a lower agreement percentage compared to the real images. However, cytological annotations are highly subjective, with the same specialist potentially providing different annotations for the same cell at different times, making it difficult for the model to learn the cell class when even the specialists who labeled the dataset have doubts. The initial image annotations of the questionnaire were annotated by a single expert and due to time constraints, only one specialist responded to the questionnaire. The results of this review revealed an 86% disagreement rate between the two specialists, as it can be seen by the results of the real images. Although the synthetic images received a low score, there was only an 8% discrepancy observed when comparing them to real images. The subjective nature of the Region-based dataset was already identified to be a possible issue in previous works [73, 53] and can influence the results that the generated images have when used as input for the Region-based detection model.

It is important to note that the model was trained using a dataset of single cells, and it was expected to understand the class of the cells from these images alone. However, specialists noted that it would be difficult to evaluate the single cell questionnaire with regard to abnormality as there is not enough information present. With the guidance of a specialist in selecting the appropriate training instances, ensuring both diversity and representativeness of the variability intrinsic to each class, the results could be improved, as the model would be directed to learn the correct classes and could be evaluated more accurately, as opposed to relying on visual inspection of an untrained eye. Additionally, the dataset used in this study was annotated by a single specialist due to resource limitations. However, for future research, it is recommended to utilize multiple specialists for annotation to reduce subjectivity in the data. This approach would significantly enhance generative models, as while Dreambooth models require relatively few samples (less than 30) for training, the higher the quality of annotations, the better the results.

6.3 AI-based Cervical Lesions Detection Algorithms

In this section, the results of the Region-based and Nuclei-based lesion detection algorithms are presented. A comprehensive comparison of the two models will be conducted at the end of this section.

6.3.1 Region-Based Detection Model

With regard to the Region-based approach, four tests were conducted. The methodology for identifying each test case is described as follows:

- TR0 - Baseline results of the original model trained without generated images [73].
- TR1 - Increasing the dataset volume using fixed-size synthetic image generation.
- TR2 - Increasing the dataset volume while controlling the size of the generated synthetic images.
- TR3 - Comparing the impact of data augmentation using fixed-size synthetic image generation versus CDA.
- TR4 - Comparing the impact of data augmentation using synthetic image generation versus CDA, while controlling the size of the generated images.

To quantitatively evaluate the performance of the Region-based model, two metrics were utilized: the mean average precision (mAP) at 50% intersection over union (IOU) for all classes, and the mean average recall for all images with a maximum of 10 detections (AR10). The results of these analyses are presented in Figure 6.23, which includes two graphs. The graph 6.23a displays the mAP@.50IOU for each abnormality class and test, while the graph 6.23b displays the

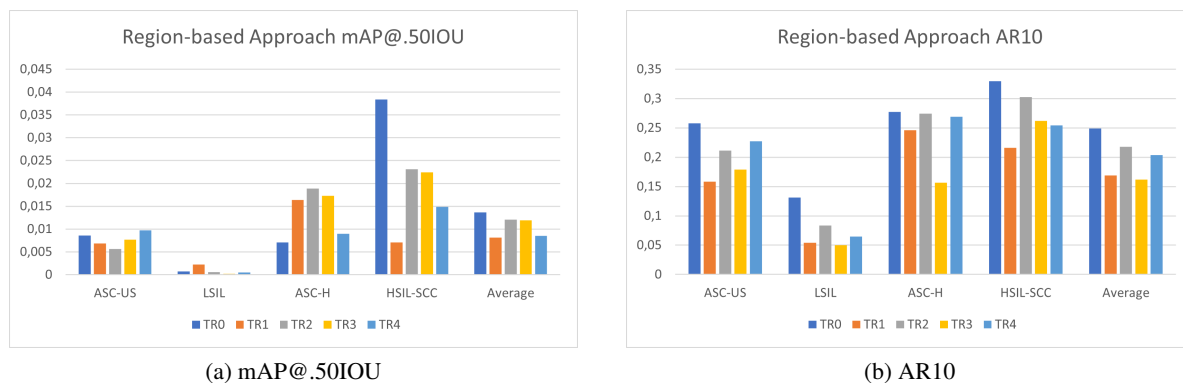


Figure 6.23: Results of the Region-based detection model. Both graphs present the values for each abnormality class and for each test. Graph (a) presents the mAP@.50IOU values and graph (b) presents the AR10.

AR10 for each class and test. Additional information regarding these results can be consulted in subsection A.1 in the Annex.

Regarding average mAP@.50IOU and AR10, Figure 6.23 illustrates that the baseline approach still presents superior results for both performance metrics. The suboptimal performance of the models trained with the generated images can be attributed to several underlying factors. One possible issue is that the developed Dreambooth model was not designed to effectively handle both single cells and regions simultaneously, as demonstrated in subsection 6.1.2.2. This limitation resulted in the utilization of a model that could only generate single cells, which caused a substantial discrepancy between the proportion of single cells and regions in the training and test sets. This likely had an adverse effect on the model's performance, as it was not exposed to an adequate number of region-based images during the training process. However, this mismatch in the ratio of single cells to regions in the training and test sets is also related to the hyperparameters used in the model. The hyperparameters chosen were based on the best results obtained by the original detection model, given time constraints. Future studies should investigate the optimal set of hyperparameters for each test, taking into account factors such as the ratio of single cells to regions in the test set in order to improve the performance of the model.

The poor performance of the generated images may have been further exacerbated by the significant variability in the format of the cells within the same class and their consequent similarity to other classes. For example, the cytoplasm can take on different shapes and sizes, even when considering cells of the same class. This similarity may have made it more challenging for the model to accurately distinguish between different classes. This hypothesis is supported by the results of the cytopathologist's multiple cell questionnaire, which indicates that real images had extremely low agreement rates regarding neoplastic changes, with a disagreement rate of 86% between two specialists, as seen in Figure 6.22. Since the annotation process and the questionnaire are different tasks administered under different conditions to both specialists, the results cannot be directly compared. However, the results highlight the subjectivity of the annotation process for

these images and the potential for variability in interpretation among different specialists. This subjectivity of the cytological images is a well-known problem in the field, and more research is needed to overcome this limitation [55, 73]. An additional factor to consider is that the Dreambooth model was trained using images that only included individual cells. However, experts in cytopathology stated that it would be impractical to classify abnormalities based solely on a single cell as it does not possess enough information for accurate classification. Cytopathologists often need to compare the cell that is being analyzed with an intermediate squamous cell of the same sample, in order to be able to infer the lesion class, along with various other techniques.

Additionally, while the overall average metrics for the original model were superior to the other tests, a closer examination reveals that this is primarily due to the high performance in the HSIL class. When examining the other classes, it is evident that at least one test outperformed the original model in each class, such as the ASC-H class, where all tests had a higher mAP@.50IOU than TR0. However, it should be noted that for the AR10 metric, TR0 showed a better performance across all classes.

Evaluating the Impact of Controlling the Size in Image Generation

In this study, two methods were employed for the image generation model: a fixed-size method and a controlled method in which the size of the generated image was modulated based on the class, as detailed in subsection 5.3.1. In order to evaluate the effectiveness of the two approaches, it was conducted experiments using test sets TR3 and TR4. Both sets were generated using the same number of patches, but TR4 was created using a controlled approach, in which the size of the generated cell was varied according to the target class.

The results, as shown in Figure 6.23, indicate that while the controlled approach had a lower average mAP@.50IOU, it had a higher AR10. This suggests that the total number of annotations increased, increasing both the number of true positives and false positives. One potential explanation for this phenomenon is that by altering the dimensions of the cells in the training dataset, the likelihood of identifying cells with more pronounced variations in size in the test dataset was heightened. In particular, the controlled approach showed improvements for the mAP@.50IOU in certain classes such as LSIL and ASC-US, however, it is important to note that the LSIL class had a very low precision, independently of the approach. One potential explanation for the observed improvement in the ASC-US class with the controlled approach may be attributed to its larger size, which constitutes a distinct characteristic compared to the other classes. Furthermore, it was observed that the inpainted cells frequently did not occupy the entire masked region, but rather only a portion of it. This phenomenon could have particularly impacted the smaller classes, namely HSIL and ASC-H (as depicted in Table 5.4), potentially resulting in even smaller inpainted cells than the already resized mask. As illustrated in Figure 5.6, it was observed that the model generated cells of varying sizes based on the class, even without the use of controlled masks. The decrease in average mAP@.50IOU for TR4 in comparison to TR3 can be attributed to the additional complexity introduced by the controlled size approach.

The evaluation of the AR10 metric revealed a significant improvement with the utilization of the resized approach, from TR3 to TR4, yielding an average increase of 20%. This improvement was observed across all classes, with the exception of HSIL, which exhibited similar performance to the other approach. Given the trade-off between the mAP@.50IOU and AR10 metrics, it is not possible to conclusively determine which approach yields the optimal performance for the Region-based method.

The results of the TR1 and TR2 tests indicate that the controlled approach outperforms TR1 in terms of mAP@.50IOU and AR10 metrics. However, it is worth noting that TR2 was trained using a dataset of isolated images, which was smaller in size, comprising only 70% of the dataset used for TR1. This reduction in synthetic images may have contributed to the improved performance observed in the controlled approach.

Comparing the Effectiveness of Conventional Data Augmentations and Generative Approaches

The tests TR3 and TR4 had the objective of comparing two different data augmentation approaches: generative and basic image manipulation (CDA). To do this, it was generated synthetic images with the same per class distribution of the CDA, which were used as input to the Region-based detection model. This way it was possible to compare the performance of both approaches. From Figure 6.23 it is possible to see that the CDA had a superior performance boost when compared to the generated images, with an average mAP@.50IOU and AR10 superior to TR3 and TR4. Since the best performance of the original detection model was obtained by adding basic image manipulations, the reasons for this superiority are the same as those previously explained at the start of this subsection 6.3.1.

Final evaluation

The utilization of synthetic images in the Region-based approach resulted in a decrease in performance, as measured by both mAP@.50IOU and AR10. This decline in performance can be attributed to various factors, including the limitations of the Dreambooth model in generating only single cells, the hyperparameters used, the similarity between the lesion classes, and the subjective nature of cytological images. Among the various tests with generative approaches conducted, the TR2 yielded the best results, in which it was enhanced the dataset volume while maintaining control over the size of the generated images. This approach demonstrated superior mAP@.50IOU and AR10 results compared to the other approaches. It is important to note that the assertion made by the original authors that the lack of images could be a bottleneck for the detection model is speculative in nature and based on assumptions. Additionally, there may be other underlying issues that are impacting the performance of the detection model, which may not be resolved through the addition of more images alone.

6.3.2 Nuclei-Based Detection model

In this section, the results of the Nuclei-based detection model are presented and discussed. The identification of each test will be made in the following manner:

- TN0 - Baseline results of the original model trained without generated images [53].
- TN1 - Increasing the dataset volume using fixed-size synthetic image generation.
- TN2 - Increasing the dataset volume while controlling the size of the generated synthetic images.
- TN3 - Comparing the impact of data augmentation using fixed-size synthetic image generation versus basic image manipulation.
- TN4 - Comparing the impact of data augmentation using synthetic image generation versus basic image manipulation, while controlling the size of the generated images.

The model's performance was quantitatively evaluated using the $mAP@.50IOU$ for all classes, and the AR100. The results of these evaluations are presented in Figure 6.24, which includes two graphs. Graph 6.24a shows the $mAP@.50IOU$ for each abnormality class and for each test, while graph 6.24b displays the AR100 in a similar format. For a more detailed breakdown of the results, refer to Table A.3 in the Annex.

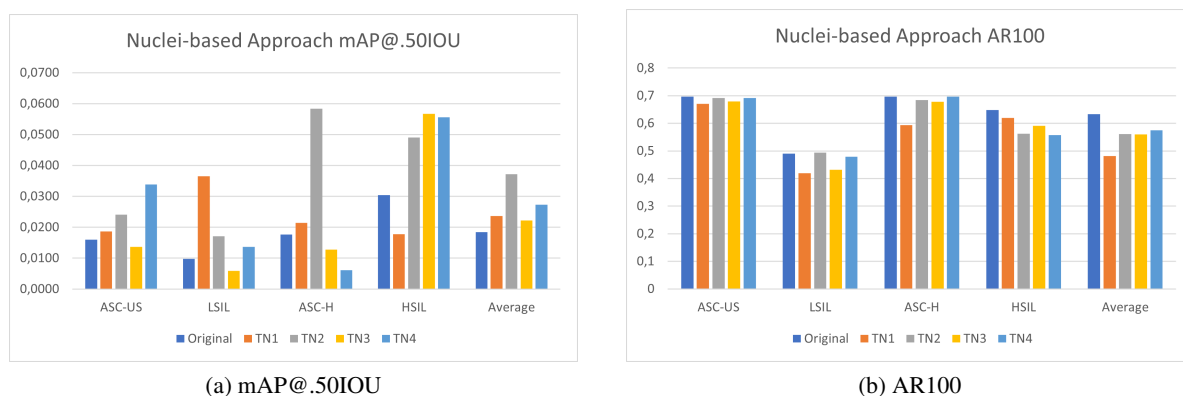


Figure 6.24: Results of the Nuclei-based detection model. Both graphs present the values for each abnormality class and for each test. Graph (a) presents the $mAP@.50IOU$ values and graph (b) presents the AR100.

The utilization of synthetic images resulted in an enhancement in the $mAP@.50IOU$ for all scenarios. The initial test, TN1, displayed a notable enhancement in the average $mAP@.50IOU$ metric, with an improvement of 28%. TN2 exhibited the highest average $mAP@.50IOU$ value among all tests with a 101% increase, while TN3 and TN4 displayed 20.5% and 48% increases respectively. On the other hand, the AR100 displayed a decline in comparison to the original experiment, particularly in TN1 with a 23% decrease. Other tests experienced a smaller decrease

of 9%-12%. The incorporation of synthetic images resulted in a substantial enhancement in the mAP@.50IOU of the model, indicating that the generated images made a significant contribution to the model's performance. Despite the decline in the AR100 for all tests, the magnitude of this decrease was relatively small compared to the major improvements in the mAP@.50IOU.

In addition to the overall enhancement in the mAP@.50IOU of all the tests, it is also important to examine the improvement of each individual class. The ASC-H class in the TN2 test exhibited particularly noteworthy results, achieving a mAP@.50IOU of 0.0584, the highest value among all tests and classes. Additionally, the HSIL class displayed higher values than the baseline model for all tests except for TN1. Conversely, the LSIL class in the TN1 test displayed the highest percentual improvement, with an increase of 275% compared to the original model. This suggests that while certain methods may be more effective overall, other methods may have advantages in specific classes. Further experimentation could be conducted with the goal of maximizing performance for each class, in order to achieve a combination of models that provide the best overall results.

The Effects of Data Volume and Controlled Distribution

In the first two tests, the primary objective was to incorporate a substantial number of synthetic images to evaluate the model's performance. In contrast, in TN3 and TN4, a more controlled distribution of the dataset was utilized. The approaches in which a large number of synthetic images were used yielded the best results for the LSIL and ASC-H classes, which were the classes with the smallest representation in the original dataset as seen in Table A.6. This boost in performance can be attributed to the addition of images to the underrepresented classes, as the lack of data for these classes was identified as one of the major limitations of the original detection model. Conversely, when a controlled number of images were added, the classes with the most representation tended to perform better. This suggests that while the most underrepresented classes benefit from a large amount of data, the ratio of original and synthetic images must be taken into account, particularly for classes with sufficient instances. An excessive number of synthetic images can reduce the overall quality of the dataset by introducing unnecessary noise. Although the optimal ratio was not studied in this research due to time constraints, it would be of interest to conduct further experiments to determine it.

Evaluating the Impact of Controlling the Size in Image Generation

In this study, two approaches were employed for the image generation model: one in which the model had complete autonomy, and another in which the size of the generated image was controlled based on the class. Results from the comparison of TN1 and TN2, as well as TN3 and TN4, suggest that the controlled approach yielded better results, with the mAP@.50IOU and AR100 of the controlled experiments being superior for both tests. However, it is important to note that it is not possible to directly compare both approaches due to various factors. Although the comparison of TN3 and TN4 is more appropriate, as both used the same number of generated patches,

the controlled approach transformed all squamous nuclei to other lesion classes, while TN3 transformed all the cells with a lesion class in another class. These factors could have also contributed to the improved performance of the controlled approach over the non-controlled one. In addition to this, the controlled approach was not better for all the classes, as there was a decrease in the mAP@.50IOU in the ASC-H class.

Comparing the Effectiveness of Conventional Data Augmentations and Generative Approaches

The objective of the tests TN3 and TN4 was to create generated images with the same distribution of the data augmentations used in the original work, in order to compare both approaches. Figure 6.25 illustrates two graphs comparing TN3 to TN4 to the original performance of the model when trained with CDA. The original model without CDA was also included in the graph as a reference. Graph 6.25a shows the mAP@.50IOU for each abnormality class and for each test, while graph 6.25b displays the AR100 in a similar format.

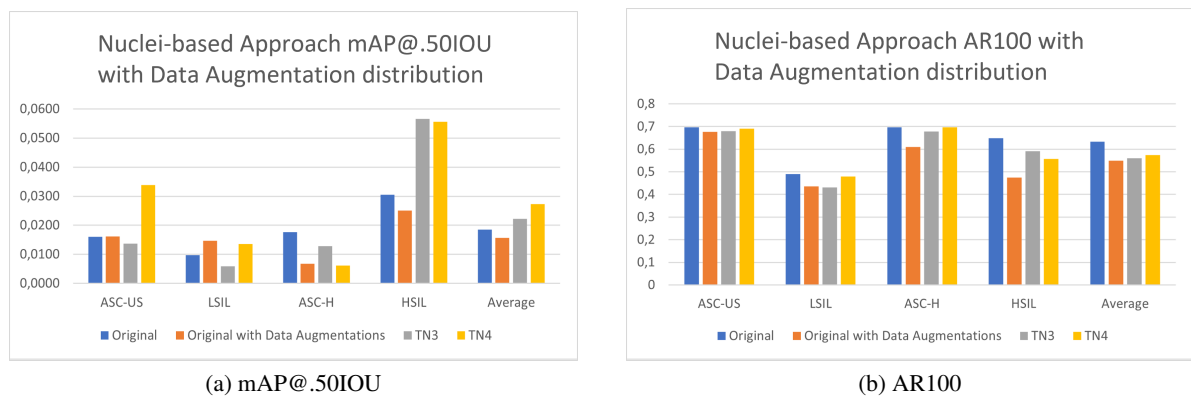


Figure 6.25: Comparison between data augmentation and synthetic images for the Nuclei-based approach. Both graphs present the values for each abnormality class and for each test. Graph (a) presents the mAP@.50IOU values and graph (b) presents the AR100.

For the Nuclei-based lesion detection model, the use of synthetic images resulted in a superior mAP@.50IOU and AR100 when compared to the use of basic image manipulations. It is important to note that when compared to the original model, the CDA approach resulted in worse performance, with lower mAP@.50IOU and AR100 scores. These results further confirm the potential of generative approaches to enhance the performance of models, as they outperformed basic image manipulation methods which have previously been shown to be effective in many studies. The improvements, however, do not apply to all the classes, as the LSIL class for example had better results when using CDA than generative approaches.

Final Evaluation

In this study, the performance of the Nuclei-based detection model was evaluated using a variety of synthetic image generation approaches. The one that yielded the most favorable results was to

increase substantially the volume of the dataset while also controlling the size of the generated images, with an increase on the $mAP@.50IOU$ of 101%, and a decrease of 11% in the AR100. Furthermore, this test had the best results for all classes except for HSIL, for which TN3 had the best performance. This suggests that the baseline model's performance was in fact limited by a lack of instances, as the addition of a substantial number of generated images significantly improved the performance of the Nuclei-based lesion detection model.

It should be noted that the proposed experiments were conducted using the same hyperparameters that were found to be optimal for the original dataset. Thus, it would be beneficial to conduct further research to optimize the hyperparameters for datasets that include generated images, in order to further improve the performance of the model.

In summary, the results of this study confirm that the incorporation of synthetic images can significantly improve the performance of the Nuclei-based lesion detection model and that controlling the distribution of the dataset can also have a positive effect. Additionally, it highlights the importance of considering the performance of individual classes and the ratio of original and synthetic images when incorporating synthetic images into the training dataset.

6.3.3 AI-based Cervical Lesions Detection Algorithms Comparison

This section compares the outcomes of the Region-based and Nuclei-based approaches. Figure 6.26 depicts two graphs that compare the baseline detection models' performance and the highest performance achieved in this study through the use of generated images. As previously described, the best results for both methods were obtained by increasing the dataset volume while controlling the size of the generated images, with tests TR2 and TN2. Graph 6.26a shows a comparison between the baseline [73, 53] and the models proposed in this dissertation regarding $mAP@.50IOU$ metric, and graph 6.26b displays the AR10 metric.



Figure 6.26: Comparison between Region-based approach and Nuclei-based approach. Both graphs present the values for the performance obtained in the original works [73, 53], and the best performance obtained with the generated images. Graph (a) presents the $mAP@.50IOU$ values and graph (b) presents the AR10.

From the analysis of graph 6.26a, it can be deduced that the use of generated images resulted in state-of-the-art outcomes for this dataset. The Nuclei-based approach achieved superior results for all classes when trained with synthetic images, particularly for the ASC-H class. Although the baseline performance of the Nuclei-based approach was higher than that of the Region-based approach, the latter achieved a better mAP@.50IOU for the HSIL class, which is a high-risk category and hence crucial. However, with the use of generated images, TN2 had the best performance across all classes, eliminating the need for multiple models. Despite the fact that the AR10 of the original Nuclei-based approach was slightly better than the one obtained with the generated images, the improvement in mAP@.50IOU was significantly more substantial, demonstrating the potential of the generated images to enhance the performance of cervical screening lesion detection models.

The impact of the generated images varied substantially when used in the different approaches. The superior performance of the Nuclei-based generated images can be somehow related to the superior performance of the original Nuclei-based detection model when compared to the Region-based one. Although the Region-based approach has more information regarding each cell, as more factors are included such as the nuclei-cytoplasm ratio, the high variability of the cells in the same class makes it very hard for the detection model to correctly identify each type of lesion class and for Dreambooth to accurately create cells which belong to a specific class. Other factors can also have influenced the inferior results of the Region-based approach as explained in subsection 6.3.1. Furthermore, although Nuclei-based images showed an improvement in performance, the Region-based Dreambooth model produced higher-quality images as determined by visual inspection by a non-expert observer. To further validate these results and determine the optimal parameters for the Dreambooth and Inpainting models, the involvement of trained cytopathologists is necessary for future studies. It is noteworthy that the parameters used to train the Nuclei-based Dreambooth model were primarily based on the results from the Region-based approach, due to time limitations. Future studies should aim to optimize the Nuclei-based Dreambooth model parameters for further improvement in results.

While the generative approach has shown promise in improving lesion detection algorithms, further refinement is necessary. Future studies should explore new hyperparameters for optimal performance and assess the ideal ratio of real and synthetic images. Additionally, ongoing annotations for the private dataset used in this study regarding the explanation for each finding, i.e. which criteria were more relevant for each annotation (e.g. augmented nuclei or reduced cytoplasm), will provide a justification for why cytopathologists chose a given class, providing a more robust evaluation of the dataset, and making it possible to control synthetic image generation in a more granular way (e.g. controlling the presence of different morphological cellular criteria for each class) to further align with the literature.

Chapter 7

Conclusion and Future work

In this work, we explored the potential of using Latent Diffusion Models to enhance cytological imaging collections for cervical cancer screening. We focused on the use of the open source latent text-to-image diffusion model Stable-Diffusion, which was fine-tuned to generate single cervical cancer cells with different classes of neoplastic changes from an input textual prompt. The model was also used to inpaint multi-cellular images with new abnormality classes.

The generated images were evaluated by two cytopathologists for realism and neoplastic changes and were also used to retrain two distinct cervical cancer lesion screening detection models. The results showed that the synthetic images were highly realistic and retained the important features of the original images, and were even indistinguishable from real cytological images. Furthermore, the use of these synthetic images in combination with real ones resulted in state-of-the-art performance in cervical cancer screening on the dataset used in this research work.

In addition to expanding the overall volume of training data and increasing the number of samples of underrepresented classes, the use of Latent Diffusion Models also allowed to balance the data volume among the different data classes. This is particularly valuable in low-resource settings where high-quality, well-annotated training data can be scarce. The study also highlighted the importance of evaluating synthetic images for realism and neoplastic changes, which is crucial for ensuring the reliability and trustworthiness of the generated images.

Despite the successful results achieved in this research, there are still several areas of improvement that can be explored in future work. One potential avenue for improvement is further experimentation with the ideal number of images generated. An optimal ratio between the synthetic images and the real ones should be studied, to accomplish a more diverse dataset, which does not affect the quality of the dataset. Different models with varying amounts of generated images could be used for each type of abnormality to determine the optimal number of synthetic images needed to achieve the best results. Another area of interest is the use of newer versions of Stable-Diffusion models, which have been developed since this research was conducted and may offer more control over the generation process and better results.

Additionally, incorporating the input of a cytopathologist during the development process could be beneficial in guiding the model to create more accurate images. By having a specialist

evaluate the synthetic images and provide feedback, the model can be fine-tuned to better capture the characteristics of each abnormality class. Another important aspect to consider in future research is the quality of the annotations provided in the datasets used in this study. The subjectivity of cytological images can vary depending on the specialist evaluating the instance, and can even vary for the same specialist at different times. This highlights the importance of thoroughly evaluating the quality of the annotations, as the generated images depend heavily on the original dataset. In the ambit of TAMI project, Fraunhofer research team is in the process of generating additional annotations that will encompass the justification for the cytopathologist's selected lesion classification. This will offer a deeper insight into what are the most important characteristics when generating a cell, being possible to better control certain aspects of the generated cells.

Overall, this research provides a promising foundation for further exploration and development of Latent Diffusion Models for synthetic generation of cytological images in cervical cancer screening. With continued research and improvement, this approach has the potential to significantly impact the development and deployment of accurate and reliable cervical cancer screening systems, ultimately leading to improved patient outcomes.

References

- [1] Stability AI. Stable diffusion 2.0 release. <https://stability.ai/blog/stable-diffusion-v2-release>, 2022. Accessed: 2022-12-15.
- [2] Marwan Ali Albahar. Skin lesion classification using convolutional neural network with novel regularizer. *IEEE Access*, 7:38306–38313, 2019.
- [3] Jill Albritton, Tania Day, Debra S Heller, Claudia Perrera, Gianluigi Radici, Darion Rowan, Maria Angelica Selim, James Scurry, Kathryn Welch, Edward Wilkinson, et al. Journal of lower genital tract diseases-july 2020. *J Low Genit Tract Dis*, 24:317–329, 2020.
- [4] Hazrat Ali, Shafaq Murad, and Zubair Shah. Spot the fake lungs: Generating synthetic medical images using neural diffusion models. *arXiv preprint arXiv:2211.00902*, 2022.
- [5] Ahmed Alrajjal, Vaishali Pansare, Moumita Saha Roy Choudhury, Mir Yousufuddin Ali Khan, and Vinod B Shidham. Squamous intraepithelial lesions (sil: Lsil, hsil, ascus, asc-h, lsil-h) of uterine cervix and bethesda system. *CytoJournal*, 18, 2021.
- [6] STABLE DIFFUSION ART. Beginner’s guide to inpainting. https://stable-diffusion-art.com/inpainting_basics/, 2022. Accessed: 2022-12-01.
- [7] AssemblyAI. Stable diffusion 1 vs 2e. <https://www.assemblyai.com/blog/stable-diffusion-1-vs-2-what-you-need-to-know/>, 2022. Accessed: 2022-12-02.
- [8] Andrew Beers, James Brown, Ken Chang, J Peter Campbell, Susan Ostmo, Michael F Chiang, and Jayashree Kalpathy-Cramer. High-resolution medical image synthesis using progressively grown generative adversarial networks. *arXiv preprint arXiv:1805.03144*, 2018.
- [9] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.
- [10] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing*, 12(8):882–889, 2003.
- [11] BIRME. Bulk image resizing made easy 2.0. <https://www.birme.net/>, 2022. Accessed: 2022-12-03.
- [12] Alceu Bissoto, Eduardo Valle, and Sandra Avila. The six fronts of the generative adversarial networks. *arXiv preprint arXiv:1910.13076*, 2019.

- [13] NB Byju, Vilayil K Sujathan, Patrik Malm, and R Rajesh Kumar. A fast and reliable approach to cell nuclei segmentation in pap stained cervical smears. *CSI transactions on ICT*, 1(4):309–315, 2013.
- [14] Darius Chira, Ilian Haralampiev, Ole Winther, Andrea Dittadi, and Valentin Liévin. Image super-resolution with deep variational autoencoders. *arXiv preprint arXiv:2203.09445*, 2022.
- [15] François Chollet. Generating images with variational autoencoders. <https://gaussian37.github.io/deep-learning-chollet-8-4/>, 2018. Accessed: 2022-10-08.
- [16] CompVis. Stable diffusion v1-4 model card. <https://huggingface.co/CompVis/stable-diffusion-v1-4>, 2022. Accessed: 2022-11-04.
- [17] CompVis. Stable diffusion v1-5 model card. <https://huggingface.co/runwayml/stable-diffusion-v1-5>, 2022. Accessed: 2022-11-04.
- [18] Teresa Conceição, Cristiana Braga, Luís Rosado, and Maria João M Vasconcelos. A review of computational methods for cervical cells segmentation and abnormality classification. *International journal of molecular sciences*, 20(20):5114, 2019.
- [19] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.
- [20] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II. IEEE, 2003.
- [21] Eduardo Luís Pinheiro da Silva. Combining machine learning and deep learning approaches to detect cervical cancer in cytology images. 2021.
- [22] Serviço Nacional de Saúde. Hospital professor doutor fernando fonseca. <https://hff.min-saude.pt/>, 2021. Accessed: 2022-12-23.
- [23] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [24] Hugging Face. Stable diffusion with diffusers. https://huggingface.co/blog/stable_diffusion, note = Accessed: 2022-12-16, 2022.
- [25] Hugging Face. Training stable diffusion with dreambooth using diffusers. <https://huggingface.co/blog/dreambooth>, 2022. Accessed: 2022-12-03.
- [26] Shervan Fekri-Ershad. Pap smear classification using combination of global significant value, texture statistical features and time series features. *Multimedia Tools and Applications*, 78(22):31121–31136, 2019.
- [27] Centers for Disease Control and Prevention. Basic information about cervical cancer. https://www.cdc.gov/cancer/cervical/basic_info/index.htm, 2021. Accessed: 2022-11-15.
- [28] David Foster. *Generative deep learning: teaching machines to paint, write, compose, and play*. O’Reilly Media, 2019.

- [29] Fraunhofer. Computer-aided cervical cancer screening, 2022.
- [30] Fraunhofer. Fraunhofer center for assistive information and communication solutions – aicos. <https://www.aicos.fraunhofer.pt/en/home.html>, 2022. Accessed: 2023-01-15.
- [31] Fraunhofer. Transparent artificial medical intelligence. https://www.aicos.fraunhofer.pt/en/our_work/projects/tami.html, 2022. Accessed: 2023-01-15.
- [32] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [33] Srishti Gautam, Arnav Bhavsar, Anil K Sao, and KK Harinarayan. Cnn based segmentation of nuclei in pap-smear images with selective pre-processing. In *Medical Imaging 2018: Digital Pathology*, volume 10581, pages 246–254. SPIE, 2018.
- [34] Atif A Hashmi, Samreen Naz, Omer Ahmed, Syed Rafay Yaqeen, Muhammad Irfan, Muhammad Ghani Asif, Anwar Kamal, and Naveen Faridi. Comparison of liquid-based cytology and conventional papanicolaou smear for cervical cancer screening: An experience from pakistan. *Cureus*, 12(12), 2020.
- [35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [36] Xianxu Hou, L. Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141, 2016.
- [37] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.
- [38] National Cancer Institute. Hpv and pap testing. <https://www.cancer.gov/types/cervical/pap-hpv-testing-fact-sheet>, 2019. Accessed: 2022-11-09.
- [39] National Cancer institute. Next steps after an abnormal cervical cancer screening test: Understanding hpv and pap test results. <https://www.cancer.gov/types/cervical/understanding-abnormal-hpv-and-pap-test-results>, 2022/03. Accessed: 2022-11-12.
- [40] Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. *Nature inspired Smart Information Systems (NiSIS 2005)*, pages 1–9, 2005.
- [41] JoePenna. The repo formerly known as "dreambooth". <https://github.com/JoePenna/Dreambooth-Stable-Diffusion>, 2022. Accessed: 2022-12-02.
- [42] Yessi Jusman, Siew Cheok Ng, and Noor Azuan Abu Osman. Intelligent screening systems for cervical cancer. *The Scientific World Journal*, 2014, 2014.
- [43] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

- [44] Kitai Kim and Bernard Naylor. *Practical guide to surgical pathology with cytologic correlation: a text and color atlas*. Springer Science & Business Media, 2012.
- [45] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [48] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018.
- [49] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, pages 725–741. Springer, 2020.
- [50] Tim Loossens, Kristof Meers, Niels Vanhasbroeck, Nil Anarat, Stijn Verdonck, and Francis Tuerlinckx. Efficient estimation of bounded gradient-drift diffusion models for affect on cpu and gpu. *Behavior Research Methods*, 54(3):1428–1443, 2022.
- [51] Zhi Lu, Gustavo Carneiro, Andrew P Bradley, Daniela Ushizima, Masoud S Nosrati, Andrea GC Bianchi, Claudia M Carneiro, and Ghassan Hamarneh. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE journal of biomedical and health informatics*, 21(2):441–450, 2016.
- [52] Midjourney. Midjourney. <https://midjourney.com/>, 2022. Accessed: 2022-10-13.
- [53] Vladyslav Mosiichuk, Ana Filipa Sampaio, Luís Rosado, Paula Viana, and Tiago Oliveira. Deep learning for mobile-based cervical cytology: From automated adequacy assessment to lesions detection. "2023 (In Preparation)".
- [54] Vladyslav Mosiichuk, Paula Viana, Tiago Oliveira, and Luís Rosado. Automated adequacy assessment of cervical cytology samples using deep learning. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 156–170. Springer, 2022.
- [55] Ritu Nayar and David C Wilbur. *The Bethesda system for reporting cervical cytology: definitions, criteria, and explanatory notes*. Springer, 2015.
- [56] Kamyar Nazari, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edge-connect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.
- [57] Ryan O'Connor. Introduction to diffusion models for machine learning. <https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/>, note = Accessed: 2022-10-16, 2022.

- [58] American Society of Clinical Oncology. Cervical cancer: Statistics. <https://www.cancer.net/cancer-types/cervical-cancer/statistics>, 2022/01. Accessed: 2022-11-23.
- [59] World Health Organization. Cervical cancer. https://www.who.int/health-topics/cervical-cancer#tab=tab_11, 2021. Accessed: 2022-10-07.
- [60] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.
- [61] Carlos Pereira, Paulo T Silva, Luís Rosado, Luís Mota, and João Martins. The design thinking process in the development of an intelligent microscopic equipment. In *International Conference on Design and Digital Communication*, pages 170–182. Springer, 2021.
- [62] Hady Ahmady Phoulady and Peter R Mouton. A new cervical cytology dataset for nucleus detection and image classification (cervix93) and methods for cervical nucleus detection. *arXiv preprint arXiv:1811.09651*, 2018.
- [63] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.
- [64] Marina E Plissiti, Panagiotis Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, O Krikoni, and Antonia Charchanti. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3144–3148. IEEE, 2018.
- [65] Federico Pollastri, Federico Bolelli, Roberto Paredes, and Costantino Grana. Augmenting data with gans to segment melanoma skin lesions. *Multimedia Tools and Applications*, 79(21):15575–15592, 2020.
- [66] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [67] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [68] Mariana T Rezende, Raniere Silva, Fagner de O Bernardo, Alessandra HG Tobias, Paulo HC Oliveira, Tales M Machado, Caio S Costa, Fatima NS Medeiros, Daniela M Ushizima, Cláudia M Carneiro, et al. Cric searchable image database as a public platform for conventional pap smear cytology data. *Scientific data*, 8(1):1–8, 2021.
- [69] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [70] Luís Rosado, José M Correia Da Costa, Dirk Elias, and Jaime S Cardoso. Mobile-based analysis of malaria-infected thin blood smears: automated species and life cycle stage determination. *Sensors*, 17(10):2167, 2017.

- [71] Luís Rosado, Paulo T Silva, José Faria, João Oliveira, Maria João M Vasconcelos, Dirk Elias, José M Costa, and Jaime S Cardoso. μ smartscope: Towards a fully automated 3d-printed smartphone microscope with motorized stage. In *International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 19–44. Springer, 2017.
- [72] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.
- [73] Ana Filipa Sampaio, Luís Rosado, and Maria João M Vasconcelos. Towards the mobile detection of cervical lesions: a region-based approach for the analysis of microscopic images. *IEEE Access*, 9:152188–152205, 2021.
- [74] ShivamShrirao. Diffusers. <https://github.com/ShivamShrirao/diffusers>, 2022. Accessed: 2022-12-02.
- [75] J. Rafid Siddiqui. Diffusion models made easy. <https://towardsdatascience.com/diffusion-models-made-easy-8414298ce4da>, 2022. Accessed: 2022-12-08.
- [76] Emrick Sinitambirivoutin. An introduction to variational auto encoders (vae). <https://towardsdatascience.com/an-introduction-to-variational-auto-encoders-vae-803ddfb623df>, 2020/07. Accessed: 2022-11-16.
- [77] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [78] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. Spynet: Segmentation prediction and guidance network for image inpainting. *arXiv preprint arXiv:1805.03356*, 2018.
- [79] AI Summer. How diffusion models work: the math from scratch. <https://theaisummer.com/diffusion-models/>, 2022. Accessed: 2022-10-11.
- [80] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022.
- [81] Mai Bui Huynh Thuy and Vinh Truong Hoang. Fusing of deep learning, transfer learning and gan for breast cancer histopathological image classification. In *International Conference on Computer Science, Applied Mathematics and Applications*, pages 255–266. Springer, 2019.
- [82] Arash Vahdat. Improving diffusion models as an alternative to gans, part 1. https://developer.nvidia.com/blog/improving-diffusion-models-as-an-alternative-to-gans-part-1/?ncid=so-link-914978-vt42&=&linkId=100000124819186#cid=nr01_so-link_en-us, 2022. Accessed: 2022-10-14.
- [83] W Patrick Walters and Mark Murcko. Assessing the impact of generative ai on medicinal chemistry. *Nature biotechnology*, 38(2):143–145, 2020.

- [84] Lilian Weng. What are diffusion models? <https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>, 2021. Accessed: 2022-10-16.
- [85] Hakan Wieslander, Gustav Forslid, Ewert Bengtsson, Carolina Wahlby, Jan-Michael Hirsch, Christina Runow Stark, and Sajith Kecheril Sadanandan. Deep convolutional neural networks for detecting cellular changes due to malignancy. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 82–89, 2017.
- [86] Wasswa William, Andrew Ware, Annabella Habinka Basaza-Ejiri, and Johnes Obungoloch. A pap-smear analysis tool (pat) for detection of cervical cancer from pap-smear images. *Biomedical engineering online*, 18(1):1–22, 2019.
- [87] XavierXiao. Dreambooth on stable diffusion. <https://github.com/XavierXiao/Dreambooth-Stable-Diffusion>, 2022. Accessed: 2022-12-02.
- [88] Yawen Xiao, Jun Wu, and Zongli Lin. Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. *Computers in Biology and Medicine*, 135:104540, 2021.
- [89] Yuan Xue, Qianying Zhou, Jiarong Ye, L Rodney Long, Sameer Antani, Carl Cornwell, Zhiyun Xue, and Xiaolei Huang. Synthetic augmentation and feature-based filtering for improved cervical histopathology image classification. In *International conference on medical image computing and computer-assisted intervention*, pages 387–396. Springer, 2019.
- [90] Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. *arXiv preprint arXiv:2212.08228*, 2022.
- [91] Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit. *Multimedia tools and applications*, pages 1–36, 2022.
- [92] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 30:4855–4866, 2021.
- [93] Xiaohui Zhu, Xiaoming Li, Kokhaur Ong, Wenli Zhang, Wencai Li, Longjie Li, David Young, Yongjian Su, Bin Shang, Linggan Peng, et al. Hybrid ai-assistive diagnostic model permits rapid tbs classification of cervical liquid-based thin-layer cell smears. *Nature communications*, 12(1):1–12, 2021.

Appendix A

AI-based Cervical Lesions Detection Algorithms Detailed Results

In this section, the detailed results of the lesion detection models are presented, including the values obtained for each test and class for the mAP@.50IOU and AR metrics.

A.1 Region-based Detection Model

Test	ASC-US	LSIL	ASC-H	HSIL-SCC	Average
TR0	0.008573	0.000744	0.00706	0.03836	0.01368
TR1	0,00687	0,00221	0,01641	0,00708	0,00814
TR2	0,00566	0,00057	0,01887	0,02307	0,01204
TR3	0.00769	0.00021	0.01729	0.02243	0.01191
TR4	0.0097	0.00046	0.00895	0.01483	0.00849

Table A.1: Region-based approach detection model mAP@.50IOU for the different classes and for each test.

Test	ASC-US	LSIL	ASC-H	HSIL-SCC	Average
TR0	0.25754	0.13125	0.27692	0.32973	0.24886
TR1	0,1581	0,05417	0,24615	0,21622	0,16866
TR2	0,21117	0,08333	0,27436	0,3027	0,21789
TR3	0.17877	0.05	0.15641	0.26216	0.16184
TR4	0.22737	0.06458	0.26923	0.25405	0.20381

Table A.2: Region-based approach detection model AR10 for the different classes and for each test.

A.2 Nuclei-based Detection Model

Test	ASC-US	LSIL	ASC-H	HSIL	Average
TN0	0.0160	0.0097	0.0177	0.0305	0.0185
TN1	0.0186	0.0365	0.0214	0.0177	0.0236
TN2	0.0241	0.0171	0.0584	0.0490	0.0372
TN3	0.0137	0.0059	0.0128	0.0566	0.0222
TN4	0.0338	0.0136	0.0061	0.0556	0.0273

Table A.3: Nuclei-based approach detection model mAP@.50IOU for the different classes and for each test.

Test	ASC-US	LSIL	ASC-H	HSIL	Average
TN0	0.697041	0.490164	0.696774	0.648571	0.6331375
TN1	0.671006	0.419672	0.593548	0.62	0.48195775
TN2	0.691716	0.493443	0.683871	0.562857	0.56150325
TN3	0.67929	0.431148	0.677419	0.591429	0.560109
TN4	0.691124	0.478689	0.696774	0.557143	0.57449525

Table A.4: Nuclei-based approach detection model AR100 for the different classes and for each test.

Table A.5 presents the highest mAP@.50IOU for each class obtained by the Nuclei-based approach, regardless of the number of epochs and its corresponding loss.

Class	LSIL	HSIL	ASC-H	ASC-US
mAP@.50IOU	0,036696	0,056647	0,061397	0,033826
Epochs	29000	11000	14000	9000
Loss	0,426268	0,43773	0,485378	0,450803

Table A.5: The highest mAP@.50IOU values for each class independently of the number of epochs and the test. Its also presented the corresponding loss and number of epochs

Test	ASC-H	HSIL	LSIL	ASC-US
TN0	101	315	82	596
TN1	1459	1673	1440	1954
TN2	707	1261	772	1954
TN3	443	542	346	678
TN4	461	658	405	980

Table A.6: Total number of training annotations for the different tests of the Nuclei-based approach.

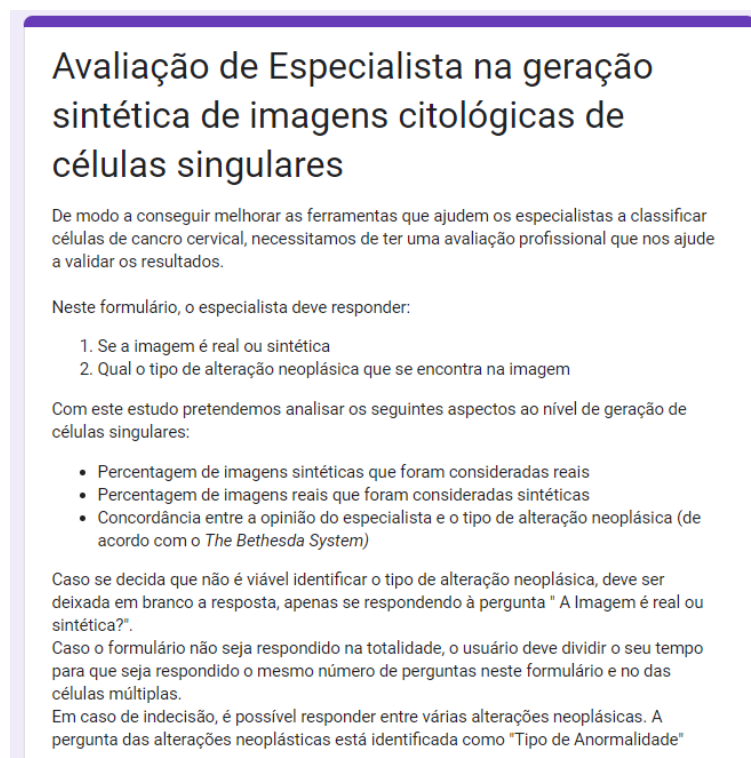
Appendix B

Cytopathologists Questionnaire

In this section, we present the questionnaires utilized to validate the generated images, which were given to cytopathologists. The questionnaires were designed in Portuguese, the specialists' native language. It is also presented the results of the questionnaires for each specialist.

B.1 Single Cell Questionnaire

B.1.1 Screenshots



Avaliação de Especialista na geração sintética de imagens citológicas de células singulares

De modo a conseguir melhorar as ferramentas que ajudem os especialistas a classificar células de cancro cervical, necessitamos de ter uma avaliação profissional que nos ajude a validar os resultados.

Neste formulário, o especialista deve responder:

1. Se a imagem é real ou sintética
2. Qual o tipo de alteração neoplásica que se encontra na imagem

Com este estudo pretendemos analisar os seguintes aspectos ao nível de geração de células singulares:

- Percentagem de imagens sintéticas que foram consideradas reais
- Percentagem de imagens reais que foram consideradas sintéticas
- Concordância entre a opinião do especialista e o tipo de alteração neoplásica (de acordo com o *The Bethesda System*)

Caso se decida que não é viável identificar o tipo de alteração neoplásica, deve ser deixada em branco a resposta, apenas se respondendo à pergunta "A Imagem é real ou sintética?".

Caso o formulário não seja respondido na totalidade, o usuário deve dividir o seu tempo para que seja respondido o mesmo número de perguntas neste formulário e no das células múltiplas.

Em caso de indecisão, é possível responder entre várias alterações neoplásicas. A pergunta das alterações neoplásicas está identificada como "Tipo de Anormalidade"

Figure B.1: Introduction of the Single cell questionnaire.

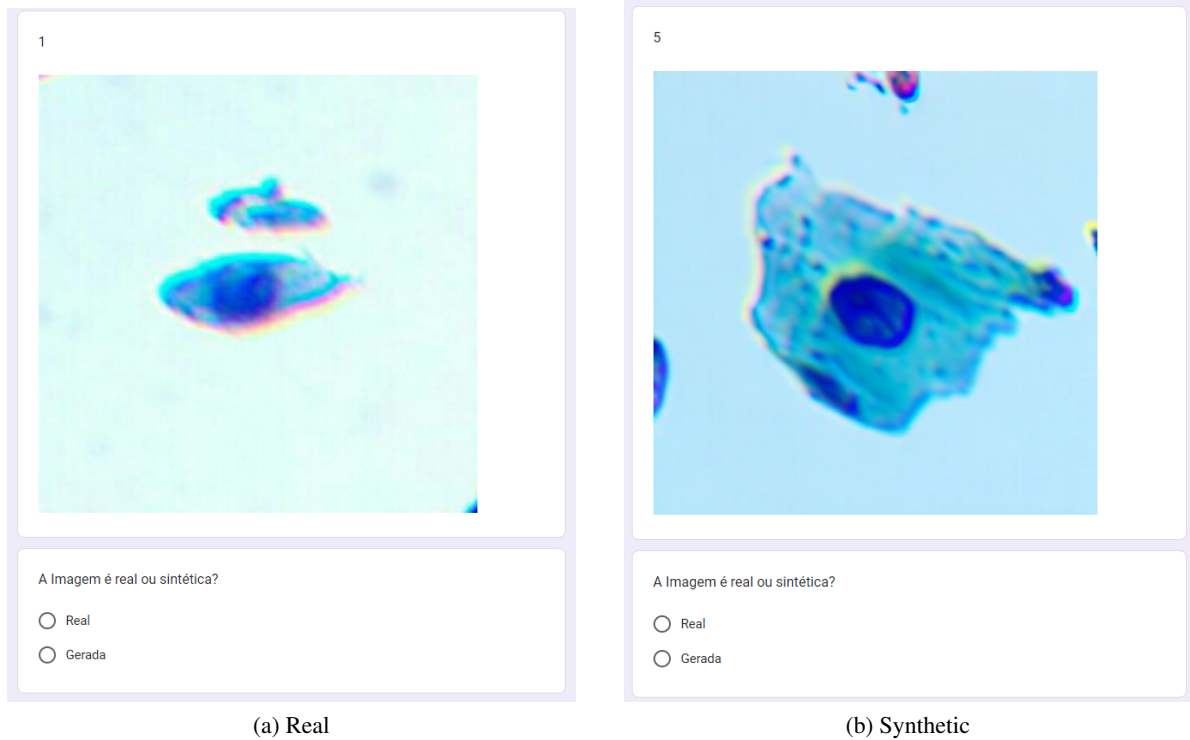


Figure B.2: Example of two questions included in the single cell questionnaire. In (a) its represented a real image, while in (b) its represented a synthetic image.

B.1.2 Results by Specialist

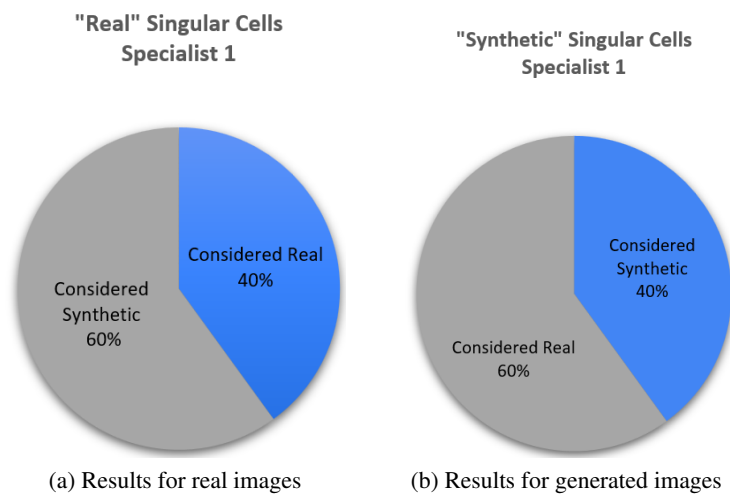


Figure B.3: Questionnaire results of specialist 1 regarding the realism of single cell images. (a) Results regarding the synthetic images. (b) Results regarding real images.

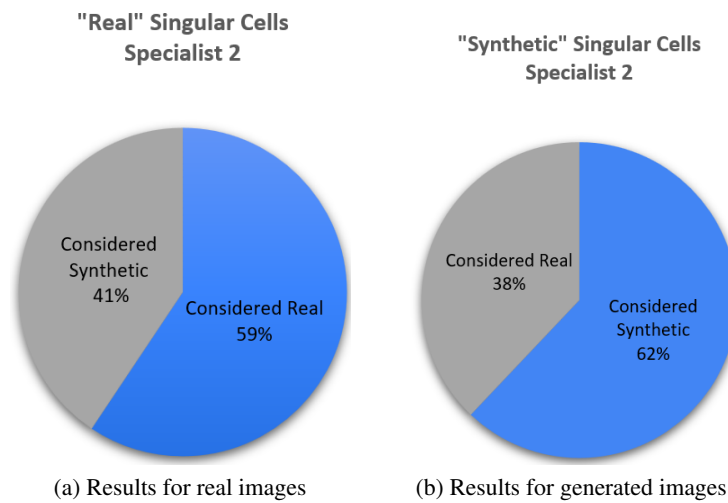


Figure B.4: Questionnaire results of specialist 2 regarding the realism of single cell images. (a) Results regarding the synthetic images. (b) Results regarding real images.

B.2 Multiple Cell Questionnaire

B.2.1 Screenshots

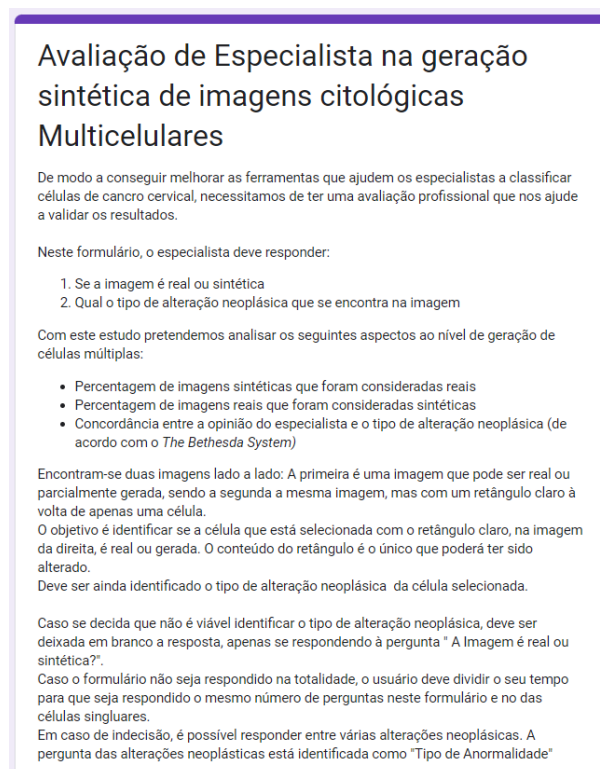
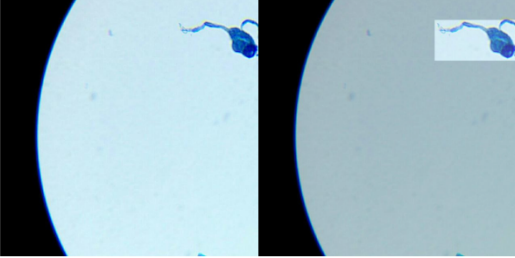


Figure B.5: Introduction of the Multiple cell questionnaire.

1



A Imagem é real ou sintética?

Real

Gerada

Tipo de Anormalidade

Atypical squamous cell of undetermined significance (ASC-US)

Atypical squamous cell, cannot exclude high-grade lesion (ASC-H)

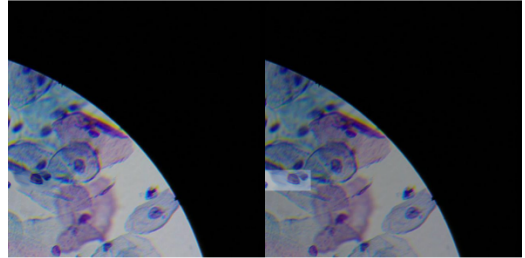
Low-grade squamous intraepithelial lesion (LSIL)

High-grade squamous intraepithelial lesion (HSIL)

Squamous cell carcinoma (SCC)

(a) Real

0



A Imagem é real ou sintética?

Real

Gerada

Tipo de Anormalidade

Atypical squamous cell of undetermined significance (ASC-US)

Atypical squamous cell, cannot exclude high-grade lesion (ASC-H)

Low-grade squamous intraepithelial lesion (LSIL)

High-grade squamous intraepithelial lesion (HSIL)

Squamous cell carcinoma (SCC)

(b) Synthetic

Figure B.6: Example of two questions included in the Multiple cell questionnaire. Each image has two corresponding questions, the first which asks regarding the realism of the image, and the second which asks for the lesion class. In (a) it is represented a real image, while in (b) it is represented a synthetic image.

B.2.2 Results by Specialist

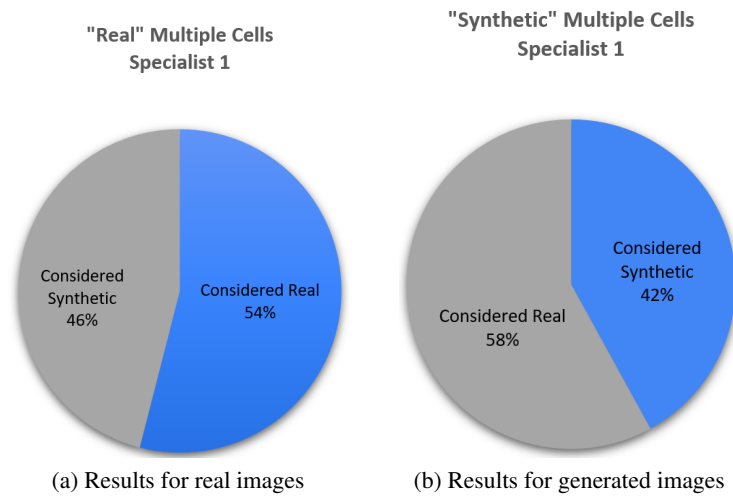


Figure B.7: Questionnaire results of specialist 1 regarding the realism of multiple cell images. (a) Results regarding the synthetic images. (b) Results regarding real images.

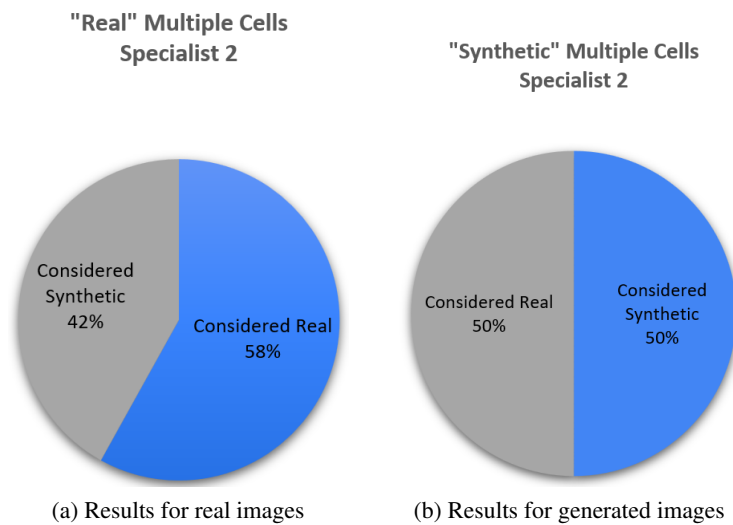


Figure B.8: Questionnaire results of specialist 2 regarding the realism of multiple cell images. (a) Results regarding the synthetic images. (b) Results regarding real images.