FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Model-agnostic generation of example-based explanations for the post-hoc assessment of cervical cytology models

Luís Henrique Condado Marques

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Luís Filipe Teixeira

Co-Supervisor: Luís Filipe Caeiro Margalho Guerra Rosado

July 31, 2022

Model-agnostic generation of example-based explanations for the post-hoc assessment of cervical cytology models

Luís Henrique Condado Marques

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: João Paulo Fernandes
External Examiner: Kelwin Correia
Supervisor: Luís Filipe Teixeira

July 31, 2022

Abstract

The prevalence of Cervical Cancer in women worldwide has motivated the research of effective screening methods which aim to ensure an earlier and more accurate diagnosis of the disease [1]. The development and use of procedures such as liquid-based cytology have resulted in a significant decrease in this cancer's mortality rate [2]. However, the screening process is still a challenging task due to the ambiguity and length of the procedure. In order to tackle these difficulties, there has been an increasing interest in the research and implementation of decision support systems for computer-aided diagnosis[3]. However, the opaque nature of machine learning algorithms does not allow for a full understanding of their decision process, reducing trust in the predicted results. This project aims to introduce explainability to medical decision support systems by providing the users with visual examples that show similar instances to justify the algorithm's prediction, thus increasing transparency and acceptance by medical professionals.

Through a model-agnostic approach, we establish the introduction of example-based explanations in two pre-trained models [4][5] developed in the scope of the CLARE[1] and TAMI[2] projects.

This approach consists of a pipeline, which, through the extraction and comparison of feature maps, produces a set of images that are similar to a given cell. The selection of these images is based on the model's decision process. The parameters for the proposed pipeline are analysed and compared in order to guarantee the best performance for each of the models.

The overall analysis and obtained results prove that the proposed model-agnostic pipeline can be used to generate example-based explanations, and allow us to understand the impact of these explanations during the development phase of medical decision support systems.

Keywords: cervical cancer, computer-aided diagnosis, explainable AI, machine learning

[1] 'CLARE', Fraunhofer Portugal AICOS, accessed 5 July 2022, https://www.aicos.fraunhofer.pt/en/our_work/projects/clare.html

[2] 'TAMI', Fraunhofer Portugal AICOS, accessed 5 July 2022, https://www.aicos.fraunhofer.pt/en/our_work/projects/tami.html

Resumo

A prevalência a nível mundial de cancro cervical em mulheres tem motivado a pesquisa de mecanismo de deteção eficazes, que garantam um diagnóstico mais atempado e preciso da doenca [1]. O desenvolvimento e uso de métodos como citologia em meio líquido tem resultado numa diminuição significativa da taxa de mortalidade do cancro [2]. Contudo, o processo de deteção ainda apresenta desafios devido à sua ambiguidade e duração. Com o objectivo de ultrapassar estas dificuldades, tem havido um aumento no interesse por investigação e implementação de sistemas de suporte de diagnóstico assistido por computador [3]. Contudo a natureza opaca dos modelos de machine learning não permite um entendimento completo das suas decisões, o que reduz a confiança nos resultados previstos. Este projecto tem como objectivo a introdução de explicabilidade aos sistemas de decisão médica através da fornecimento aos utilizadores de exemplos visuais semelhantes, de modo a justificar a previsão do algoritmo, aumentando a transparência e a aprovação por parte dos profissionais de saúde. Através de uma metodologia agnóstica ao modelo, estabelecemos a introdução de explicações baseadas em exemplos em dois modelos pré-treinados [4][5], desenvolvidos no âmbito dos projetos CLARE[1] e TAMI[2]. Esta abordagem consiste numa pipeline, que através de extração e comparação de feature maps, produz um conjunto de imagens que são semelhantes a uma dada célula. A seleção destas imagens é baseada no processo de decisão do modelo. Os parâmetros para a pipeline proposta são analisados e comparados com o objectivo de garantir a melhor performance para cada um dos modelos. A análise geral e resultados obtidos provam que a pipeline proposta pode ser utilizada para gerar explicações baseadas em exemplos, e permite compreender o impacto destas explicações durante a fase de desenvolvimento de sistemas de suporte à decisão médica.

Keywords: cancro cervical, diagnóstico assistido por computador, inteligência artificial explicável, machine learning

Acknowledgements

I would like to thank Fraunhofer Portugal for the opportunity to undertake this project. In particular, Luís Rosado and Ana Sampaio for the constant assistance throughout these last months. I would also like to thank Professor Luís Teixeira for all the help he provided in the development of this project.
To my friends, for all the motivation and constant exchange of ideas.
To my girlfriend Natália for her support and medical knowledge.
And to my parents, siblings and grandparents, for the patience and support they have shown throughout all my academic journey.

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)

Luís Henrique Condado Marques

"There is nothing like looking, if you want to find something."

J R R Tolkien

Contents

List of Figures

List of Tables

Abbreviations

ACA	Average Class Accuracy
AGC	Atypica Glandular Cells
ASC	Atypical Squamous Cells
AUC	Area Under Curve
CNN	Convolutional Neural Network
FN	False Negative
HPV	Human Papillomavirus
IPO	Instituto Português de Oncologia
LBC	Liquid-Based Cytology
LIME	Local Interpretable Model-Agnostic Explanations
ML	Machine Learning
PCA	Principal component analysis
SHAP	Shapley Additive Explanations
SSD	Single-Shot Detectors
SVM	Support Vector Machine
TBS	The Bethesda System
TCGFE	Texture-Color-Geometry Feature Extraction
VAE	Variational autoencoder
VIA	Visual Inspection with Acetic Acid
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Context and Motivation

According to the data provided by the World Health Organization (WHO), cervical cancer presents itself as the fourth most common cancer in women worldwide. It is estimated that more than 570,000 women are diagnosed with this type of cancer every year, and more than 311,000 fall victim to the disease [1].

The detection and screening of this type of cancer are particularly challenging due to the complexity and ambiguity of the process. Different methods have been recommended by the WHO, but all of them rely on the decision of a medical professional and, as such, are susceptible to the limitations of the environment in which the test is performed, such as the availability of specific resources or the proficiency of the medical professional performing the screening.

The most common of these screening methods is the conventional Papanicolaou (Pap) smear, which consists on the collection of cells at the outer opening of the cervix, which are later analysed under the microscope to detect the presence of cervical lesions. The appearance of this particular screening method had a very significant impact on the cancer's incidence, reducing it by 60% to 90% and reducing the mortality rate by 90% [7].

Additionally, in the last years, the appearance of Liquid-Based Cytology has contributed to the increase of the sample quality and has allowed for a faster and more accurate screening process, but even with the general improvements made to the screening process, it remains a difficult and time-consuming task.

This complication has prompted the research and development of decision support systems for computer-aided diagnosis. The goal of these systems is to employ Machine Learning models to facilitate the prediction and detection of cervical lesions in cells obtained during methods such as the Pap Smear. However, when applying the model, the medical professional does not receive information regarding the decision process, consequently reducing the trust in the provided results.

To reduce the effects of this problem, some methods of Explainable AI (XAI) have been proposed which aim to introduce explanability to Machine Learning models by providing insight into the model's decision process.

1.2 Objectives

This project aims to implement and analyze the introduction of example-based explainability to medical decision support systems by ensuring that when using these systems, the medical professional has access to the model's final prediction as well as to visual examples that show similar instances to justify that exact prediction. These visual instances allow the user a greater comprehension of the model's decision process, thus increasing the trust in the result.
In order to guarantee the desired example-based explanations, a novel model-agnostic approach is proposed. It consists of a pipeline composed of 4 separate blocks: i) Feature map extraction; ii) Dimensionality reduction; iii) Feature map similarity; and iv) Interpretable features similarity filtering. Thus, the proposed approach aims to bring scientific contributions to the field of example-based XAI systems, with a focus on the cervical cytology use case. In particular, we envision that this work will support further research regarding essential aspects of these type of systems, such as the impact of example-based explanations during the development phase of medical decision support systems (e.g. identifying system limitations), or the potential to improve the interaction between AI systems and the medical professionals (e.g. increase user trust and acceptance of AI system via example-based explanations).

1.3 Document Structure

This document is divided into eight chapters. Chapter 1 consists of the Introduction and the description of the project's objectives. Chapter 2 provides some context regarding Cervical Cancer and its screening methods. Chapter 3 consists of a theoretical overview of Explainable AI methods. Chapter 4 consists of the State-of-the-art review of Cervical Cancer screening, the use of Explainable AI in diagnosis, and the use of Example-based methods. Chapter 5 provides a description of pre-trained models which will be used in this work and corresponding datasets. Chapter 6 provides an in-depth description of the proposed methodology. Chapter 7 consists of an analysis and exposition of the obtained results. And finally, Chapter 8 consists of some final conclusions and indications for possible future works.
The document also contains some appendices containin the various model architectures, loss plots, and relevant images, all of which are referenced throughout the various chapters.

Chapter 2

Background Concepts of Cervical Cancer

2.1 Cervical Cancer Disease Characterization

Cervical cancer is defined as the abnormal growth of malignant cells in the cervix's lining. Most of the cases of Cervical Cancer originate from an infection by the Human papillomavirus (HPV), which can be transmitted either by sexual activities or by skin-to-skin contact of the genital areas [8]. Around 90% of these infections are short-lived, usually resolving spontaneously after a few months. However, there is a risk that the infection may become chronic and cause pre-cancerous lesions that later progress to invasive cervical cancer. This process usually takes 10 to 20 years [9], and as such, early cervical lesions might not show any signs or symptoms, which prompts the necessity for regular screening of precancer lesions. The cure rate for invasive cervical cancer is closely related to the stage of disease when the diagnosis occurs.
When the lesions progress to invasive cancer, they might cause symptoms such as vaginal bleeding, unusual vaginal discharge, pelvic pain, or pain during sexual intercourse [10]. If left untreated, Cervical Cancer is almost always fatal [9].

2.2 Cervical Cancer Screening Characterization

Given the vital importance of the screening process in the reduction of the cancer's mortality rate, the WHO has proposed several different methods, which can be divided into three different types.

2.2.1 HPV Testing

The HPV Testing aims to detect the presence of the HPV in the patient's cells. The test is done by collecting cells from the woman's cervix and searching for traces of the virus's DNA [11].
It is important to note that the presence of HPV in the cells does not mean that the cells contain cancerous or precancerous lesions. However, it is the most accurate way of finding people at risk

of developing cervical cancer. If a given patient tests positive, some follow-up visits would be necessary, along with more tests to look for a precancer or cancer. [7].

2.2.2 Visual inspection with acetic acid

The Visual inspection with acetic acid (VIA) procedure allows the medical professionals to directly visualize cervical lesions by applying acetic acid to the lining of the cervix. In the presence of acetic acid, the damaged tissues will become white after a short amount of time, while the normal cervical tissues remain unchanged [8]. These lesions usually appear in an area of the cervix called the squamocolumnar junction. The position of this zone changes through age and, as such, is difficult to identify in postmenopausal women, which means that WHO does not recommend the use of this method in older women, as it provides poor performance. [12].
Additionally, this method has a limited accuracy when detecting pre-cancerous lesions [13].

2.2.3 Cervical Cytology

Cytology has been the standard screening method in the last decades, and its appearance has contributed to a drastic reduction in the cancer's incidence, reducing it by 60% to 90% and reducing the mortality rate by 90% [7].
There are two different methods of cervical cytology currently being employed, (i) the conventional Papanicolau Smear (Pap Smear) and (ii) Liquid Based Cytology (LBC).
In general, LBC provides better quality samples by allowing a much clearer and consistent spread of smears, reducing the number of unsatisfactory cases, and allowing the screening to take much less time [14]. However, it also involves aggravated economic costs. For this reason, LBC is more prevalent in developed countries while the Pap Smear is most commonly used in less developed countries where LBC is not affordable [15].
It becomes clear that the choice of screening method heavily depends on the setting and context in which the tests are performed and must be adapted to each patient's particular situation.

2.3 The Bethesda System

There are several classification systems available for cervical cancer and precancer. These systems aim to classify the different lesions for various different reporting and clinical purposes, such as screening, diagnosis, and risk assessment [3]
TBS is a classification system for reporting the results of Conventional Pap Smears and Liquid-Based Cytology tests. This system reports both sample adequacy, and cytology results [16].
Even though none of the models used in this work employ the TBS as classification system, it is the standard system for cervical cytology and should be the object of a closer analysis. Also, the used nuclei detection model used the sample adequacy standards of the TBS as reference.

2.3.1 Sample adequacy

Specimen quality assessment is an essential step in ensuring that the results provided are reliable, as such, TBS takes into account the quality of the sample when performing the screening by establishing objective guidelines and minimum criteria [17], that must be followed to ensure the validity of the results:

- **Cellularity**: The prepared samples should have a minimum amount of visible cells. A Liquid-Based preparation should have an estimated minimum amount of at least 5000 well visualized and well-preserved cells. In the case of a conventional Pap-Smear, the preparation should contain an estimated minimum of 8000 - 12000 cells [17].
- **Obscuring Factors**: Excessive presence of blood, inflammation, bacteria, or lubricant can hinder the interpretation of certain specimens, not allowing for correct visualization of the respective cells [17].
- **Evidence of Transformation Zone**: It is recommended that the samples contain 10 well-preserved endocervical cells or squamous metaplastic cells. This is an optional criterion, but it is often included in reports [17].

2.3.2 Cytology Results

If the sample is considered adequate, then a cytological result can be inferred. If the analyzed cells show any malignant lesions or abnormalities, an abnormal result will be reported. Specific findings such as benign infections or inflammations are not considered abnormal and are reported as a normal result [9].
Abnormal cells can either be atypical squamous cells (ASC) or atypical glandular cells (AGC), which means that a different grading system will be required for each case.

2.3.2.1 Atypical Squamous Cells

In terms of ASC, we can classify these cells according to 3 essential features: squamous differentiation, increased nuclear to cytoplasmic ratio, and minimal nuclear changes. This classification can be done in the following types:

- **Atypical squamous cells of undetermined significance (ASC-US):** ASC-US refers to changes that are suggestive of LSIL. Nuclei are approximately two and one half to three times the area of the nucleus of a typical intermediate squamous cell or twice the size of a squamous metaplastic cell nucleus [17].
- **Atypical squamous cells, cannot exclude a high-grade squamous intraepithelial lesion (ASC-H):** ASC-H is a designation reserved for the minority of ASC cases (representing less than 10% of all ASC interpretations) in which the cytologic changes are suggestive of HSIL [17].

- **Low-grade squamous intraepithelial lesions (LSIL):** LSIL is a designation for cells that usually occur singly, in clusters, and sheets. Overall cell size is large, with reasonably abundant "mature" well-defined cytoplasm. They also show nuclear enlargement more than three times the area of normal intermediate nuclei [17].

- **High-grade squamous intraepithelial lesions (HSIL):** HSIL is a designation for smaller cells that show less cytoplasmic maturity than cells of LSIL. Usually, these cells occur singly, in sheets, or syncytial-like aggregates. While overall cell size is variable, in general, the cells of HSIL are smaller than those of LSIL. Higher-grade lesions often contain quite small basal-type cells.[17] These cells have a high likelihood of progressing to cancer if not treated. [9].

- **Squamous cell carcinoma (SCC):** SCC are the most prevalent malignant neoplasm of the cervix, being defined as an invasive epithelial tumor composed of squamous cells of varying degrees of differentiation [18]

2.3.2.2 Atypical Glandular Cells

The classification for AGC labels each cell according to their site of origin, dividing them into four different types:

- **Atypical endocervical cells, not otherwise specified (AGC-NOS):** AGC-NOS cells occur in sheets and strips with some cell crowding, nuclear overlap, and pseudo-stratification. They can show nuclear enlargement up to three to five times the area of normal endocervical nuclei. The cytoplasm may be relatively abundant, but the nuclear to cytoplasmic ratio is increased [17].

- **Atypical glandular cells, favor neoplastic:** In these cells, the cell morphology, either quantitatively or qualitatively, falls just short of an interpretation of endocervical adenocarcinoma in situ or invasive adenocarcinoma. Abnormal cells occur in sheets and strips with nuclear crowding, overlap, and pseudo-stratification. Nuclei are enlarged and often elongated with some hyperchromasia. It is also noticeable an increase in the nuclear to cytoplasmic ratios [17].

- **Endocervical adenocarcinoma in situ (AIS):** A noninvasive high-grade endocervical glandular lesion that is characterized by nuclear enlargement, hyperchromasia, chromatin abnormality, pseudostratification, and mitotic activity [17].

- **Adenocarcinoma (invasor):** In this case, the cytologic criteria match those identified for AIS but may contemplate additional signs of invasion.

Chapter 3

Explainable AI

3.1 Overview

Explainable AI encompasses all the AI methods and approaches in which the results of the solution can be understood by humans. This notion emerged as a contrast to the more opaque and black-box nature of regular AI models. These methods and algorithms allow for a better understanding of the decision process of traditional ML models, which lead to advantages both for the developer and the end-user.

3.2 Explainability

The notion of XAI is deeply based on the concept of explainability. This term has been the object of much research in the last years as academics aimed to find a determinate definition [19], and although no definitive answer was found, several scholars have proposed ways to define it. In [20] for instance, explainability is defined as *"the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model"*.
However, in practical terms, explainability is harder to define and achieve, as it is difficult to identify the necessary features for a result to be considered an explanation. In [21], T. Miller reviews relevant works from philosophy, cognitive psychology, and social psychology to identify what can be considered a 'good' explanation. The result was the delineation of different sets of attributes and objectives which try to define the construct of 'explainability':

- **Explanations are contrastive** — "They are sought in response to particular counterfactual cases. People do not ask why event P happened, but rather why event P happened instead of some event Q." [21]
- **Explanation are selected (in a biased manner**) — "People rarely, if ever, expect an explanation that consists of an actual and complete cause of an event. Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation. However, this selection is influenced by certain cognitive biases." [21]

- **Probabilities probably don't matter** — "While truth and likelihood are important in explanation and probabilities really do matter, referring to probabilities or statistical relationships in explanation is not as effective as referring to causes. The most likely explanation is not always the best explanation for a person, and importantly, using statistical generalisations to explain why events occur is unsatisfying, unless accompanied by an underlying causal explanation for the generalisation itself" [21].

- **Explanations are social** — "Explanations are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs." [21]

Miller finished this explanation by concluding that these four points all converge around a single point: *"explanations are not just the presentation of associations and causes, they are contextual"*. This is particularly relevant for it shows that the attributes that define each explanation are dependent on the context on which its presented.

3.3 Main Advantages

The addition of explainability in regular ML models brings general advantages to the model and the users. The most noticeable and arguably the most important of these advantages is the increase in trust [22]. In [23], the authors identify trust as the primary way to increase a user's confidence in the system's decision process and predictions.
This increase in trust can happen as a consequence of other advantages, such as the presence of causality. In [24] the author presents the view that causal attribution is human nature, and thus a system that provides casual explanation is perceived more human-like by end-users.
This concept is particularly relevant for this work, as the increase of confidence is the basis for all other proposed objectives, and it is also valuable for ML models of all areas by serving as a tool for tasks such as:

- **Ensuring safety measures:** In a situation where there is risk for the user, explanations can be used to assure the users that the environment is being correctly interpreted [25].

- **Increasing social acceptance:** As seen before, explanations are a social progress. Given that people attribute beliefs, desires, intentions and so on to objects, that means that a machine or algorithm that explains its predictions will find more acceptance [25].

- **Debugging:** The ability to interpret is valuable, both in the research and development phase as well as after deployment. An interpretation for an erroneous prediction helps to understand the cause of the error, and it delivers a direction for how to fix the system.

3.4 Explanation Methods Classification

There have been several attempts to categorize the various explanation methods according to different criteria.

3.4.1 Intrinsic vs Post-hoc

The models can be classified into intrinsic and post-hoc methods based on whether explainability is achieved through constraints imposed directly to the model (intrinsic) or by applying methods that analyse the model after training (post-hoc) [26].

3.4.2 Model-specific vs Model-agnostic

They can also be classified into model-specific and model-agnostic approaches:
The **model-specific** approach consists of the use of inference models that are limited to specific model classes.
In a practical context, a **model-specific** approach has to ensure that the chosen model is interpretable, as the interpretation of intrinsically interpretable models is always model-specific [25]. This interpretability depends solely on being able to 'interpret' a single instance in the dataset. [25], which then allows for the representation of the corresponding data in a humanly understandable way.
The **model-agnostic** approaches on the other hand, can be used on any machine learning model and are always applied after the model has been trained, making the models intrinsically post-hoc. This type of approach usually works by analyzing the model's feature spaces [25].

3.4.3 Technical classification

It is also possible to classify each explanation method based on the techniques used. Such classification was proposed in by Arya et al. in [26], where the most popular classes of explainability methods are categorized as follows:

- **Saliency Methods:** Explanation methods that highlight different portions in an image to understand the classification tasks. This category includes the popular Local Interpretable Model-Agnostic Explanations (LIME) [27] and SHapley Additive exPlanations (SHAP) [28] methods.
- **Neural Network Visualization Methods:** Explanation methods that allow the visualization of intermediate representations/layers of a neural network.
- **Feature Relevance Methods:** Explanation methods that use partial dependence plots and sensitivity analysis to study the effects of different input features on the output value.
- **Example-based Methods:** Explanation methods that explain the predictions of test instances based on similar or influential training instances.

- **Knowledge Distillation Methods:** Explanation methods that learn a simpler model based on a complex model's predictions, and explain that simpler model.
- **High-Level Feature Learning Methods:** Explanation methods that learn a high level of interpretable features in an unsupervised manner through variational autoencoder or generative adversarial network frameworks.
- **Methods that Provide Rationales:** Explanation methods used mostly in the natural language processing and computer vision domains that generates rationales/explanations derived from input text.
- **Restricted Neural Network Architectures:** Explanation methods that propose certain restrictions on the neural network architecture to make it interpretable

Chapter 4

State-of-the-art Review

4.1 Cervical Cancer Detection

The purpose of this work is to dynamize and facilitate the diagnosis of cervical cancer, but the model agnostic approaches used to ensure the desired explainability are meant to be applied to existing models for the detection of cervical lesions. As such, the existing methodologies must be considered and reviewed.

4.1.1 Cell Segmentation

In order to accurately detect and later classify the cells contained in the cytological images, and as many identifying features of cancerous cells are morphological, it is crucial that a precise **cell segmentation** is performed. This is one of the first necessary tasks in any Cervical Cancer screening algorithm and has been the subject of much research, ranging from traditional image segmentation techniques to more complex ML models.

For single-cell instances, simple algorithms have been applied with much success, as in [29] where the proposed approach uses a customized Laplacian of Gaussian (LoG) filter to segment free lying cell nuclei, or in [30] which uses a simples histogram thresholding method. However, these methods do not obtain satisfactory results when applied to images obtained in Pap Smears, and as such, some more complex methods have been proposed. In [31] the authors use three watershed passes to segment both nucleus and cytoplasm from large cell masses of overlapping cervical cells. In [32] a grid-based approach is proposed where the model first approximates the boundaries using a similarity metric and later refines them by reducing concavity, iterative smoothing, and outlier removal. In [33] a level set method is proposed that uses a joint optimization of multiple level set functions, where each function represents a cell within a clump that has both intracell and intercell constraints.

Some more complex ML approaches have also been proposed, such as [34] which consists of a two-stage approach. The first stage consists of using a Support Vector Machine (SVM) to segment the image into nuclei, cellular clusters, and background. In the second stage, a robust shape deformation framework constructs the cytoplasmic shape of each overlapping cell.

All these proposed methods have achieved good results in the segmentation task, even for overlapping cells.

4.1.2 Cell Classification

After the cell segmentation is done, the next step is to classify the cell according to its abnormality level. This classification can be done according to different classification systems, but as mentioned, the TBS is currently the standard system for cervical cytology. However, the literature review includes models using other systems, mostly due to the limitations imposed by the public datased.

The classification methods proposed have steadily achieved good performances, with many achieving accuracy results nearing the 99% mark. In [35], Zhang et al. propose an approach that uses a CNN pre-trained on the ImageNet dataset [36], followed by the application of transfer learning. This model achieved an accuracy of 98.3 % and AUC of 0.99. In [37] Gautam et al. propose two different approaches: (i) The use of patch-based CNNs and pre-trained CNNs using the AlexNet architecture, and (ii) Decision-tree based classification using transfer learning. This work relates the cell classification in both 2 and 7 classes and reports accuracy results of 99.3% and 93.75%, respectively. In [38] Jith et al. also use Deep Learning to propose a method that aims to overcome the necessity for an accurate segmentation by directly classifying the image patches with cells, using a CNN based on the AlexNet architecture. The model obtained an accuracy result of 99.6%. The attempts to reduce the importance of the segmentation step is also noticeable in [39], where Zhao et al. propose a solution where the image is divided into blocks instead of segmented cells and is later classified using a Support Vector Machine achieving an accuracy result of 98.98%

A hybrid approach was also proposed by Zhang et al. [40] where a CNN is used for feature extraction and an SVM for the classification task, reporting an accuracy of 99.3%.

4.2 Explainable AI in Diagnosis

As previously mentioned, the complex and black-box nature of Machine Learning models contributes to the reduction of trust in the provided results. In models used for Medical purposes, this decrease of trust has a significantly higher impact. In [41], the author goes so far as to classify this black-box nature as dangerous when deployed in clinical applications because they may not provide reliable justification to the medical experts. As such, it is crucial to ensure that the results provided by each model are accompanied by valid explanations. This issue has prompted the integration of explainability in Machine Learning models used for the diagnosis of the most varied diseases.

In [42], the authors proposed an interpretable model for Alzheimer's diagnosis and progression through the development of a two-layer model with Random Forest as a classifier algorithm. The first layer consists of a multi-class classification for the early diagnosis of Alzheimer's patients. The second layer performs a binary classification to detect possible progression within three years from a baseline diagnosis. Besides the use of the SHAP framework to ensure explainability, the

model also implements 22 explainers based on decision trees to provide complementary justifications for every Random Forest decision in each layer. In [43] a model is proposed that aims to enhance the interpretation of COVID diagnosis through X-ray scans. The model uses SqueezeNet [44] to recognize Covid-19 in regular lung images and uses both SHAP and LIME to achieve the desired explainability. In [45] the authors use topological features to detect Glioblastomas (GBM) and use the only LIME method to ensure explainability.
However, some more complex methods are also used. In [46] the author's work focuses on the automatic detection of Parkinson's and proposes an explainable approach based on Genetic Programming, called Grammar Evolution. A context-free grammar is used to describe the language of the programs to be generated and evolved. These correspond to the explicit classification rules for the diagnosis of the subjects. In [47] it is presented an XAI framework for breast cancer therapies. This framework is based on an adaptive dimensional reduction and outlines the most important clinical features for oncological patient profiling, and based on these features determines the profile of each patient. In [48] the authors present a computer-aided framework for allergy diagnosis. Several algorithms were implemented and then cross-validated to choose the optimal trained model for multi-label classification. The transparency of the machine learning models was ensured using post-hoc approaches for the extraction of interpretable rules represented by the random forest method.

4.3 Example-based Explainable AI

The advantages provided by the introduction of explainability in diagnosis are undeniable, however due to the sensitive nature of the diagnostic procedures it is important to establish which types of explanations are more valuable to the diagnostic process. In [49] the authors provide a in-depth analysis of the use of XAI in digital pathology. Through interviews with certified pathologists they conclude that the introduction of explainability in these models should be made using an "parallel approach", which consists of providing elements based on components of the AI decision-making process that are identified as most closely matching the users' decision-making processes. As an example, the authors mention providing the user with "samples from the training data that are most influential, or at least most similar, to a given model outcome". In practical terms this corresponds to an example-based approach, which aims to mimic the process of referring to textbook cases. In [50] the author also concludes that the use of example-based explanations improves the trust in the predicted results, however it can reveal certain limitations of the algorithm.
In terms of practical applications, literature shows that the use of example-based methods is steadily increasing [51]. In the medical field, it has not yet reached its full potential. But even though there have been few studies, the literature shows that the use of these methods brings genuine improvements, as seen in [52] where the authors analyze several explanation methods to aid in the diagnosis of Heart Diseases by comparing example-based methods such as Anchors, Counterfactuals, Counterfactuals guided by Prototypes, Contrastive Explanation Method among others. In [53] an approach is used in which the authors ensure explainability by providing the user with

adaptive reference cells, obtained through the comparison of Feature Maps extracted from models used for the classification of Blood Cells. A more complex solution is also presented in [54] where the author proposes a counterfactual multi-granularity graph supporting facts extraction for lymphedema diagnosis. First, the model structures the sequence of Electronic Medical Records into a hierarchical graph network and then obtains the causal relationship between multi-granularity features and diagnosis results through counterfactual intervention on the graph.
The use of Example-based XAI stretches also to other various fields of study. In [55] Gade et al. analyses the theoretical impact of XAI and Example-based methods in Industrial environments, while in [56] the authors used Example-based explanations to increase transparency and trust in a ChatBot system. These advantages spread even to areas such as agriculture, as seen in [57] where the authors use example-based pos-hoc approaches for Predicting Grass Growth for Sustainable Dairy Farming.

Chapter 5

Detection and Classification Models

This project aims to ensure a model-agnostic approach to provide Example-based explanations of the predictions of a given model. To verify this, in order to accurately validate the proposed pipeline, two different models will be used. Even though both models are applied to cervical cytology, they have different architectures, use different datasets and are used to detect and classify different objects within an image.
The use of different models for the testing and validation of the pipeline ensures that the proposed model agnosticism is fulfilled in accordance with the main objectives. Both models were developed and provided by Fraunhofer Portugal AICOS.

5.1 Region Detection Model

The Region Detection model [4] was developed by Ana Sampaio, Luís Rosado and Maria Vasconcelos in the scope of the CLARE[1] project. The model was created to serve as a basis for an automated solution for the detection and classification of cervical lesions in images acquired by a mobile device. It consists of a Single Shot Detector [58] with a MobileNetV2 [59] backbone for convolutional feature map extraction and it detects the position of individual cells within the image, while also classifying them according to their type.
A visual representation of the model's backbone architecture can be seen in Figure A.1 in the appendices.

5.1.1 Dataset

The dataset used to train this model was the public SIPaKMeD dataset [60]. It consists of 966 cluster cell images of Pap smear slides, obtained by a CCD camera adapted to an optical microscope. An example of these images can be seen in Figure 5.1.

The dataset also includes annotations made by experts which defined the contours for each cell's nuclei and cytoplasm, as well as a classification for each of the cells according to their type, using the following categories: Superficial-Intermediate, Parabasal, Koilocytotic, Dysketarotic and

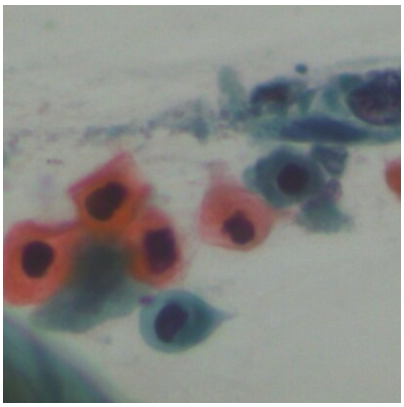

Figure 5.1: Example of an uncropped image from the SIPaKMeD dataset

Metaplastic. These categories are exemplified in Figure 5.2.

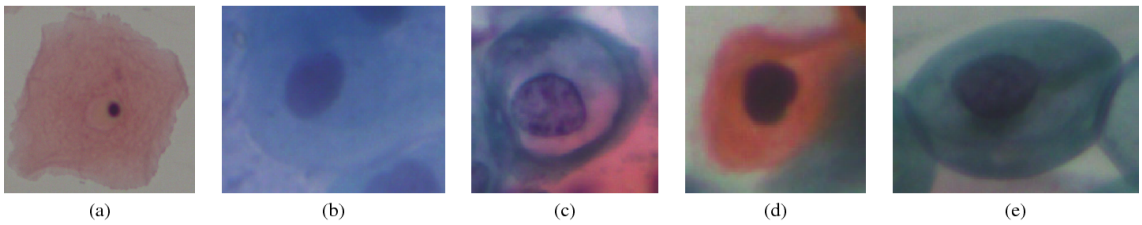

Figure 5.2: Examples of the five categories: (a) Superficial-Intermediate; (b) Parabasal; (c) Koilocytotic; (d) Dysketarotic; (e) Metaplastic

The full distribution of cells for each class can be seen in Table 5.1, as well as the number of images.

Table 5.1: Distribution of cells in the SIPaKMeD dataset

Category	Number of images	Number of cells
Superficial-Intermediate	126	813
Parabasal	108	787
Koilocytotic	238	825
Metaplastic	271	793
Dyskeratotic	223	813
Total	966	4049

As it is necessary to ensure that the images provided to the detection network have a fixed dimension, and due to the size of the original images in the dataset, the images need to be cropped into smaller 320x320 patches, taking into account the annotated region. This process is explained in full in [4]

5.1.2 Results

This particular model was submitted to a cross-validation and as such the results available correspond to the average validation results of the five splits. These results can be seen in Table 5.2.

Table 5.2: Results obtained by the Region Detection Model

Class	Dyskeratotic	Koilocytotic	Metaplastic	Parabasal	Sup-Int	Average
AR@100	0.62014	0.63065	0.75044	0.61435	0.71971	0.66706
mAP@.50IOU	0.56760	0.41967	0.57412	0.53607	0.46209	0.51191

The model does not show a lot of discrepancy between the several classes, but it seems to behave better when detecting Metaplastic and Superficial-Intermediate cells.

5.2 Nuclei Detection Model

The Nuclei Detection Model [5] was developed by Vladyslav Mosiichuk, Paula Viana, Tiago Oliveira and Luís Rosado in the scope of the TAMI^2 project.
The model aims to automate the assessment of sample adequacy according to the guidelines established in the Bethesda System. Unlike the region detection model, its ultimate goal is not to correctly detect and categorize whole cells, but to detect the nucleus regions of squamous cells. It consists of a Retinanet [61] model that detects and classifies each nuclei according to its adequacy. A visual representation of the model's backbone architecture can be seen in Figure A.2 in the appendices.

5.2.1 Dataset

The private dataset used to train this model is composed of 765 images obtained using the µSmartScope prototype [4].

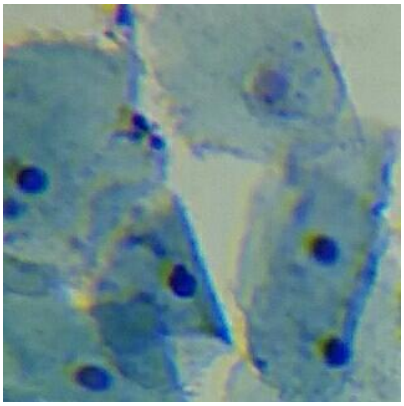

Figure 5.3: Example of an image from the private dataset

The nuclei in each of the images was manually annotated by experts, which defined a bounding

box for each cell's nuclei and classified it according to the following types: Squamous nucleus, Inflammatory cell, Glandular nucleus, Undefined nuclei, Undefined all and Artifact. The split in train and test sets was done manually to ensure a consistent distribution of class. In total, the dataset contains 42387 annotations, which are distributed and split as shown in Table 5.3.

Table 5.3: Distribution of cells in the private dataset

Category	Nº of annotation	Train Split	Test Split
Squamous nucleus	21973	79%	21%
Inflammatory cell	17914	71%	29%
Glandular nucleus	684	81%	19%
Undefined nuclei	667	74%	26%
Undefined all	997	84%	16%
Artifact	152	81%	19%
Total	42387		

The images were obtained and annotated by experts from Hospital Fernando Fonseca and Instituto Português de Oncologia in Porto (IPO).

5.2.2 Results

Regarding detection and classification, the model performed 5216 predictions for 5483 existing annotations, which resulted in a True Positive Rate of 74% and a total of 473 False Positives. However the annotations present in the dataset were incomplete, as the specialists did not annotate certain cells due to doubts about poor visibility of the object. As such, some of these False Positives turned out to be correct. The full results of the model can be seen in Table 5.4

Table 5.4: Results obtained by the Nuclei Detection model

Class	AP	Recall	Accuracy	TP	FP
Squamous nucleus	0.8236	0.7380	0.7984	4743	473
Inflammatory cell	0.7072	0.6274	0.8310	2091	566
Glandular nucleus	0.2257	0.0347	0.9864	5	7
Undefined nuclei	0.0	0.0	0.9837	0	1
Undefined all	0.0	0.0	0.9847	0	0
Artifact	0.0	0.0	0.9973	0	0

As would be expected the model showed a better behaviour in classifying cells belonging to the Squamous and Inflammatory classes, although as mentioned, the main purpose of the models was only to classify each cell between Squamous and non-Squamous.

Chapter 6

Methodology

6.1 Overview

The proposed methodology consists of a model-agnostic pipeline composed of 4 separate blocks. Initially, a batch of images is sent through the previously trained model, and the model's feature extractor is applied to each image generating an intermediate representation of the image's feature space. These extracted feature maps will then serve as input for the dimensionality reduction module, whose output will be a reduced representation of the initial feature map. Taking advantage of the reduced dimension, the maps are then compared using different metrics in order to obtain the cells in the train set that the model considers the most similar to each image in the test set.
After the similar cells are obtained, the results can be filtered by more specific morphological characteristics of the cell.

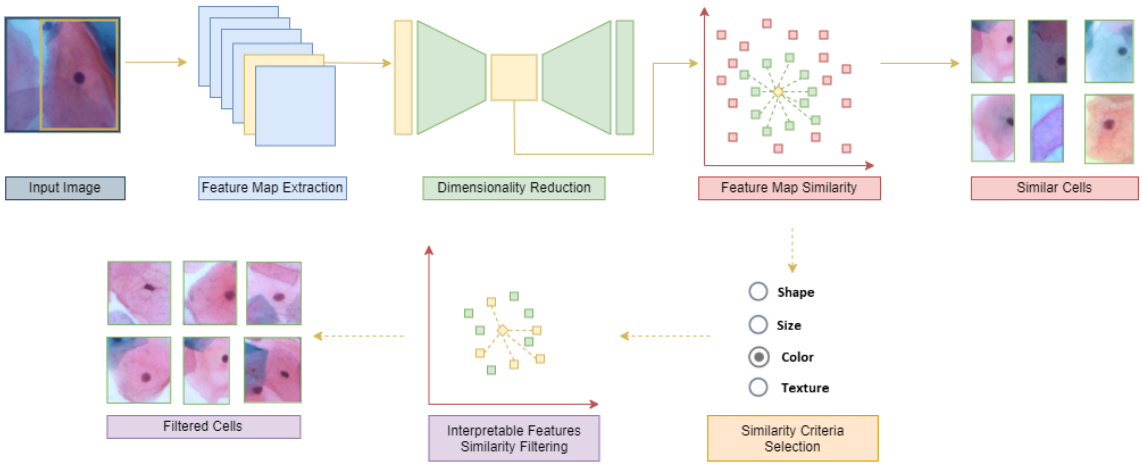

Figure 6.1: Visual representation of the proposed pipeline and respective blocks

6.2 Data Processing and Feature Map Extraction

As mentioned previously, the images used in the pipeline come from two different datasets depending on the pre-trained model being used. These images were already processed and labelled in accordance with the necessity of the models, and as such, no further processing is necessary.
As explained in Chapter 5, the images composing the datasets correspond to patches of the original microscopic field. As such, each image contains several individual cells. It is necessary to extract the feature map for every corresponding cell present in the dataset images in order to be able to compare them. In this case, we retrieve the feature maps generated by the backbone network of the previously trained detection model.
It was observed that the pipeline obtained considerably better results when the feature maps were extracted after a batch normalization layer, however as a batch normalization shifts and scales the data, it would be expected that the results from other layers would be just as good if they were normalized in their turn. However, since no concrete gain came from using a different neighbouring layer, the extraction from a batch normalization layer was chosen to avoid any additional data processing.
In this case, we use the last batch normalization layer of the feature extractor. Choosing a layer closer to the input would guarantee that the feature maps maintained more morphological information regarding the cell, however the use of a layer closer to the output guarantees that: (i) Only the most relevant morphological structures are maintained, (ii) The feature maps are closer to the prediction, and as such are more relevant when trying to explain the model's decisions.
In the case of both models, the obtained feature maps corresponded to the whole area of the initial picture, and as such, the feature map has to be cropped in terms of the detected bounding box in order to obtain the feature map of every corresponding detection. The cropping is done by normalizing the positions of the detected bounding boxes in relation to the feature map size. This process will result in several feature maps with different dimensions, each corresponding to an individual cell present in the original patch.
The lack of consistency regarding the dimensions of each map would present itself as a problem in the following steps, as the dimensionality reduction methods require a batch of feature maps with identical shape. To counter this, each map is subject to a Region Of Interest (ROI) Alignment.
ROI Alignment is an operation that allows the extraction of small feature maps from each region, producing fixed-size feature maps from non-uniform inputs. This operation is similar to the ROI Pooling layer proposed in the Fast R-CNN architecture [62]. In this case, it extracts feature maps whose dimension are in accordance to the necessity of the pipeline.
As a final step, the obtained feature maps are flattened into feature vectors.
After every feature vector is obtained, the data is then split into train and test sets according to the original dataset. The test set will contain the initial cells, and the train set will contain the images that will serve as the possible similar cells.

The Region Detection model takes as input an image of size 320x320x3, which consists of a

patch containing several different individual cells. The corresponding feature maps have a size of 10x10x1280 and are extracted after the last Batch Normalization layer of the model's feature extractor. The necessary cropping for each single cell results in a set of feature spaces with 1280 channels and variable width and height. After the ROI Alignment, these spaces will have a size of 5x5x1280. The flattening of these feature spaces will result in a 1x32000 feature vector.

The Nuclei Detection model also takes as input an image of size 320x320x3, and the feature maps are extracted after the last Batch Normalization layer with a size of 10x10x512. After the necessary cropping and ROI Alignment are performed, the feature spaces will have a size of 5x5x512, resulting in a feature vector of size 1x12800 after being flattened.

6.3 Dimensionality Reduction

Due to the size of the Feature vectors obtained from the intermediate layers of the model, it is not advisable to compare them without additional processing. The high-dimensional space must be reduced to avoid problems related to the phenomena commonly known as the "Curse of Dimensionality" [63], hence it is necessary to obtain low-dimensional spaces which maintain the relevant properties of the original feature vector.
To achieve this, two different methods are proposed, namely Principal Component Analysis and Variational Auto Encoder:

6.3.1 Principal Component Analysis

Principal Component Analysis (PCA) presents itself as one of the most common methods for dimensionality reduction. For this work, the scikit-learn's [64] PCA implementation was used.
The number of components that should be kept in each model was decided based on the trade-off between the number of components and the cumulative variance. These comparisons can be seen in Figures 6.2 and 6.3 for each corresponding model. Regarding the Region Detection model, it was observed through the plot present in Figure 6.2 that a 95% threshold would be an acceptable value to establish as minimum cumulative variance whilst maintaining a reasonably small amount of components. In this case, the 95% cumulative variance translates to 110 components.
In the Nuclei Model, however, it can be observed in Figure 6.3 that the PCA applied to this model results in a more significant loss of information. As such, the 95% threshold would result in a number of components larger than advised. Thence, for this model a 90% threshold was established, resulting on 120 components.

6.3.2 Variational Autoencoder

Although the PCA does indeed reduce the feature vectors to low-dimensional spaces, it becomes clear that for a reduced number of components the information loss is still very high. It then becomes necessary to explore other options for Dimensionality Reduction, and in accordance with

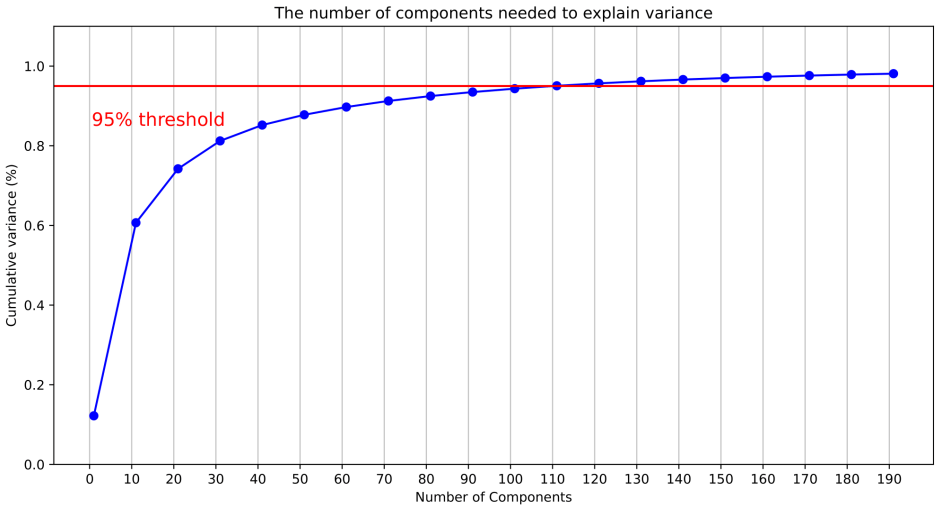

Figure 6.2: Relation between the number of PCA components and the cumulative variance when applied the Region Model

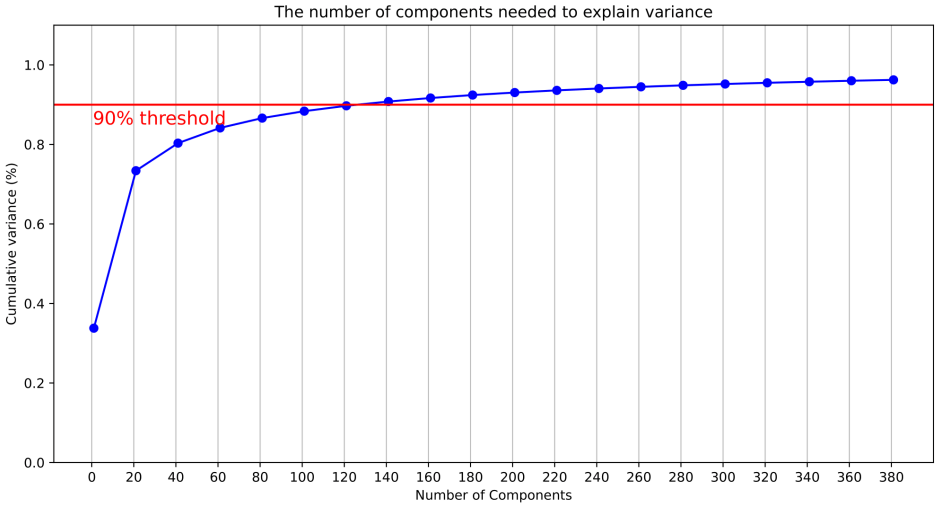

Figure 6.3: Relation between the number of PCA components and the cumulative variance when applied the Nuclei Model

the review of the state-of-the-art in Chapter 4 we propose the use of a Variational Autoencoder (VAE) to achieve this goal.

The implemented VAE is based on the architecture proposed in [65], which achieved good results in the previously mentioned article [53]. The full architecture of both the encoder and decoder can be seen in the appendices in Figures A.3 and A.3 respectively.

The VAE was trained using a machine with 16 GB of RAM, equipped with a 16-core AMD® Epyc 7302 CPU and a NVIDIA A16-8C GPU.

In these settings, the training of the VAE took over 30 hours, and due to this it was not possible

to perform an extensive amount of experiments with different batch sizes in order to tune the parameter. However, it became apparent that larger batch sizes resulted in a greater loss for the model. As such, a mini-batch size of 32 was chosen, as it yielded the best results.
The original VAE architecture defined a value of 512 for the latent dimension however, this value was too large for our purpose, as it would hinder the feature map comparison and would not be seen as a valid alternative to the PCA. For this reason, a new valid value had to be defined closer to the number of components used in the PCA. Through some experiments, it was observed that a value of 100 for the latent dimension was satisfactory.
The full parameter setting for the VAE training can be seen in Table 5.1
As both models used different datasets and produced feature vectors with different sizes, two

Parameters	Values
optimizer	Adam
	learning_rate = 0.001
batch size	32
latent dimension	100
early stopping	min delta = 0.1
	patience = 100

Table 6.1: Parameters used for the training of the VAE

VAEs had to be trained. The main architecture of the VAE's are the same except for the input shape. As the feature vector from the region model is significantly larger than the one obtained from the nuclei model, the possibility of increasing the latent dimension was considered. However, there were no practical results that evidenced a rise in performance from the model when this was attempted.
The graphs representing the training loss for both models can be seen in the annexes, in Figures B.1 and B.2 for the region detection model, and in Figures B.3 and B.4 for the nuclei detection model.

6.4 Feature Map Similarity

Having the feature vectors reduced to a small dimensional space, they must now be compared in order to find the ones that are most similar. Specifically, for each image present in the test set, the N most similar images will be obtained by comparing the feature vectors, thus providing the user with a number of example-based explanations for the classification of each image. It is important to note that these images will indicate a similarity within the model's prediction rather than a purely visual one.
For the required purpose, the K-nearest neighbours (KNN) algorithm stands as the most obvious solution to the problem, as it is intrinsically sensitive to the local structure of the data. Thus, the necessary comparisons can be made by using the feature space generated by the KNN.
To perform these comparisons, it is necessary to select valid distance metrics that ensure a correct matching. In [66] the authors perform an extensive study to identify the most valuable metrics

when comparing feature spaces. Based on the results of their work, four different metrics were chosen and compared in order to identify the most valuable for our pipeline.

Euclidean Distance: Measures the square root of the sum of the absolute differences between two feature points.

$$Euclidean\,(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan Distance: Measures the distance between two points through the sum of the absolute differences of their Cartesian coordinates.

$$Manhattan\,(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Chebyshev distance: Measures the greatest of the differences between two points along any coordinate dimension.

$$Chebyshev\,(x,y) = max(|x-y|)$$

Cosine distance: Measures the normalized dot product of the two feature points.

$$Cosine\,(x,y) = \sum_{i=1}^{n} \frac{(x_i \cdot y_i)}{||x_i|| \cdot ||y_i}$$

The used metrics and implementation of KNN are from the scikit-learn library.

6.5 Interpretable Features Similarity Filtering

The Feature Filtering is an optional stage of the pipeline which applies additional filtering to the similar images, based on morphological features of the given cell. If this stage is active, then the "Feature Map Similarity" stage will provide the 50 closest images instead of the default 8.

Those 50 images will then be subject to several similarity criteria based on the domain knowledge. Fraunhofer Portugal had previously obtained these criteria in the scope of the TAMI project. Through user research meetings with medical specialists from IPO-Porto, some relevant morphological characteristics were identified, which were considered valuable when analysing the similarity between cellular structures in cervical cytology. These criteria were then mapped with specific metrics, which could be extracted through the use of a previously developed computer vision library named Texture-Color-Geometry Feature Extraction (TCGFE) [6].

The TCGFE is a library developed by Fraunhofer Portugal. This library can be used to extract three main types of features: geometry, texture and color. To perform this extraction, the library uses the original image and a binary mask which delimits the corresponding regions from where the extraction occurs.

The extracted features are interpretable, as they quantify something that is visually interpretable (such as shape, color or texture). These criteria and corresponding metrics can be seen in Table 6.2.

Table 6.2: Considered similarity criteria and metrics. [6]

Criterion	**Sub-Criterion**	**Interpretable Features (TCGFE Metrics)**
Nuclei	Size	Maximum Diameter, Bounding Box Area
	Shape	Principal Axis Ratio, Irregularity Indexes
	Color	Energy, Mean, Entropy
	Texture	Mean, Stardard deviation, Maximum, Short run emphasis, Long run emphasis, Run percentage, Long run high grey level emphasis, Low grey level runs emphasis, high grey level runs emphasis, Short run low grey level emphasis, Short run high grey level emphasis, Grey level non-uniformity, Long run low grey level emphasis, Energy, Entropy, Contrast, Homogeneity, Correlation, Maximum Probability
Citoplasm	Color	Energy, Mean, Entropy
	Texture	Mean, Stardard deviation, Maximum, Short run emphasis, Long run emphasis, Run percentage, Long run high grey level emphasis, Low grey level runs emphasis, high grey level runs emphasis, Short run low grey level emphasis, Short run high grey level emphasis, Grey level non-uniformity, Long run low grey level emphasis, Energy, Entropy, Contrast, Homogeneity, Correlation, Maximum Probability
Cell	Size	Maximum Diameter, Bounding Box Area
	Shape	Principal Axis Ratio, Bounding Box Ratio, Irregularity Indexes

If a given criterion or sub-criterion is chosen for filtering, the values of the corresponding metrics are normalized for each image and the 8 closest ones are obtained using a K-nearest neighbors algorithm, using the Euclidean distance as metric. These images will combine the previous feature map similarity with a morphological similarity.

The mapping between the similarity criteria and Interpretable Features (TCGFE Metrics) presented in in Table 6.2 were obtain through an empirical approach, and had not been tested before this work.

Chapter 7

Results and Discussion

7.1 Metrics

The pipeline's ultimate goal is to provide the user with visual explanations that represent the model's decision process. However there is no concrete way to define or evaluate the correctness of these explanations. As seen in the articles analysed in the state-of-the-art revision, most example-based XAI models are evaluated via human-centred approaches, by performing user based experiments which require the participation of a large number of specialists in the given area.
To overcome this difficulty a computer centered metric was used instead. This metric cannot fully evaluate the model's performance, but it can be used to compare the different approaches mentioned in the methodology.
Each of the experiments mentioned in the following section were made by using 1000 test images, for a different number of neighbors in the KNN algorithm.

7.1.1 Average Class Accuracy (ACA)

The most important evaluation for this particular pipeline is verifying if the similar cells are indeed similar to the initial cell. In the scope of this work, this similarity will not be purely visual, but also within the model's decision process. As there is no simple computer-based way to verify this, a metric was developed which aims to evaluate the efficiency of the pipeline of choosing the correct examples. This metric is based on two premises: (i) That the most similar cells will belong to the same class, and (ii) That when providing an example to explain a model's decision, this example must be in accordance with the decision itself. In other words, if the model provides a certain prediction, then the majority of the explanations should belong to the predicted class.
As such, the used Average Class Accuracy (ACA) metric measures the average proportion of similar cells that have the same class as the initial cell.

For a test image i:
a_i = The average of resulting images which belong to the same class as the test image.

For the whole set of test images, of size N:

$$ACA = \frac{\sum_{i=1}^{N} a_i}{N}$$

In a practical context, the images present in the test set correspond to cells detected by the respective model. Thus the ACA scores are limited to the performance of the model, as some of the images may not contain objects which are recognizable as cells, and the class of those object may not corresponds to its real class. This hinders the performance of the ACA metric, since we are evaluating the explanations of detections which are incorrect. Due to this, for each model, two experiments were made: (i) The first consists of the use of detections as test images, which corresponds to the pipeline as it would be used in a real-life context. This experiment evaluates the pipeline's ability to explain the model's decision. (ii) The second experiment consists of the use of ground-truth annotations as test images. This experiment is used to evaluate the pipeline's performance in a context where the model would work to perfection, which, although unrealistic, is necessary for a full evaluation of our pipeline.

Even though the ACA metric does not fully evaluate the similarity, we must consider that the purpose of the pipeline is to explain the model's prediction and not improve it. Thus the chosen metric is valuable as a comparative measure to indicate a similarity within the model's decision by taking into account the predicted classes. However, this metric was used along with a visual analysis of the results to ensure that the pipeline provides examples which are easy for the user to comprehend.

7.2 Feature Map Extraction

Regarding the first step of the pipeline, the Feature Map Extraction was made in a linear way, which did not depend on any parameters. As such, the extraction produced the expected results for both models.
However, some initial experiments were conducted to assess the most adequate way of extracting the feature maps for the intended purpose. These experiments were analyzed mostly through visual observations of the results, or through the intuitive interpretation of the obtained data. An example of this was the previously mentioned choice of extraction layer.
Apart from the ROI Alignment, further processing methods were attempted in order to increase performance and reduce the required time for training the models. These methods included applying average, max and min operations to the feature maps, but these attempts cause considerable data loss, which drastically decreased the pipeline's performance.

7.3 Region Detection Model

As the ACA metric is only calculated after the whole pipeline has executed, the result analysis will start with the KNN metrics, and based on those metrics a comparison will be made regarding the best dimensionality reduction method.
This analysis will be made for both experiments, using detections and annotations as test images.

7.3.1 Detections as test images

7.3.1.1 KNN metrics

Regarding the region detection model, it is necessary to identify which KNN metric provides the best results, for both the PCA and VAE methods. Figures 7.1 and 7.2 show the comparison between the four chosen metrics, as well as the progression of the ACA with the rise of the number of KNN neighbors.

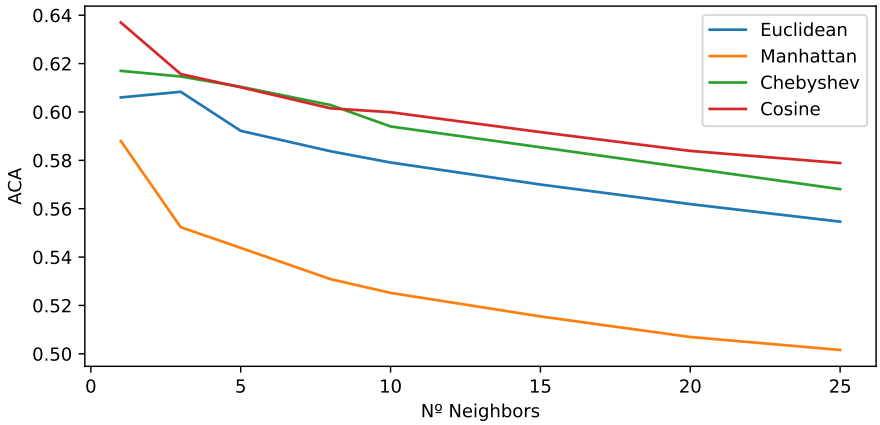

Figure 7.1: Comparison of ACA scores for each of the four KNN metrics, with a PCA for dimensionality reduction, and using the region detection model

As can be seen, the best performing metric varies according to the dimensionality reduction method, and all four metrics show slightly different results of ACA scores.
For the PCA, the cosine distance proved to be the best metric, although the difference to the Chebyshev metric is minimal. For the VAE the Euclidean or Manhattan distances provided the best results. This indicates that as would be expected, the KNN behaves differently depending on the method used to reduce the data. It also indicates that the identification of the best metrics is relevant, specially when considering a generic approach which is meant to be applied to different models and use cases.
The figures also show that, as the chosen number of neighbors gets higher, the ACA score drops, supporting the intuitive assumption that less similar images will be less likely to belong in the same class.
From a purely visual perspective, the difference between the metrics is not always noticeable,

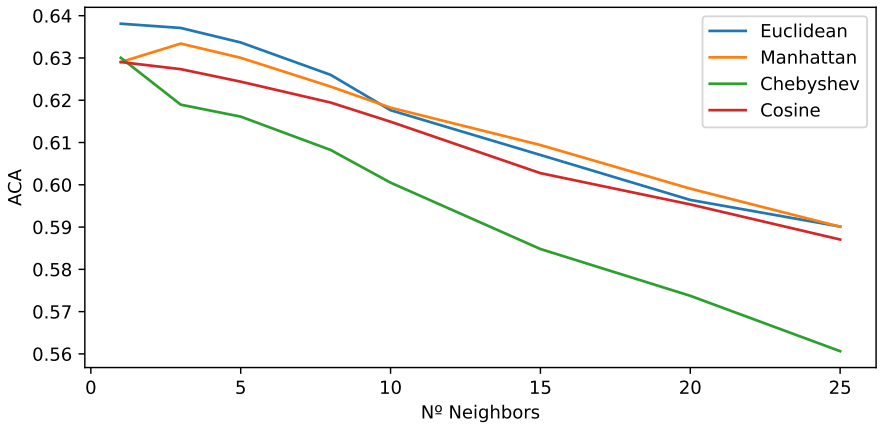

Figure 7.2: Comparison of ACA scores for each of the four KNN metrics, with a VAE for dimensionality reduction, and using the region detection model

however in some instances we can observe some disparity between the obtained images. An example of this can be observed in Figures 7.3 and 7.4 which represent the model's results for the Cosine and Manhattan metrics respectively, using PCA as the dimensionality reduction method. Additional examples of this can be seen in Appendix C.

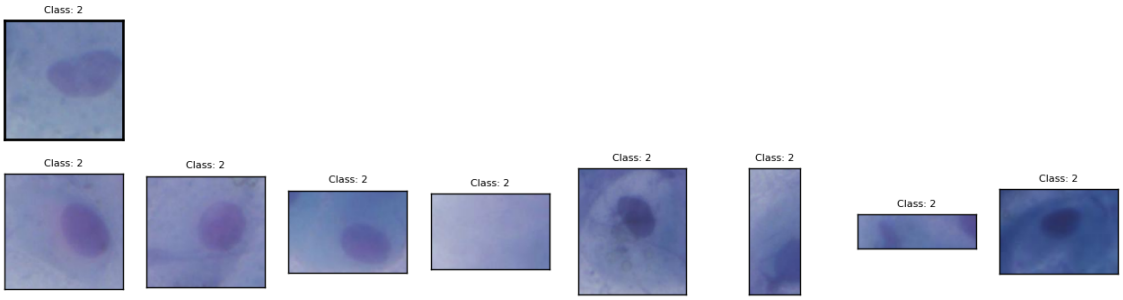

Figure 7.3: Example of a pipeline result when using PCA and Cosine distance, for the region detection model

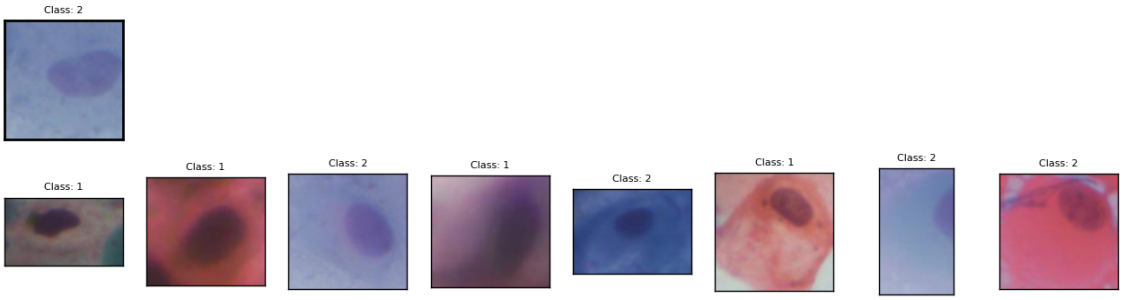

Figure 7.4: Example of a pipeline result when using PCA and Manhattan distance, for the region detection model

From Figure 7.4 we can also conclude that although the results show mainly cells from the same

class, these cells present variations in size, color and texture. In this particular case, additional filtering could help standardize these visual similarities.
We can conclude that the used metric has an impact in the overall performance, and that it may prove crucial to ensure a correct interpretation of the models explanation.

7.3.1.2 Dimensionality Reduction

Regarding the dimensionality reduction, we compare both methods by comparing the results of the distance metric which obtained the highest average ACA score. In the case of the PCA it was the Cosine distance, and in the case of the VAE it was the Euclidean distance.

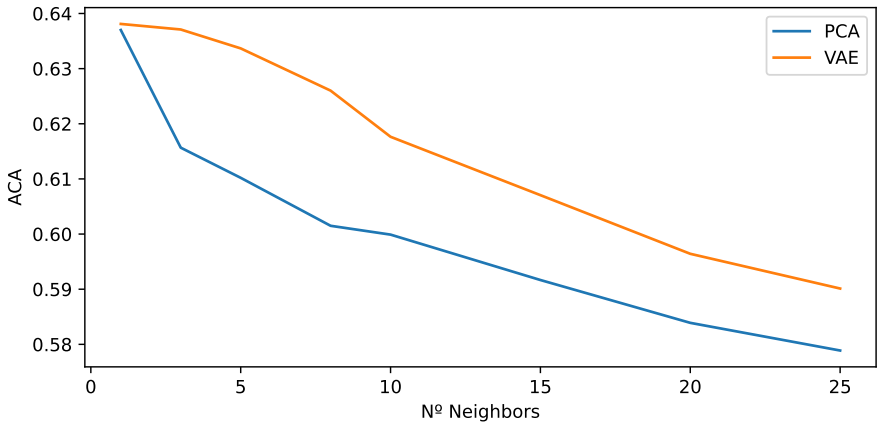

Figure 7.5: Comparison of best ACA scores for both dimensionality reduction methods, using the region detection model

As can be seen in Figure 7.5, the difference in results for both methods is small. However, the VAE obtains slightly better results for all the observed number of neighbors.
In our practical context, the number of neighbours is extremely relevant, as the bibliography indicates that a large number of explanations may hinder the objective of increasing trust. As such, for a practical comparison, more emphasis should be given to the results for a small number of neighbours, such as a number between five and ten. For this interval we can see that the VAE's advantage is more noticeable. However, visually we can observe that both methods provide similar results as can be seen in Figures 7.6 and 7.7, and in Appendix D, and it is difficult to distinguish a clear advantage in using the VAE.

7.3.2 Annotations as test images

If we perform a closer analysis of the ACA results of the pipeline for each class, and compare them to the AP100 results of the used model we can see a clear correlation between the values in Figure 7.8.

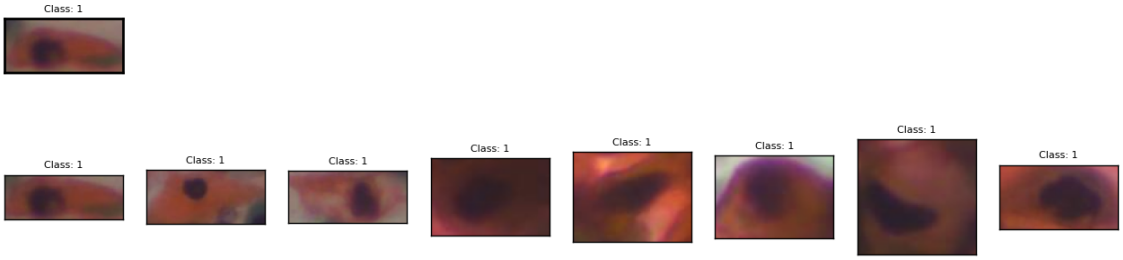

Figure 7.6: Example of a pipeline result when using PCA and Cosine distance, for the region detection model

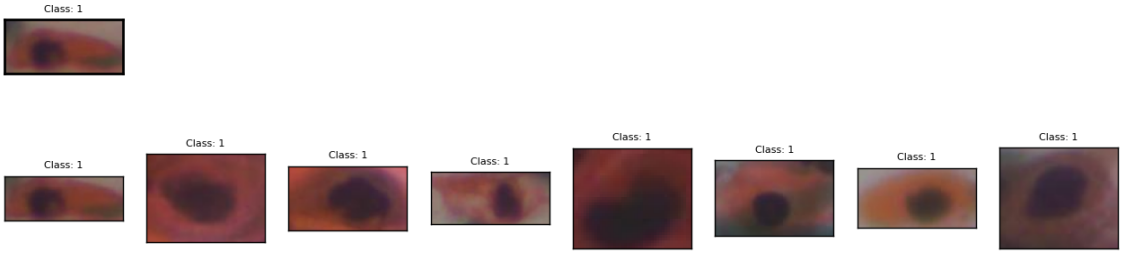

Figure 7.7: Example of a pipeline result when using VAE and Euclidean distance, for the region detection model

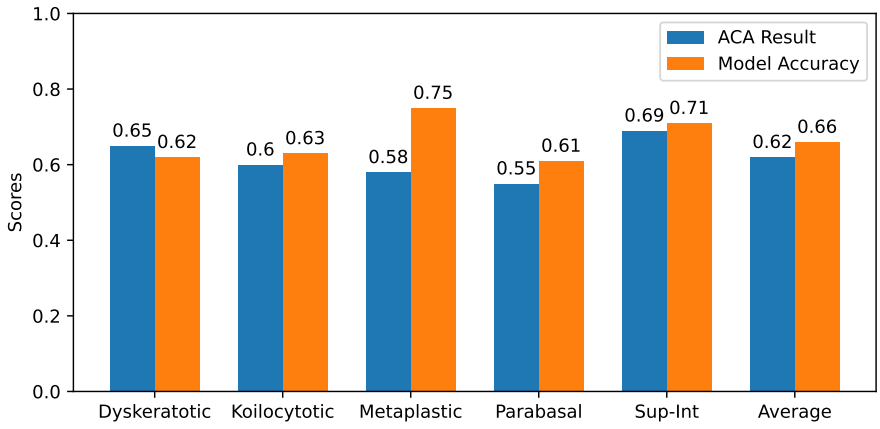

Figure 7.8: Class-by-class comparison between the results of the pipeline and the region detection model

Although the shown scores represent two fundamentally different things, the resemblance between the values serves as some indication that the pipeline is providing results that are in accordance with the model and not simply obtaining visually similar cells. However, it also indicates that the pipeline is indeed intrinsically limited by the model's performance, as would be expected.
In order to perform a technical evaluation of the pipeline's performance, we must consider a situation where the pipeline is free from the bias introduced by the model. As mentioned, this is an unrealistic setting, but its analysis is vital if we are to understand the pipeline's real performance.

7.3.2.1 KNN metrics

Since the training data used for this experiment is the same as in the previously mentioned one, it would be expected that the use of the same metrics would result on very similar comparative results between the metrics.

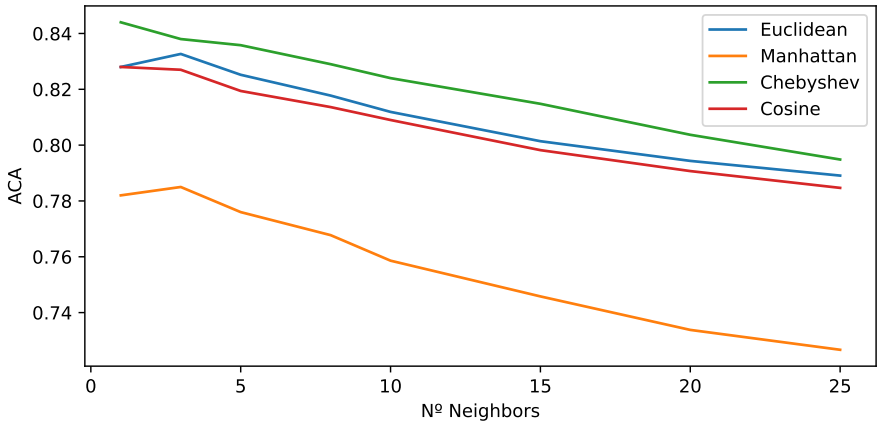

Figure 7.9: Comparison of ACA scores for each of the four KNN metrics, with a PCA for dimensionality reduction, using the region detection model and annotations as test images

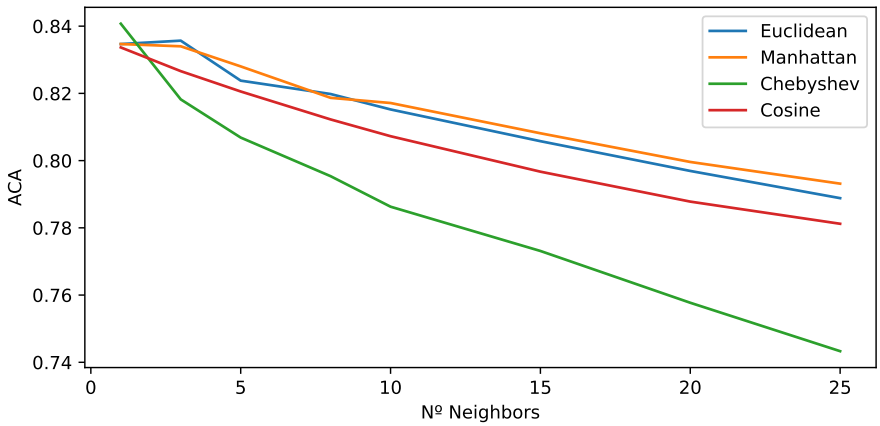

Figure 7.10: Comparison of ACA scores for each of the four KNN metrics, with a VAE for dimensionality reduction, using the region detection model and annotations as test images

Indeed, as can be seen in Figures 7.9 and 7.10, the relation between the metrics is very similar, especially in the case of the VAE, where the only main difference is the Manhattan distance producing the best results, by a minimal margin. In PCA however, there is some difference, with the Chebyshev distance obtaining the best results.
Nonetheless, it becomes immediately apparent that the pipeline obtains considerably better results without the limitation of the model, achieving ACA scores above 0.80.

7.3.2.2 Dimensionality Reduction

Once again we can compare the dimensionality reduction methods by choosing the best performing metric for each. And if in the first experiment the VAE performed slightly better, in this case the PCA obtains better results, as can be seen in Figure 7.13

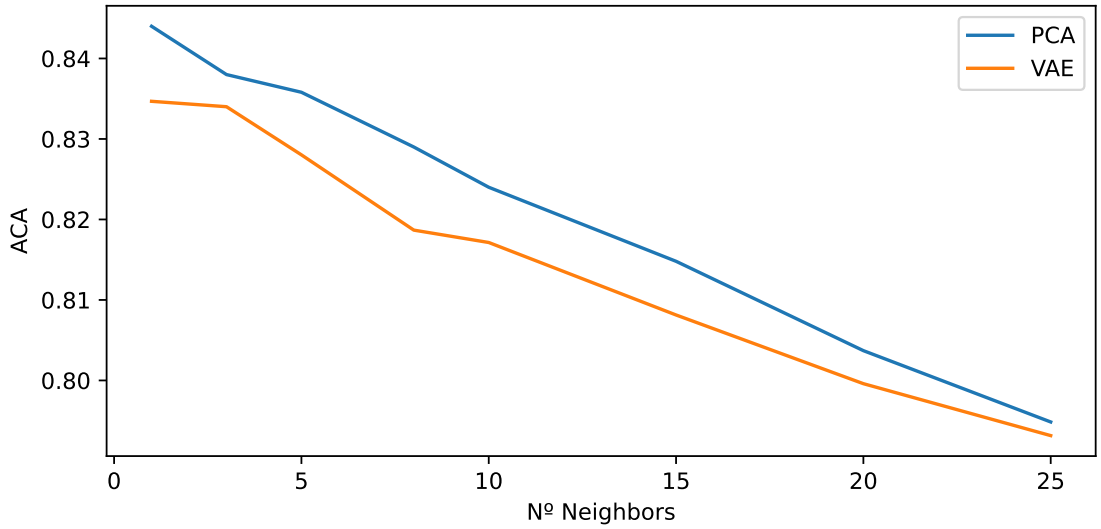

Figure 7.11: Comparison of ACA score for both dimensionality reduction methods, using the region detection model and annotations as test images

This indeed is in accordance with the visual observations made in the first experiment, showing that the difference in performance between the PCA and the VAE is minimal, and it is difficult to distinguish any as the most advantageous.

7.3.3 Interpretable Features Similarity Filtering

The last step of the pipeline is an optional phase consisting of additional filtering. As this filtering will be made only taking into consideration visual features, the use of the ACA metric is not indicated in this context. This section does not strictly deal with the notion of explainability, being merely an additional filter that processes the results of the previous steps. Due to this, only a visual representation of the results will be given.
For the cells in this dataset, we will use the filters related to shape, and cytoplasm.

7.3.3.1 Original Query

We start by having a subset of images, which were the result of the previous steps of the pipeline. For the demonstration we are considering 30 images, although usually 50 are used.
In Figure 7.12, the original images can be seen in the top left corner, and the resulting images are organized by order of similarity. When each filter is applied, the obtained images reorganize and

the 8 most similar according to the filter are provided.

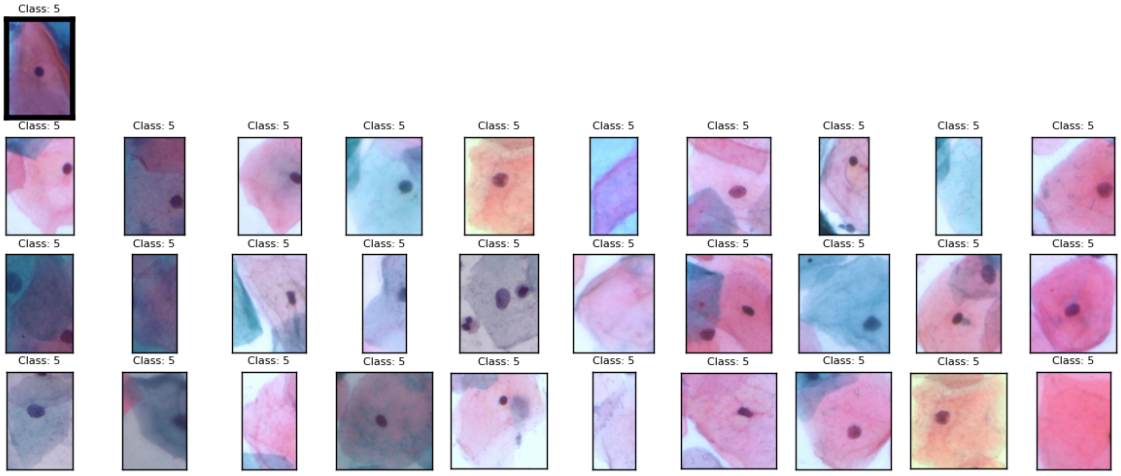

Figure 7.12: Representation of the pipeline's output, using the 30 images more similar in terms of feature maps for the region detection model

7.3.3.2 Cytoplasm Color

The first filter to be demonstrated is the cytoplasm color. We can see by the results in Figure 7.13 that the cells have been correctly filtered. The resulting cells show consistent color, and no image is present that does not fit the criteria.

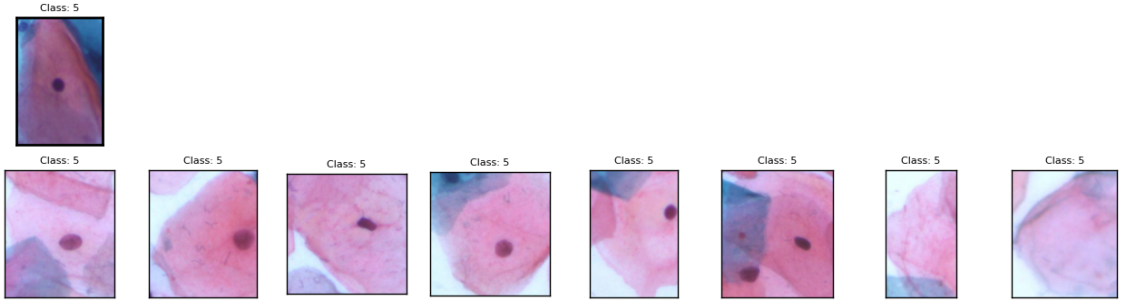

Figure 7.13: Results of the "Cytoplasm Color" filter

7.3.3.3 Citoplasm Texture

In this case the filter is still applied to the citoplasm, however the emphasis goes to its texture. Although hard to define, we can see in Figure 7.14 that the presented cells seem to have citoplasms with similar density.

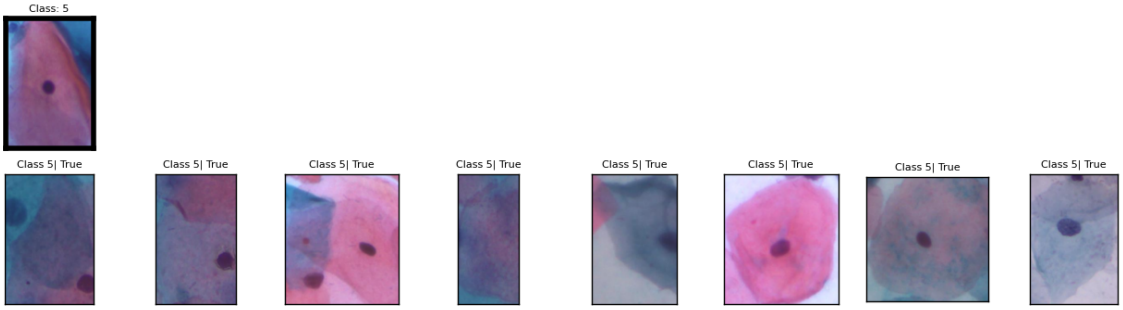

Figure 7.14: Results of the "Cytoplasm Texture" filter

7.3.3.4 Cell Size and Shape

As the dataset presents the delimitation of the cell as a bounding box, the shape-related filter uses that bounding box to estimate the size and shape. We can see in Figures 7.15 and 7.16 that although the results have been filtered, the obtained cells are similar only regarding the sizes of the bounding boxes, both in terms of area (number of pixels) and aspect ratio. With a proper segmentation around the cell however, the results would theoretically be much better.

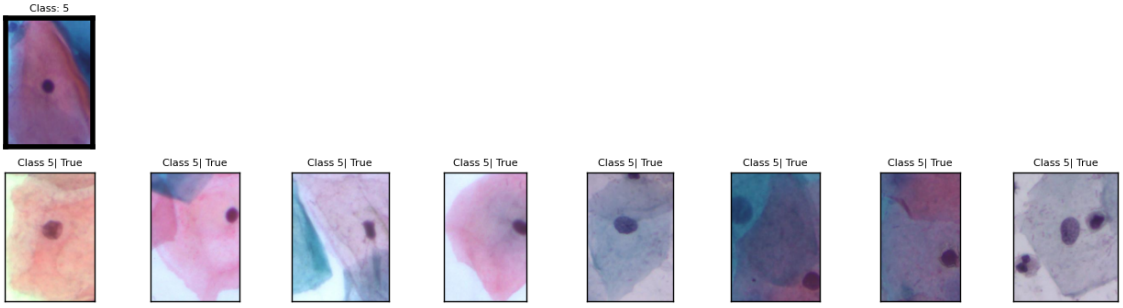

Figure 7.15: Results of the "Cell Size" filter

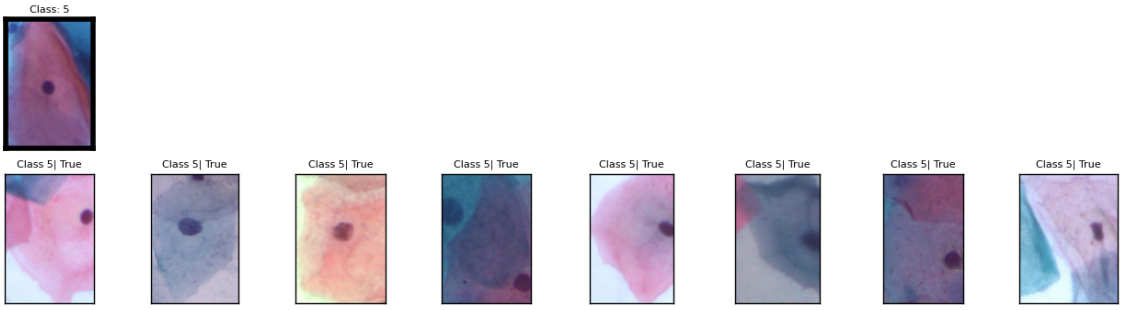

Figure 7.16: Results of the "Cell Shape" filter

7.3.4 Discussion

By analysing the results for the Region Detection Model we can observe that the distance metrics for the KNN have an impact on the general result of the model. Even though the ACA metric shows very close results, especially for the VAE, a visual observation tells us enough to conclude that the use of the best metric provides the best apparent results. From the information gathered from both experiences, we can infer that the Chebyshev or Cosine metrics are the best if using the PCA, and the Euclidean or Manhattan metrics would be the best if using the VAE.
However not many conclusions can be made regarding the dimensionality reduction methods in a general context. Even though the comparison plots show some difference between both metrics, there is no overall trend, and the visual analysis does not indicate any significant difference between the two. This is also true in terms of computational performance, as the pipeline shows very little difference in execution time when using both models. In a practical context, where only one test image is considered at a time, this difference would be immaterial, even considering that the VAE has a latent dimension of 100, and the PCA uses 110 components to represent the data.
Nevertheless, for evaluating the introduction of explainability in this particular model we should specially consider the results of the first experiment, which indicates that in a practical situation, using the real model with a reduced number of neighbors, the use of the VAE would obtain the best results.
It terms of the additional filtering, the color and texture filter appeared to obtain excellent results, proving that the filtering can bring advantages as a complementary step to the pipeline.
Overall, all the results indicate that the pipeline was able to sucessfully introduce example-based explanations to this model, being able to obtain valuable examples that take into consideration the model's decision process, as well as a visual similarity crucial to the interpretability of the decision.

7.4 Nuclei Detection Model

For the nuclei detection model the review will follow the same structure, starting with a comparison between the KNN metrics, and based on those metrics a comparison will be made between the dimensionality reduction methods.

7.4.1 Detections as test images

7.4.1.1 KNN metrics

In Figures 7.17 and 7.18 we can observe once again the comparison between the chosen metric, for the PCA and VAE respectively.
In this case we can see that the ideal metrics differ from those in the nuclei detection model. This would be expected, given the difference in the nature of the data.
For the use of PCA, the Euclidean distance is shown to provide the best results, while the VAE

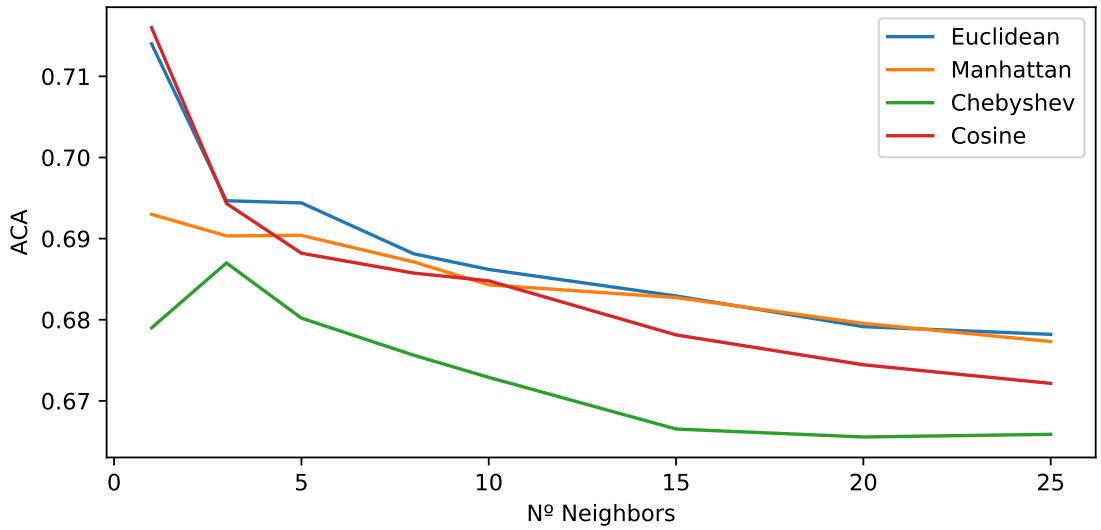

Figure 7.17: Comparison of ACA scores for each of the four KNN metrics, with a PCA for dimensionality reduction, using the nuclei detection model

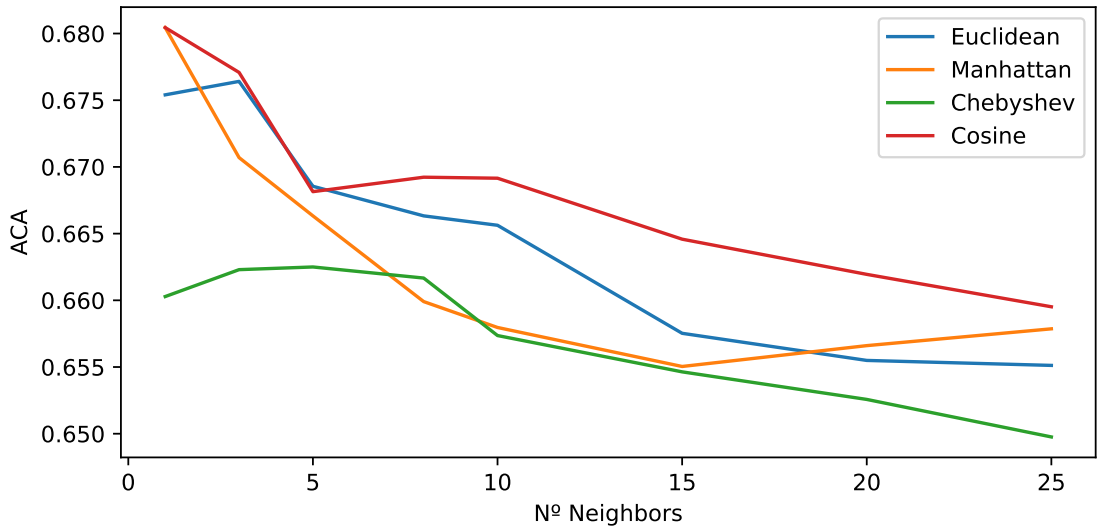

Figure 7.18: Comparison of ACA scores for each of the four KNN metrics, with a VAE for dimensionality reduction, using the nuclei detection model

method works best with a Cosine Distance.

Visually, the difference between the metrics is not very noticeable, opposite of what happened in the region detection model. For the use of PCA, we can see in Figures 7.19 and 7.20, and in Appendix E, that the Euclidean and Chebyshev distances provide very close examples:

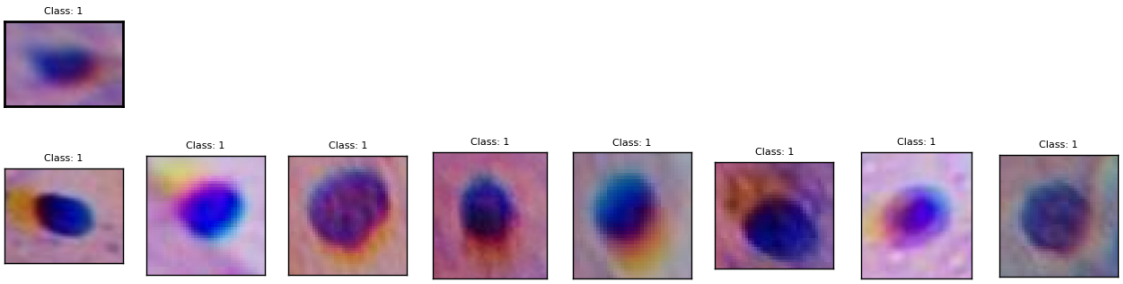

Figure 7.19: Example of a pipeline result when using PCA and Euclidean distance, for the nuclei detection model

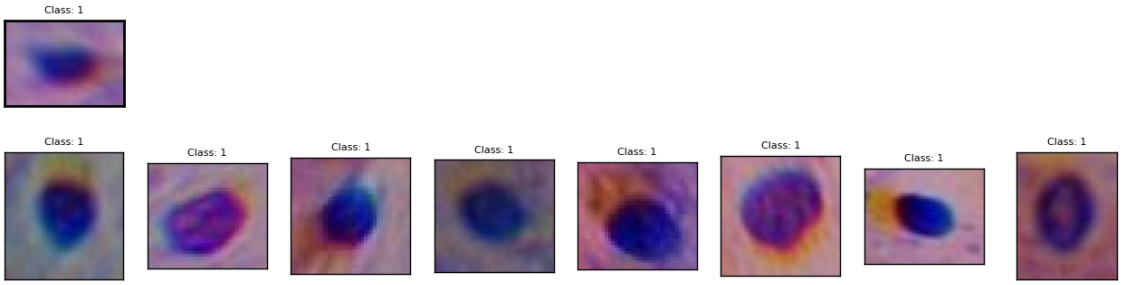

Figure 7.20: Example of a pipeline result when using PCA and Chebyshev distance, for the nuclei detection model

7.4.1.2 Dimensionality Reduction

Once again we must also compare the dimensionality reduction methods by choosing the best performing metric for each. This comparison can be seen in Figure 7.21.

In this case, and contrary to what he have seen in the region detection model, the PCA obtains

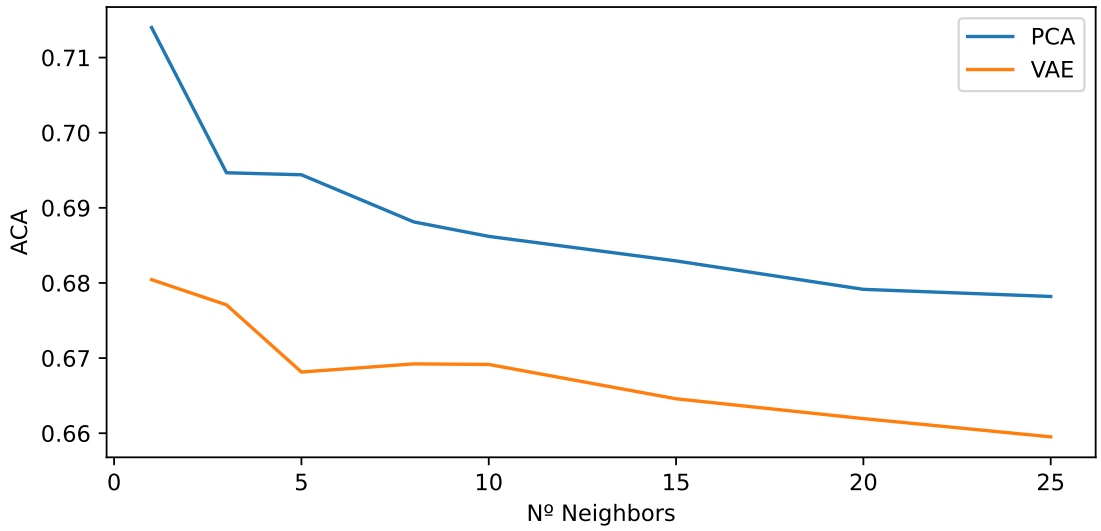

Figure 7.21: Comparison of ACA score for both dimensionality reduction methods, using the nuclei detection model

the best results when applied in the context closer to a real life application.
Also contrary to the first model, in this case a visual advantage can be seen in favour of the PCA, in some instances. An example of this is present in Figures 7.22 and 7.23, and in Appendix F.

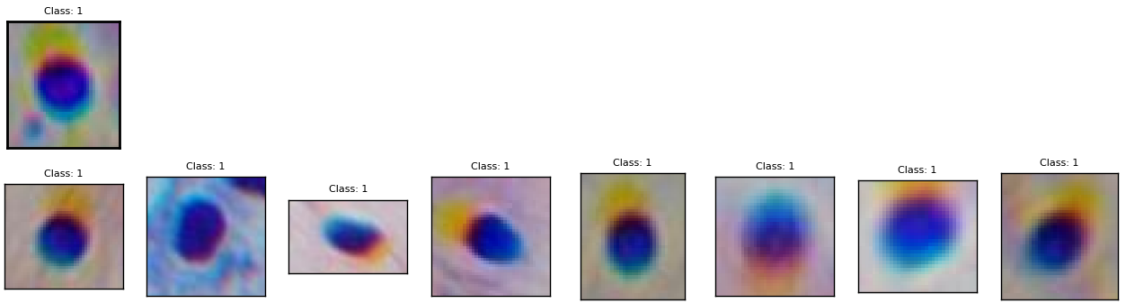

Figure 7.22: Example of a pipeline result when using PCA and Euclidean distance, for the nuclei detection model and annotations as test images

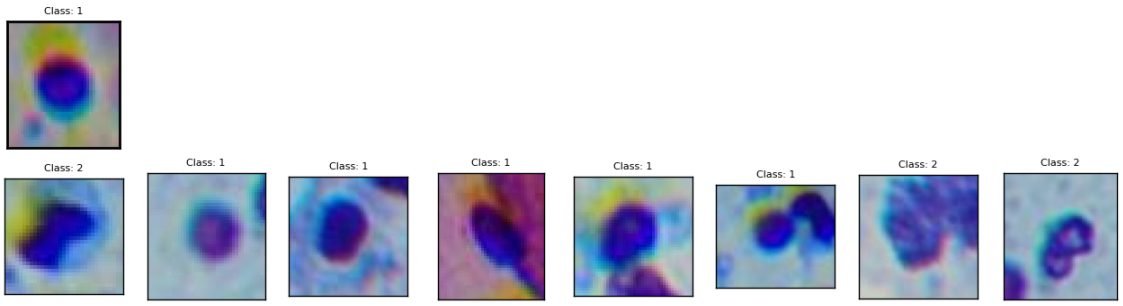

Figure 7.23: Example of a pipeline result when using VAE and Cosine distance, for the nuclei detection model

7.4.2 Annotations as test images

As mentioned before, this model's goal is not to classify cells into certain categories, but to correctly detect which cells are Squamous and which are not. As such, it does not make sense to analyse the results for the various categories. However we can observe that this model proved to be more apt in performing its desired categorization than the region detection model, obtaining considerably better Accuracy and Precision scores, thus proving more reliable. This is also noticeable in the results obtained by the pipeline.
Nonetheless it is still important to consider a situation in which the model behaves perfectly, to make a technical analysis and comparison.

7.4.2.1 KNN metrics

Through the observation of Figures 7.24 and 7.25 we understand that it is extremely difficult to make any conclusions regarding the best metric for both methods.

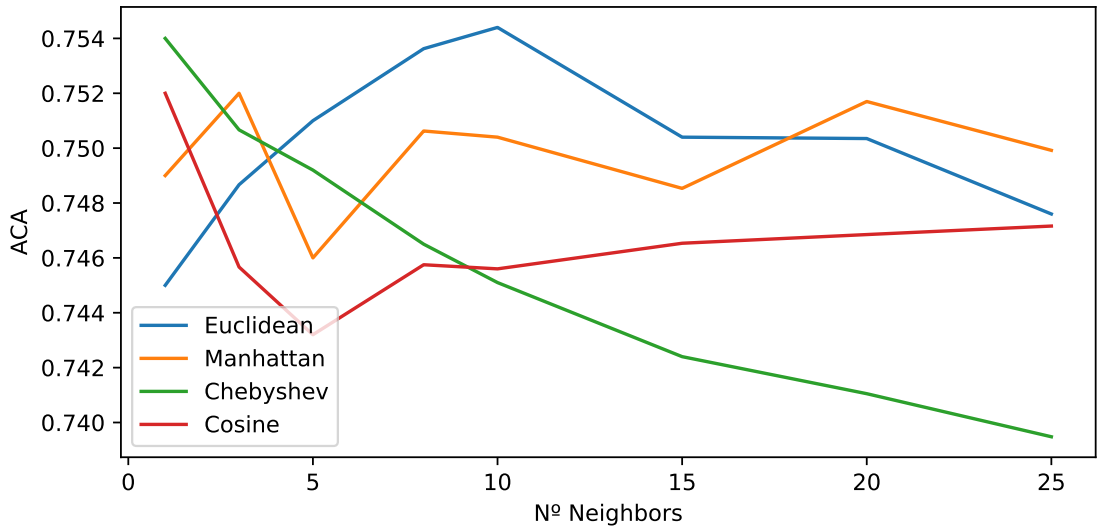

Figure 7.24: Comparison of ACA scores for each of the four KNN metrics, with PCA for dimensionality reduction, using the nuclei detection model and annotations as test images

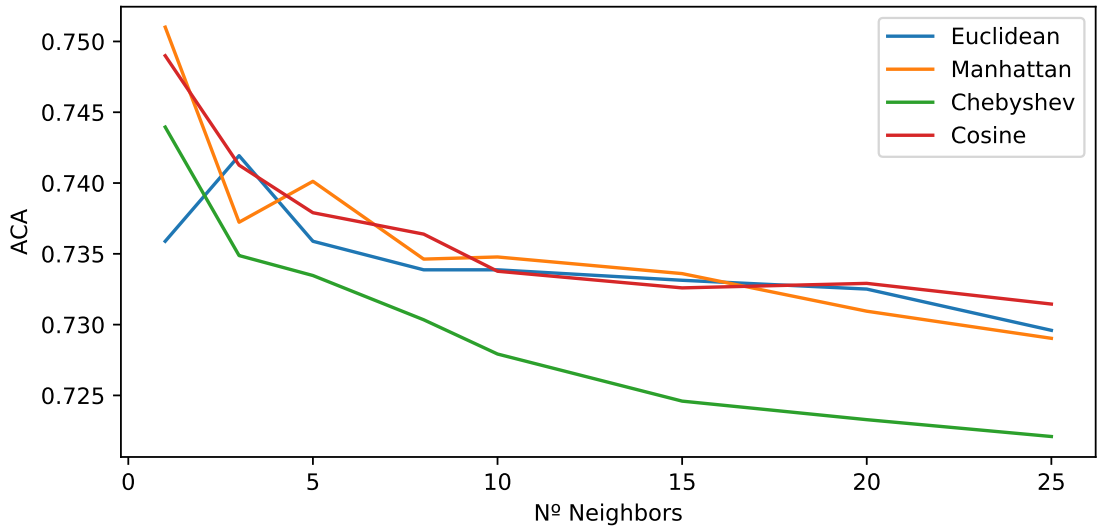

Figure 7.25: Comparison of ACA scores for each of the four KNN metrics, with a VAE for dimensionality reduction, using the nuclei detection model and annotations as test images

In the case of the PCA, the scattering of the data for different neighbors does not allow to infer any clear advantage. However, as mentioned before in a practical approach the number of neighbors would be small. This could perhaps help identify the Euclidean distance as the best

metric for this case. In the case of the VAE it is also very difficult to make any conclusion due to the closeness of the results, as the Euclidean, Manhattan and Cosine distances are separated by only 0.5 percentage points.

7.4.2.2 Dimensionality Reduction

Choosing the Euclidean distance as the best metric for the PCA, and the Cosine distance as the best metric for the VAE, we compare both in Figure 7.26:

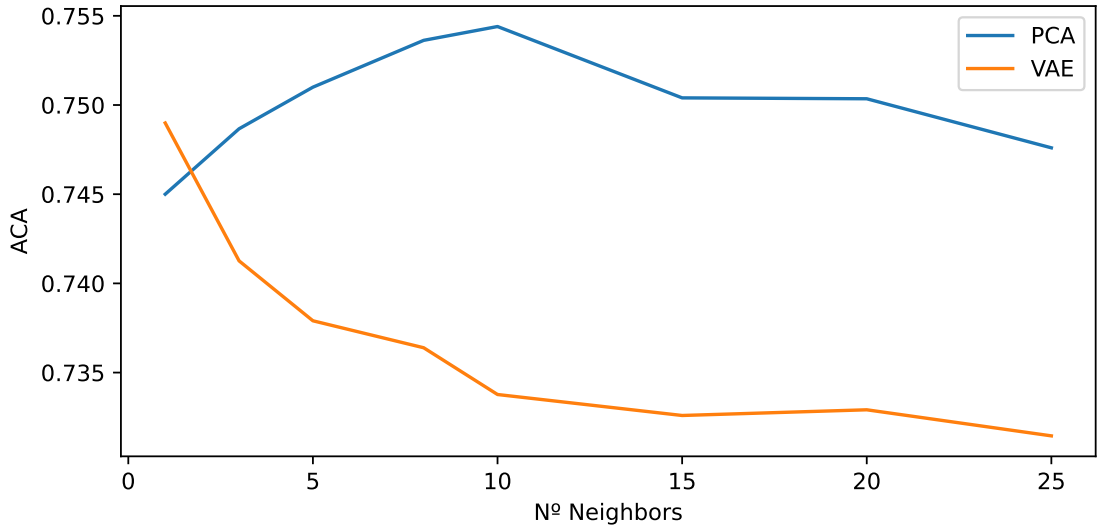

Figure 7.26: Comparison of ACA score for both dimensionality reduction methods, using the NUCLEI detection model and annotations as test images

Once more the use of PCA seems to yield the best results, in accordance to the results of the first experiment, both in the ACA metric and in the visual observation.

7.4.3 Interpretable Features Similarity Filtering

The additional filtering step should also be applicable to this particular dataset, although with different metrics. As the images only contain the cell's nuclei, for this dataset, we will use only the nuclei related filters, namely color, texture, shape and size.

7.4.3.1 Original Query

Once again, we start with a subset of 30 images obtained by the pipeline.
We can immediately see in Figure 7.27 that the images are much more similar from a visual point of view, than the images in the region detection model. Also, the images show much less focus, which can hinder the performance of the filters

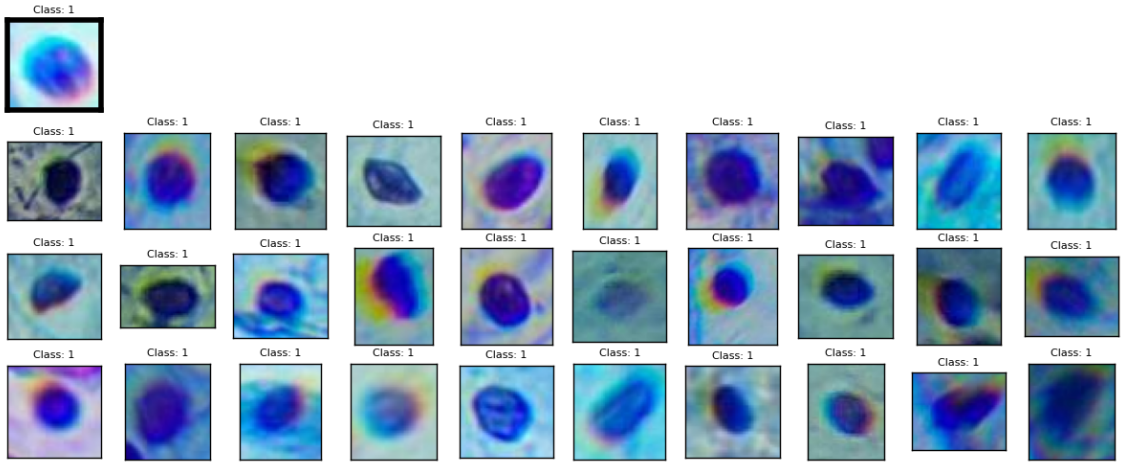

Figure 7.27: Representation of the pipeline's output, using the 30 images more similar in terms of feature maps for the nuclei detection model

7.4.3.2 Nuclei Color

In this case, as can be seen in Figure 7.28, the effect of the color filter is small, however we can see that the results show mostly lighter images.

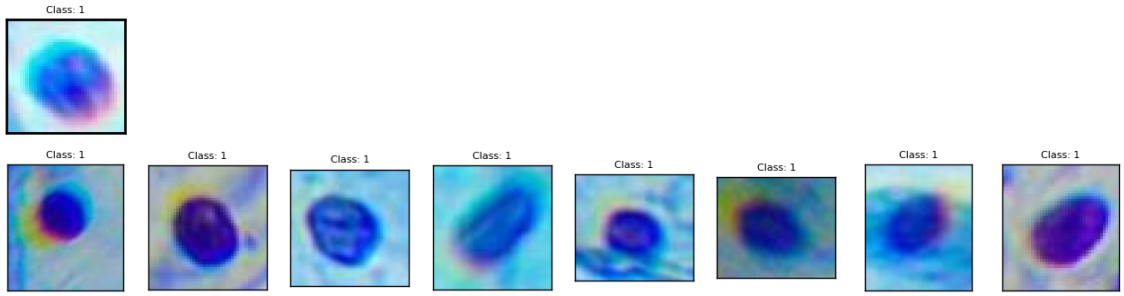

Figure 7.28: Results of the "Nuclei Color" filter

7.4.3.3 Nuclei Texture

Once more, through observation of Figure 7.29 we can see that the texture filter seems to have little effect, however we can see some images with a less dense nuclei, as the one in the original image.

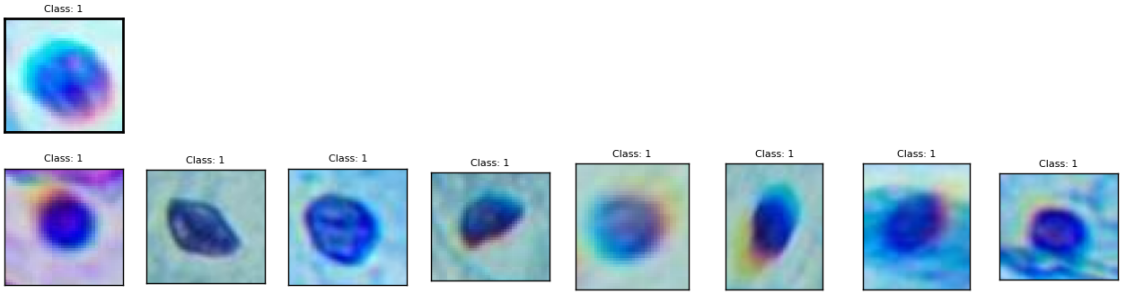

Figure 7.29: Results of the "Nuclei Texture" filter

7.4.3.4 Nuclei Size and Shape

As mentioned, in this dataset the annotations are also in the form of bounding boxes, which limits the results of the size and shape filters. In Figures 7.30 and 7.31 we see that the results show only images whose bounding boxes are roughly the same size, as in the case of the region detection mode.

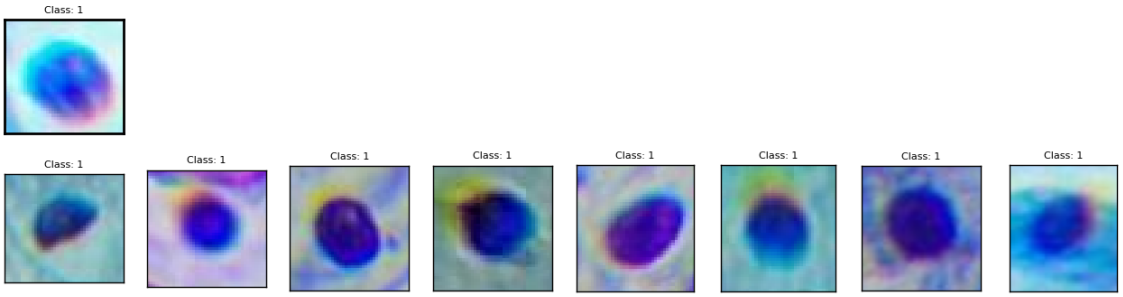

Figure 7.30: Results of the "Nuclei Size" filter

7.4.4 Discussion

Through the application of the pipeline to the nuclei detection model, we see once more that the identification of the best metrics has a positive impact on the models performance.
However, contrary to the region detection model, the best metric for each dimensionality reduction method is not as clear. Nonetheless as the main objective of this work is to provide valid explanations of the model's decision, based on an eventual practical application of the pipeline we must once more predominantly consider the first experiment. Thus, we can conclude that for the use of

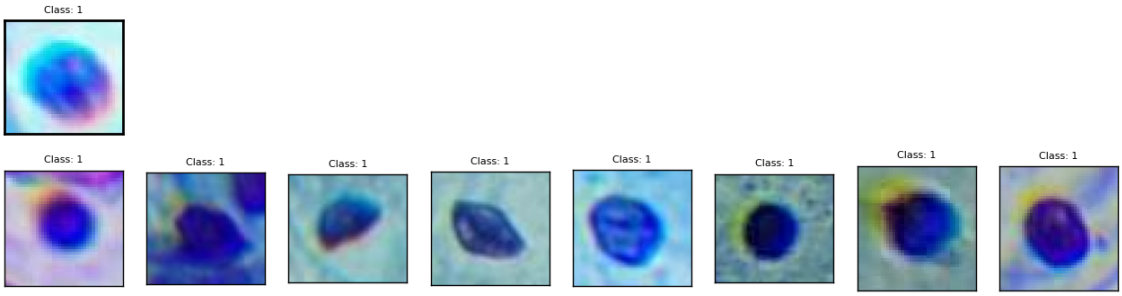

Figure 7.31: Results of the "Nuclei Shape" filter

the PCA the Euclidean Distance should be used as metric, and in the case of the use of the VAE, the Cosine distance should be used.

The use of these models also allow us to better evaluate the impact that the model's performance has in the pipeline. We can observe that the difference between the use of detections and annotations is much smaller in this context. This is due to the fact that the nuclei detection model has a considerably better performance at detecting relevant objects, which translates in much more compatible test set.

The application of filtering to the results, however, did not perform as well as in the region detection model, proving limited utility for this particular case, since the nuclei are not properly segmented.

All the results show that the pipeline is valuable when applied to different models verifying the objective of a model agnostic-aproach. In practical terms, all the evidence gathered in the bibliographic review suggests that the presence of these explanations would increase the medical professional's trust in the model, but it would also help in the decision-making by simulating the task of referring to textbook cases. In this particular context the use of the additional filtering can also be of great utility, allowing the expert to use his/her own domain knowledge to increase the pipeline's utility.

Chapter 8

Conclusion and Future Work

The primary goal of this work is the introduction of explainability to medical decision support systems by developing a model agnostic pipeline. The results obtained for two different models tell us that the proposed approach does indeed provide examples that represent in some way the model's decision process.

As most works reported in the literature overlook the evaluation of the best metrics and methods, this dissertation provides the needed information for a complete assessment of the pipeline, thus going beyond the state of the art in this aspect. We can also see that the use of different models requires a change in the pipeline's parameters, which means that although the approach is model-agnostic, the customization of the pipeline's settings for each use case is needed to ensure that the pipeline provides the best results.

The greatest challenge of this work proved to be the lack of computer-centred evaluation metrics for the required purpose, which led to some problems regarding the pipeline's full evaluation. The evaluation was made using a metric that by itself does not fully represent the intended goal. As such, it had to be combined with visual observations. Since these observations are entirely empirical, the metrics are mostly used to compare the various parameters and not to fully evaluate the pipeline's ability to increase trust, which was the ultimate goal.

To evaluate this increase of trust, some user-based experiments could be made with medical specialists and cytologists. This would give us real insight into the value of the provided explanations.

Regarding the Feature filtering based on interpretable features, we observed that it works very well when applied to full cells, especially using the metrics related to color and texture. But in the nuclei detection model the results were subpar. However, a deeper exploration of the subject could be made, especially with a model which guarantees an accurate cell segmentation. Some metrics could also be developed to evaluate the Feature filtering more correctly, so that the evaluation is not dependant on a visual analysis. To do this we could identify some metrics for each feature filter, such as the deltaE [67] for color, or a simple measurement of aspect ratio for the size and shape. This way we could evaluate if the filtered images are more similar than the initial results.

A deeper analysis could also be made of the context in which the explanations are given. The state-of-the-art review tells us that explanations are more valuable when presented in a contrastive

context. For instance, in this case, the explanations could be presented only when the model's prediction differs from the specialist's. However, there was no an opportunity to further test this notion.

Regarding the chosen data, the experiments were performed using both detections and annotations as test images. It became clear that the use of detections revealed the limitations of each model. Although not ideal for the increase of trust, this transparency can be used to aid the developer in the improvement of the models, especially when dealing with decision support systems.

References

[1] World Health Organization. *Comprehensive Cervical Cancer Control: A Guide to Essential Practice*. World Health Organization, Geneva, Switzerland, 2014.

[2] Randall K Gibb and Mark G Martens. The Impact of Liquid-Based Cytology in Decreasing the Incidence of Cervical Cancer. *Reviews in obstetrics and gynecology*, 4(Suppl 1), 2011.

[3] 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), 2019.

[4] 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.

[5] 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.

[6] 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.

[7] C Marth, F Landoni, S Mahner, M McCormack, A Gonzalez-Martin, and N Colombo. Cervical cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 28:iv72–iv83, 2017.

[8] World Health Organization. Reproductive Health, Research, World Health Organization, World Health Organization. Chronic Diseases, and Health Promotion. *Comprehensive Cervical Cancer Control: A Guide to Essential Practice*. Integrating Health Care for Sexual and Reproductive Health and Chronic Diseases. World Health Organization, 2006.

[9] Who, world health organization. human papillomavirus (hpv) and cervical cancer, fact sheet. https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer, February 2022.

[10] National cancer institute. cervical cancer treatment. 2018. https://www.cancer.gov/types/cervical/patient/cervical-treatment-pdq, February 2022.

[11] American cancer society. hpv test. 2017. https://www.cancer.org/cancer/cervical-cancer/detection-diagnosis-staging/screening-tests/hpv-test.html, February 2022.

[12] Urmila Banik, Pradip Bhattacharjee, Shahab Uddin Ahamad, and Zillur Rahman. Pattern of epithelial cell abnormality in pap smear: A clinicopathological and demographic correlation. *Cytojournal*, 8, 2011.

[13] Nguyen Vu Quoc Huy, Ngo Viet Quynh Tram, Dang Cong Thuan, Truong Quang Vinh, Cao Ngoc Thanh, Linus Chuang, et al. The value of visual inspection with acetic acid and pap smear in cervical cancer screening program in low resource settings–a population-based study. *Gynecologic oncology reports*, 24:18–20, 2018.

[14] Fatemeh Haghighi, Nahid Ghanbarzadeh, Marziee Ataee, Gholamreza Sharifzadeh, Javid Shahbazi Mojarrad, and Fatemeh Najafi-Semnani. A comparison of liquid-based cytology with conventional papanicolaou smears in cervical dysplasia diagnosis. *Advanced biomedical research*, 5, 2016.

[15] Vikrant Bhar Singh, Nalini Gupta, Raje Nijhawan, Radhika Srinivasan, Vanita Suri, Arvind Rajwanshi, et al. Liquid-based cytology versus conventional cytology for evaluation of cervical pap smears: experience from the first 1000 split samples. *Indian Journal of Pathology and Microbiology*, 58(1):17, 2015.

[16] Ritu Nayar and David C Wilbur. The pap test and bethesda 2014. *Acta cytologica*, 59(2):121–132, 2015.

[17] Ritu Nayar and David C Wilbur. *The Bethesda system for reporting cervical cytology: definitions, criteria, and explanatory notes*. Springer, 2015.

[18] RH Young. Who classification of tumours of female reproductive organs. 2014.

[19] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[20] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. Explainable software analytics. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, pages 53–56, 2018.

[21] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[22] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.

[23] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*, pages 801–810. IEEE, 2007.

[24] Taehyun Ha, Sangwon Lee, and Sangyeon Kim. Designing explainability of an artificial intelligence system. In *Proceedings of the Technology, Mind, and Society*, pages 1–1. 2018.

[25] C. Molnar. *Interpretable Machine Learning*. Lulu.com, 2020.

[26] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[28] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[29] 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.

[30] RL Cahn, RS Poulsen, and G Toussaint. Segmentation of cervical cell images. *Journal of Histochemistry & Cytochemistry*, 25(7):681–688, 1977.

[31] Afaf Tareef, Yang Song, Heng Huang, Dagan Feng, Mei Chen, Yue Wang, and Weidong Cai. Multi-pass fast watershed for accurate segmentation of overlapping cervical cells. *IEEE transactions on medical imaging*, 37(9):2044–2059, 2018.

[32] Hady Ahmady Phoulady, Dmitry Goldgof, Lawrence O Hall, and Peter R Mouton. A framework for nucleus and overlapping cytoplasm segmentation in cervical cytology extended depth of field and volume images. *Computerized Medical Imaging and Graphics*, 59:38–49, 2017.

[33] Zhi Lu, Gustavo Carneiro, and Andrew P Bradley. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. *IEEE Transactions on Image Processing*, 24(4):1261–1272, 2015.

[34] Afaf Tareef, Yang Song, Weidong Cai, Heng Huang, Hang Chang, Yue Wang, Michael Fulham, Dagan Feng, and Mei Chen. Automatic segmentation of overlapping cervical smear cells based on local distinctive features and guided shape deformation. *Neurocomputing*, 221:94–107, 2017.

[35] Ling Zhang, Le Lu, Isabella Nogues, Ronald M Summers, Shaoxiong Liu, and Jianhua Yao. Deeppap: deep convolutional networks for cervical cell classification. *IEEE journal of biomedical and health informatics*, 21(6):1633–1643, 2017.

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[37] Srishti Gautam, Nirmal Jith, Anil K Sao, Arnav Bhavsar, Adarsh Natarajan, et al. Considerations for a pap smear image analysis system with cnn features. *arXiv preprint arXiv:1806.09025*, 2018.

[38] OU Nirmal Jith, KK Harinarayanan, Srishti Gautam, Arnav Bhavsar, and Anil K Sao. Deepcerv: Deep neural network for segmentation free robust cervical cell classification. In *Computational Pathology and Ophthalmic Medical Image Analysis*, pages 86–94. Springer, 2018.

[39] Meng Zhao, Aiguo Wu, Jingjing Song, Xuguo Sun, and Na Dong. Automatic screening of cervical cells using block image processing. *Biomedical engineering online*, 15(1):1–20, 2016.

[40] A Dongyao Jia, B Zhengyi Li, and C Chuanwang Zhang. Detection of cervical cancer cells based on strong feature cnn-svm network. *Neurocomputing*, 411:112–127, 2020.

[41] Yiming Zhang, Ying Weng, and Jonathan Lund. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, 12(2), 2022.

[42] Shaker El-Sappagh, Jose M Alonso, SM Islam, Ahmad M Sultan, and Kyung Sup Kwak. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease. *Scientific reports*, 11(1):1–26, 2021.

[43] Joe Huei Ong, Kam Meng Goh, and Li Li Lim. Comparative analysis of explainable artificial intelligence for covid-19 diagnosis on cxr image. In *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 185–190, 2021.

[44] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016.

[45] Matteo Rucco, Giovanna Viticchi, and Lorenzo Falsetti. Towards personalized diagnosis of glioblastoma in fluid-attenuated inversion recovery (flair) by topological interpretable machine learning. *Mathematics*, 8(5), 2020.

[46] F. Cavaliere, A. Della Cioppa, A. Marcelli, A. Parziale, and R. Senatore. Parkinson's disease diagnosis: Towards grammar-based explainable artificial intelligence. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6, 2020.

[47] Nicola Amoroso, Domenico Pomarico, Annarita Fanizzi, Vittorio Didonna, Francesco Giotta, Daniele La Forgia, Agnese Latorre, Alfonso Monaco, Ester Pantaleo, Nicole Petruzzellis, Pasquale Tamborra, Alfredo Zito, Vito Lorusso, Roberto Bellotti, and Raffaella Massafra. A roadmap towards breast cancer therapies supported by explainable artificial intelligence. *Applied Sciences*, 11(11), 2021.

[48] Ramisetty Kavya, Jabez Christopher, Subhrakanta Panda, and Y. Bakthasingh Lazarus. Machine learning and xai approaches for allergy diagnosis. *Biomedical Signal Processing and Control*, 69:102681, 2021.

[49] Theodore Evans, Carl Orge Retzlaff, Christian Geißler, Michaela Kargl, Markus Plass, Heimo Müller, Tim-Rasmus Kiehl, Norman Zerbe, and Andreas Holzinger. The explainability paradox: Challenges for xai in digital pathology. *Future Generation Computer Systems*, 133:281–296, 2022.

[50] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 258–262, 2019.

[51] Ruth MJ Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, pages 6276–6282, 2019.

[52] Devam Dave, Het Naik, Smiti Singhal, and Pankesh Patel. Explainable ai meets healthcare: A study on heart disease dataset. *arXiv preprint arXiv:2011.03195*, 2020.

[53] Oscar Odestål and Anna Palmqvist Sjövall. Adaptive reference images for blood cells using variational autoencoders and self-organizing maps. *Master's Theses in Mathematical Sciences*, 2020.

[54] Haoran Wu, Wei Chen, Shuang Xu, and Bo Xu. Counterfactual supporting facts extraction for explainable medical record based diagnosis with graph network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1942–1955, Online, June 2021. Association for Computational Linguistics.

[55] Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. Explainable ai in industry. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3203–3204, 2019.

[56] Anjali Khurana, Parsa Alamzadeh, and Parmit K Chilana. Chatrex: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–11. IEEE, 2021.

[57] Eoin M Kenny and Mark T Keane. Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in xai. *Knowledge-Based Systems*, 233:107530, 2021.

[58] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[59] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[60] 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.

[61] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[62] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[63] Mario Köppen. The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pages 4–8, 2000.

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[65] Variational autoencoder svhn. https://github.com/bjlkeng/sandbox/blob/master/notebooks/variational_autoencoder-svhn/model_fit.ipynb, June 2022.

[66] K Kavitha and B Thirumala Rao. Evaluation of distance measures for feature based image registration using alexnet. *arXiv preprint arXiv:1907.12921*, 2019.

[67] Werner GK Backhaus, Reinhold Kliegl, and John S Werner. *Color vision: Perspectives from different disciplines*. Walter de Gruyter, 2011.

Appendix A

Model Architectures

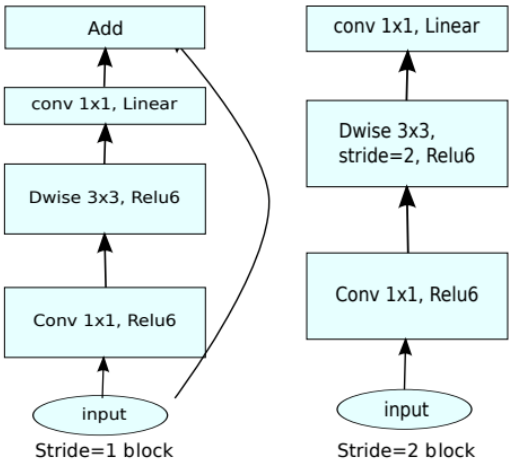

Figure A.1: Mobilenet v2 model architecture

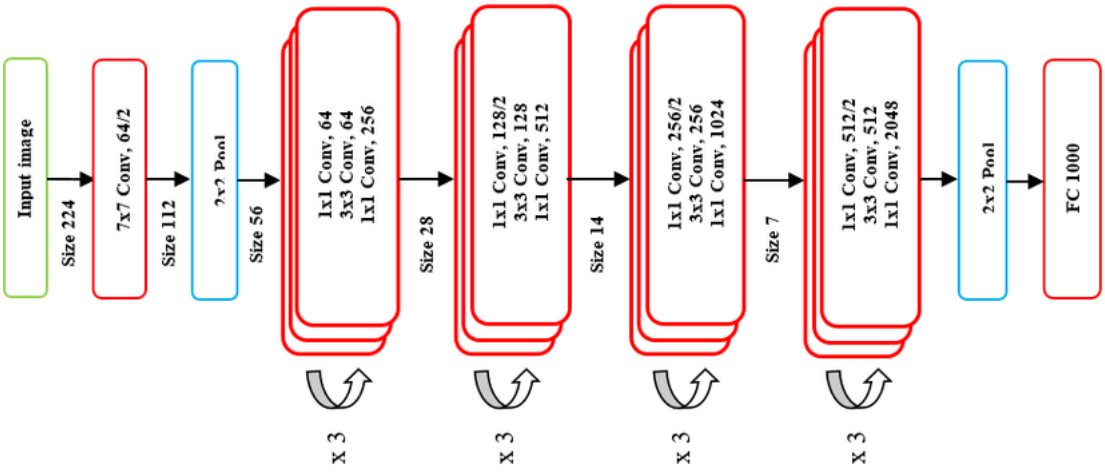

Figure A.2: Resnet50 model architecture

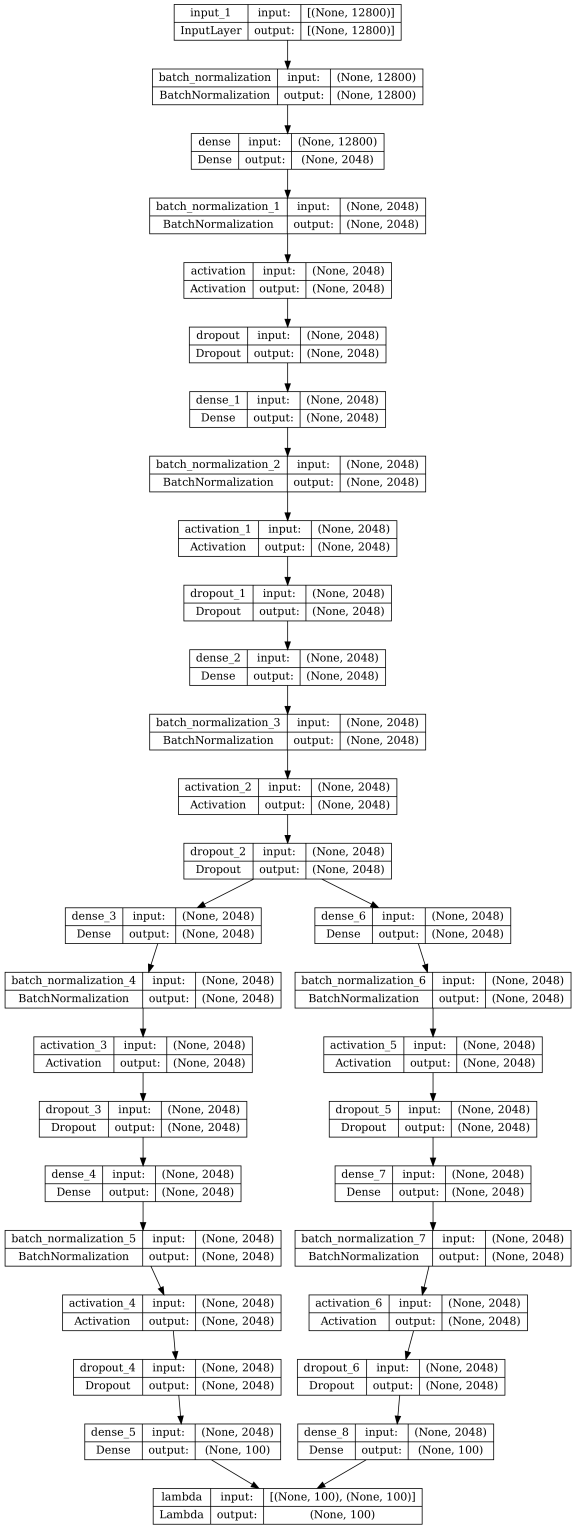

Figure A.3: VAE Encoder architecture

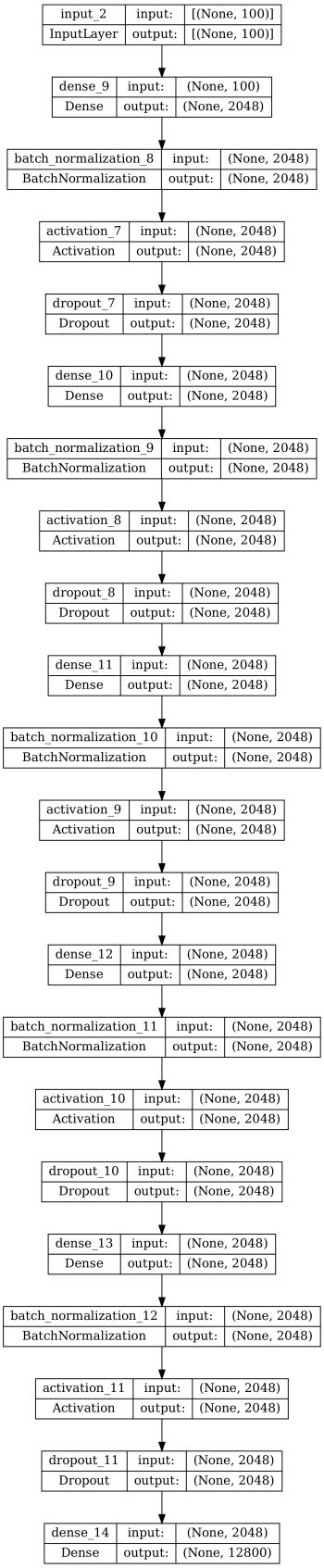

Figure A.4: VAE Decoder architecture

Appendix B

Loss plots from VAE training

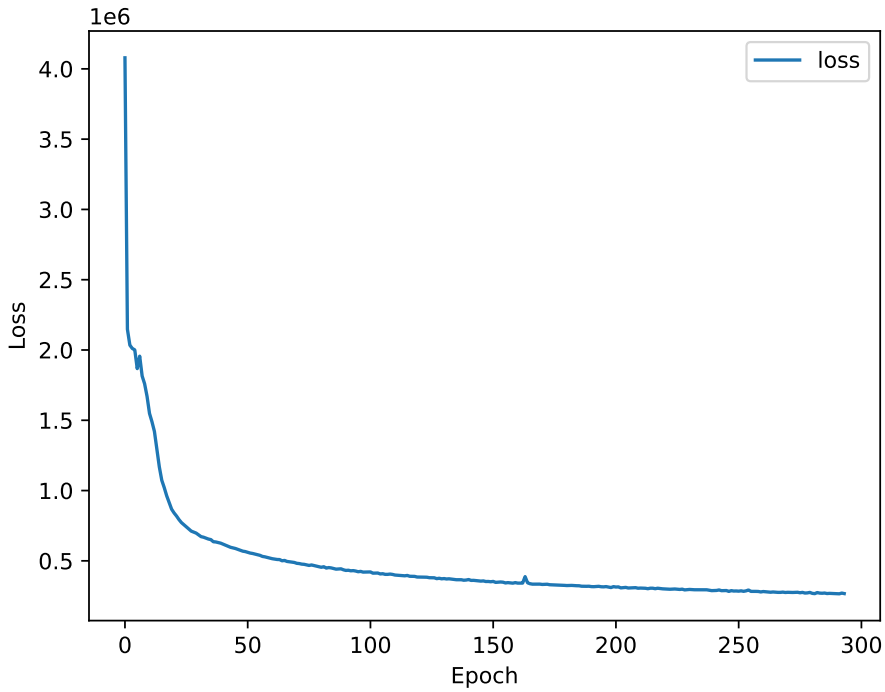

Figure B.1: Training loss for VAE training, using the region detection model

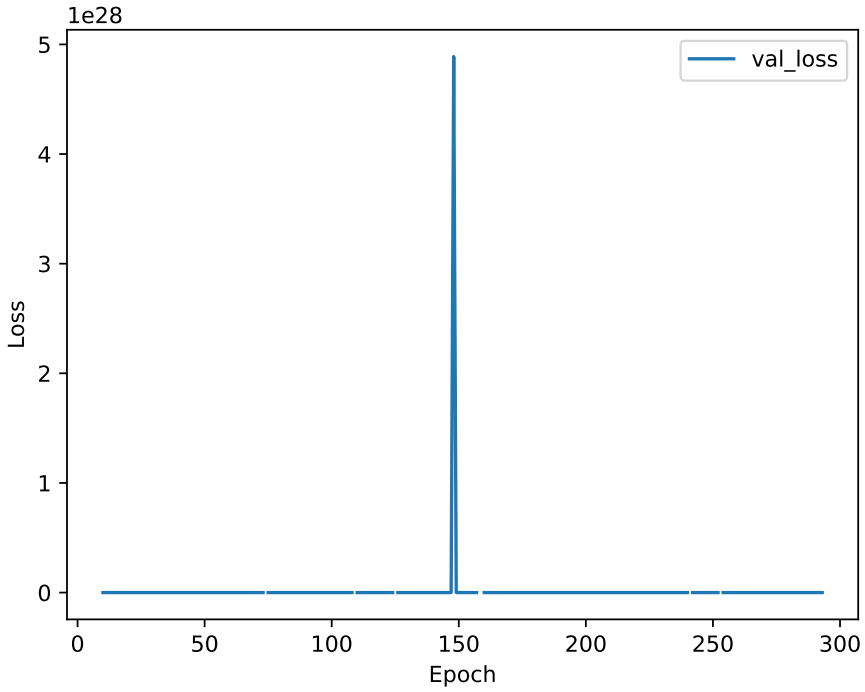

Figure B.2: Validation loss for VAE training, using the region detection model

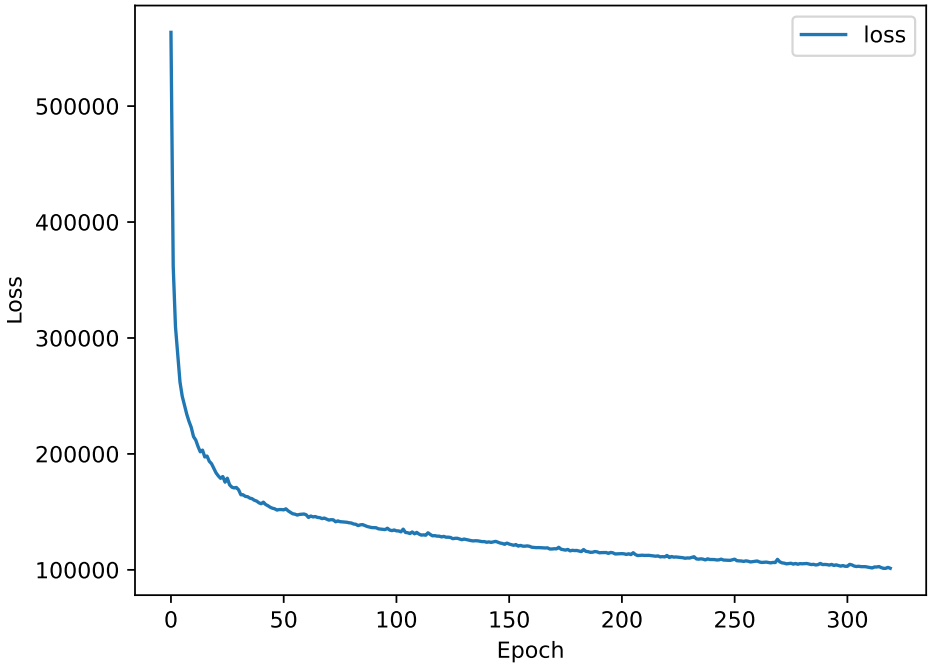

Figure B.3: Training loss for VAE training, using the nuclei detection model

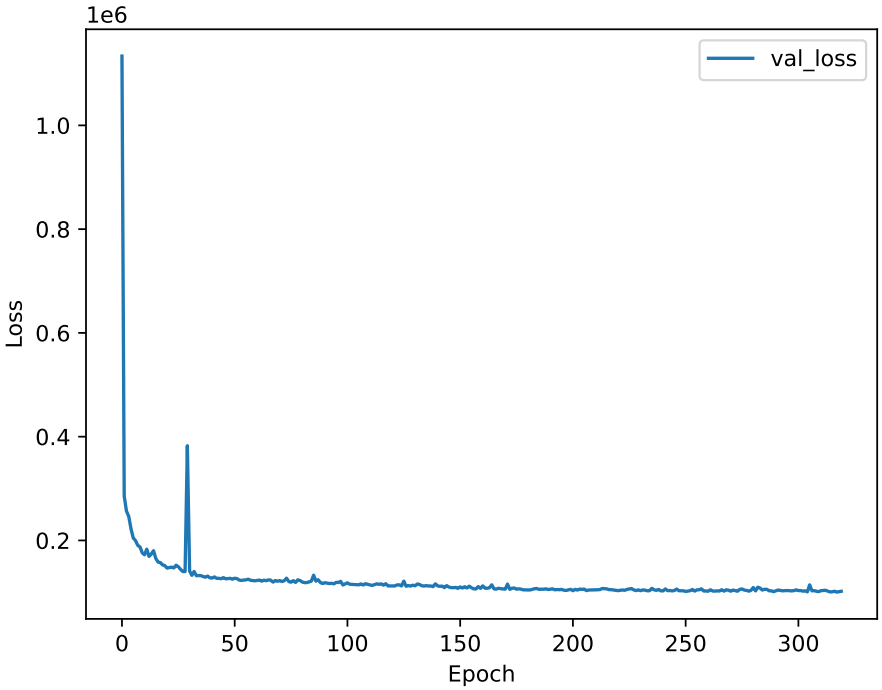

Figure B.4: Validation loss for VAE training, using the nuclei detection model

Appendix C

Additional results between Cosine and Manhattan distances, for the use of PCA

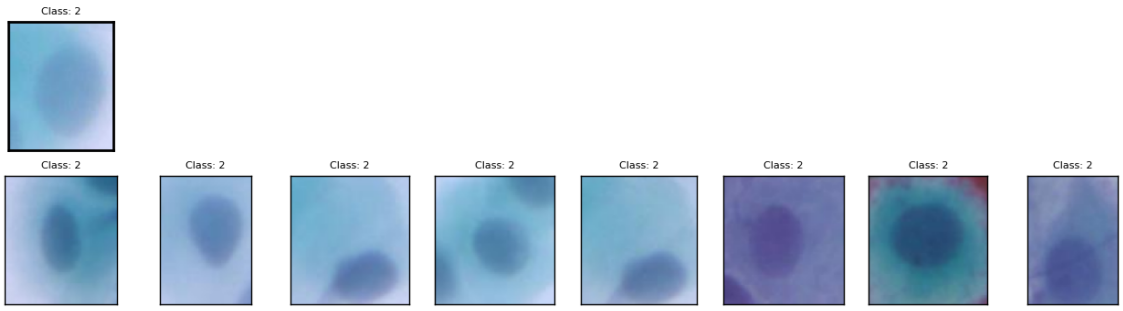

Figure C.1: Results of PCA and Cosine distance for example cell 1

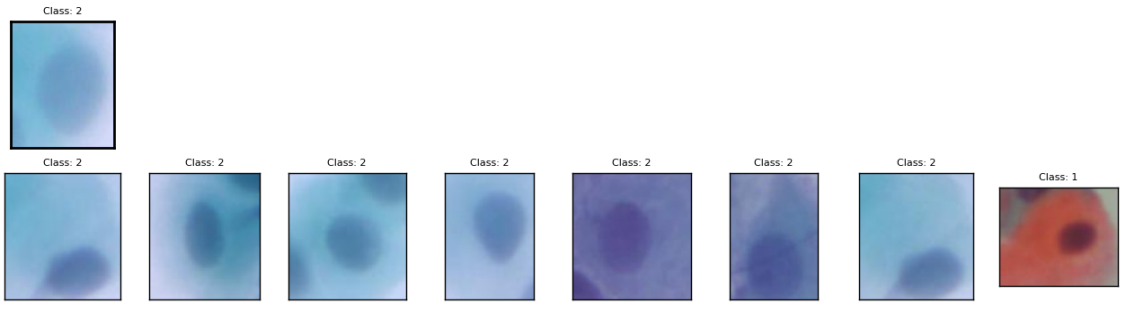

Figure C.2: Results of PCA and Manhattan distance for example cell 1

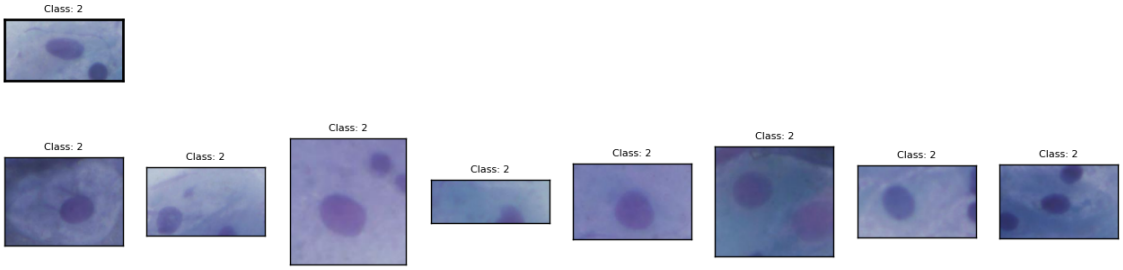

Figure C.3: Results of PCA and Cosine distance for example cell 2

Figure C.4: Results of PCA and Manhattan distance for example cell 2

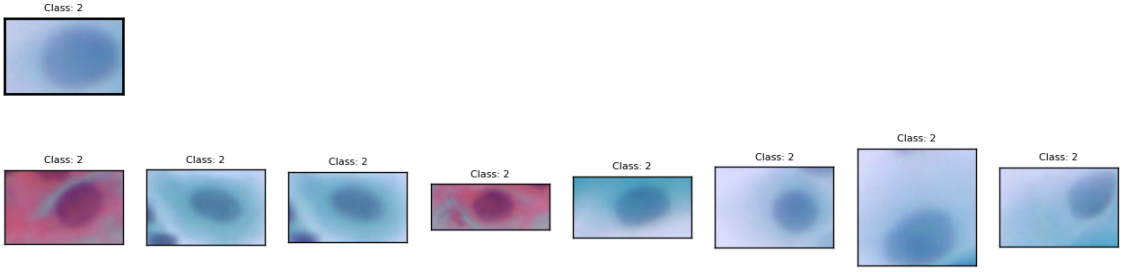

Figure C.5: Results of PCA and Cosine distance for example cell 3

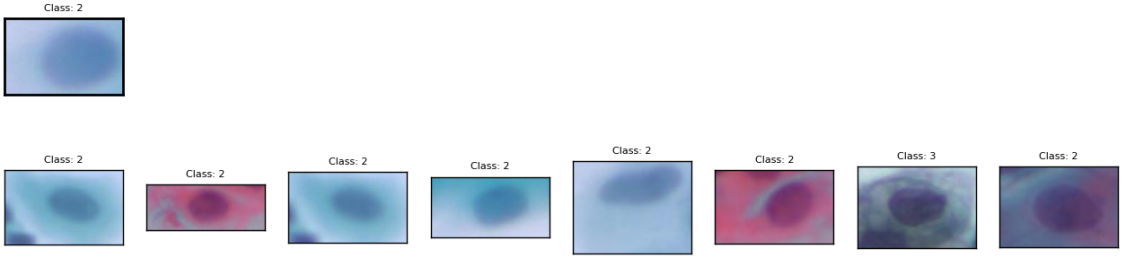

Figure C.6: Results of PCA and Manhattan distance for example cell 3

Appendix D

Additional results between PCA and VAE, for the region detection model

Figure D.1: Results of PCA and Cosine distance for example cell 1

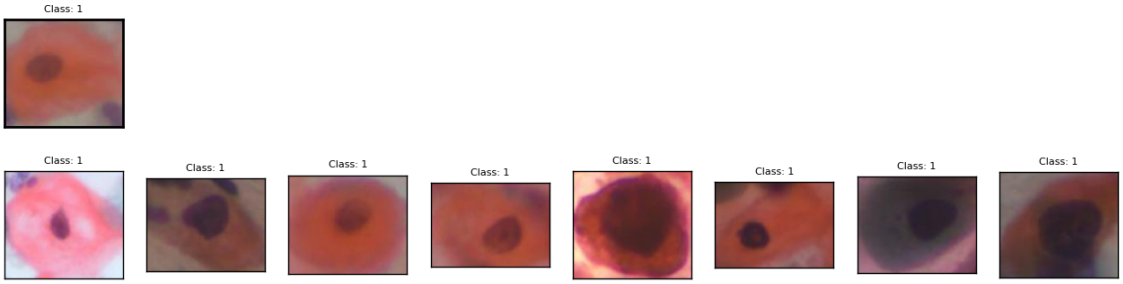

Figure D.2: Results of VAE and Manhattan distance for example cell 1

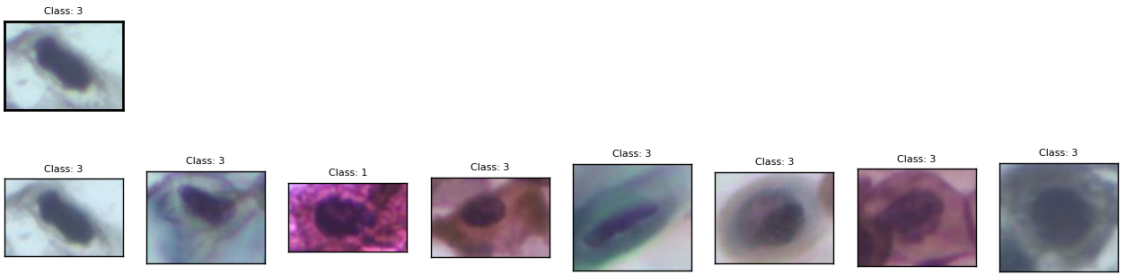

Figure D.3: Results of PCA and Cosine distance for example cell 2

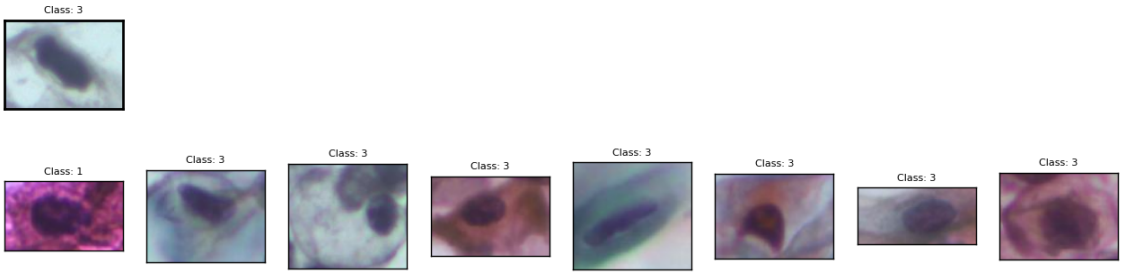

Figure D.4: Results of VAE and Manhattan distance for example cell 2

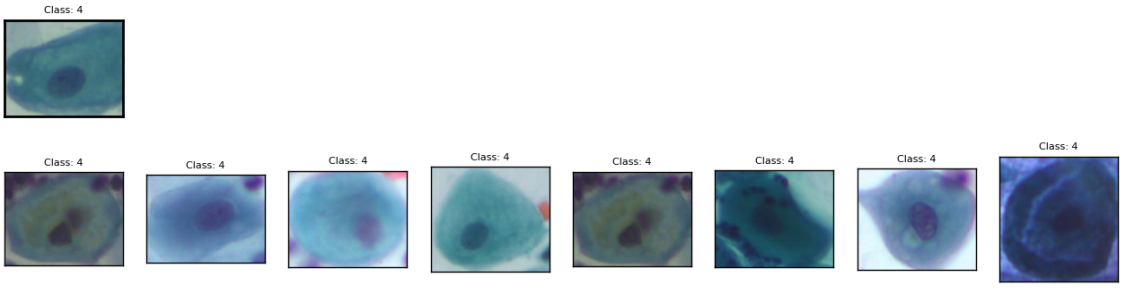

Figure D.5: Results of PCA and Cosine distance for example cell 3

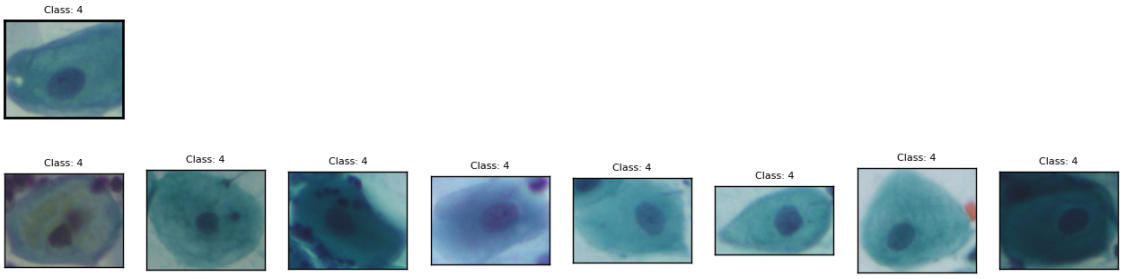

Figure D.6: Results of VAE and Manhattan distance for example cell 3

Appendix E

Additional results between Euclidean and Chebyshev distances, for the use of PCA

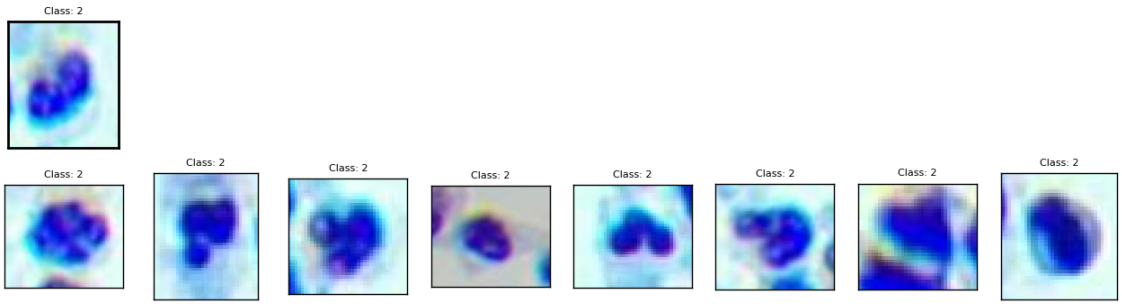

Figure E.1: Results of PCA and Euclidean distance for example cell 1

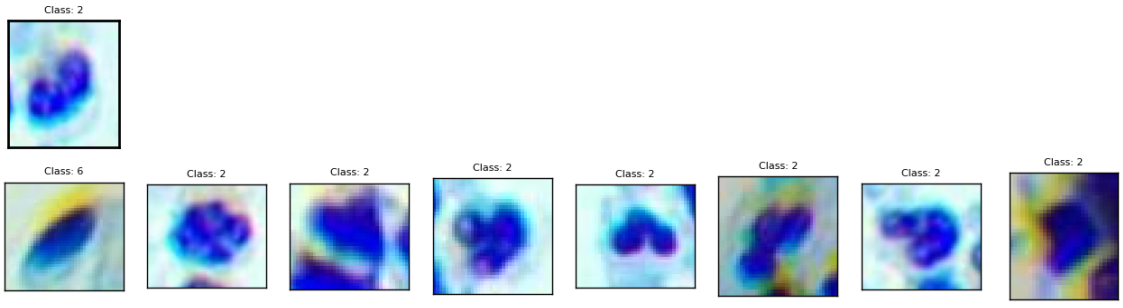

Figure E.2: Results of PCA and Chebyshev distance for example cell 1

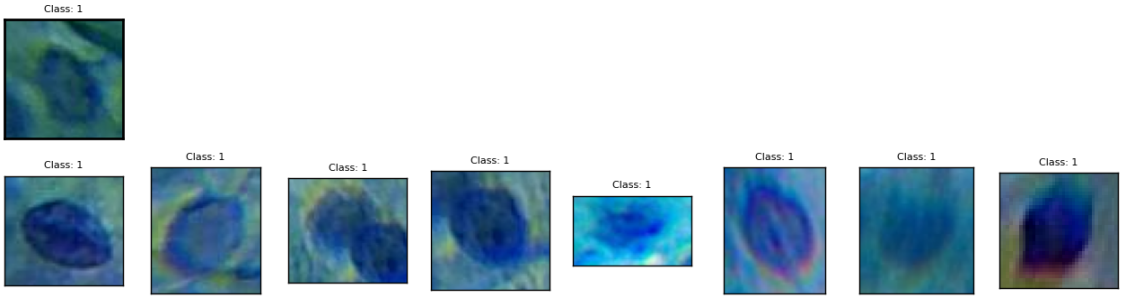

Figure E.3: Results of PCA and Euclidean distance for example cell 2

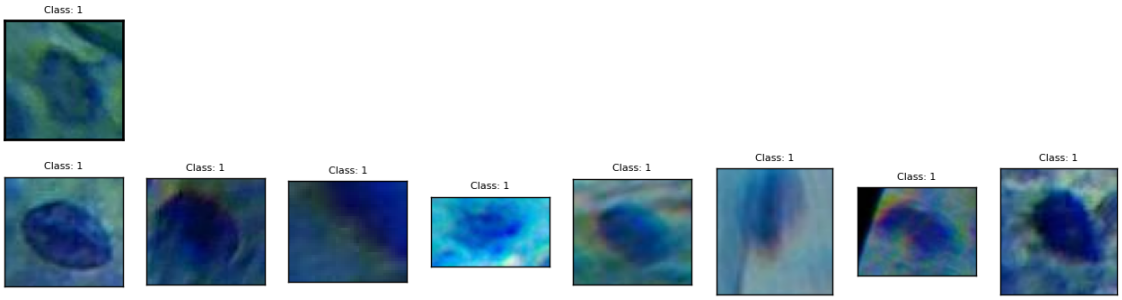

Figure E.4: Results of Chebyshev distance for example cell 2

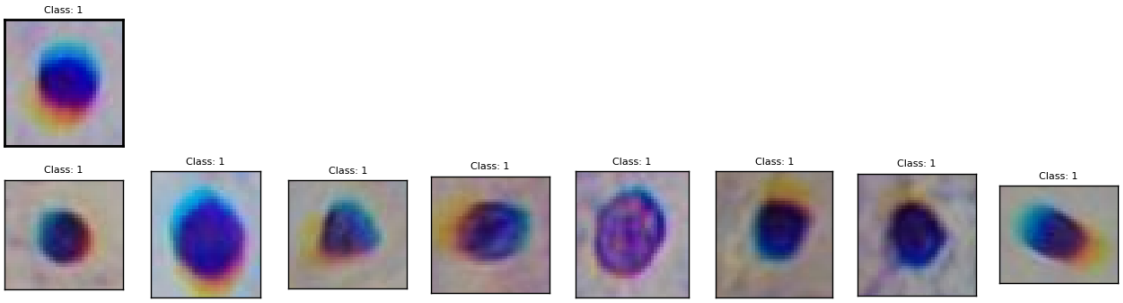

Figure E.5: Results of PCA and Euclidean distance for example cell 3

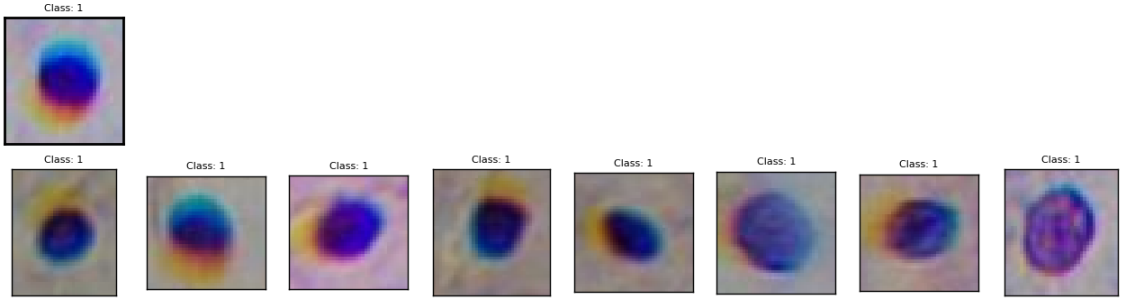

Figure E.6: Results of PCA and Chebyshev distance for example cell 3

Appendix F

Additional results between PCA and VAE, for the nuclei detection model

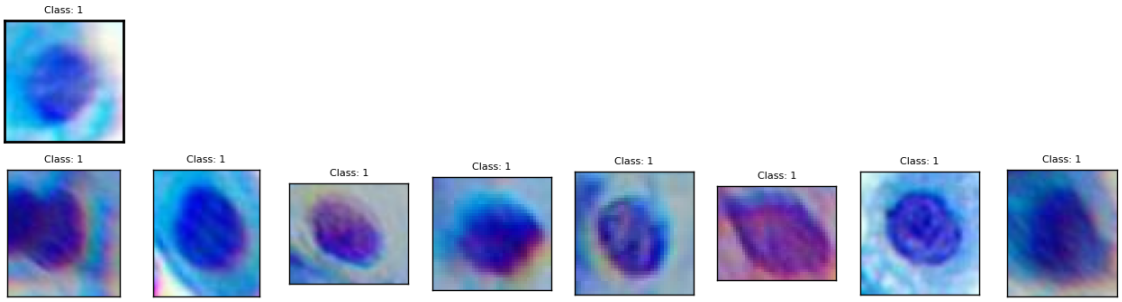

Figure F.1: Results of PCA and Euclidean distance for example cell 1

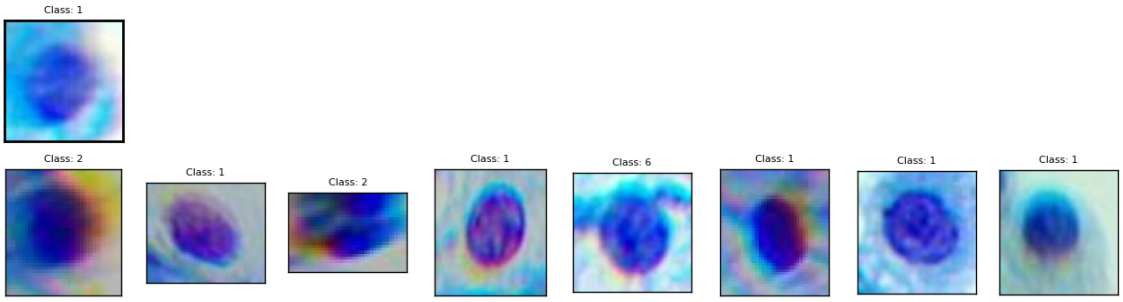

Figure F.2: Results of VAE and Cosine distance for example cell 1

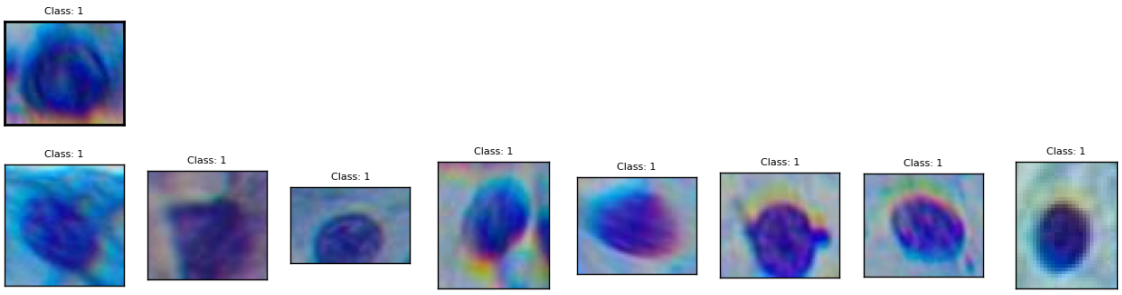

Figure F.3: Results of PCA and Euclidean distance for example cell 2

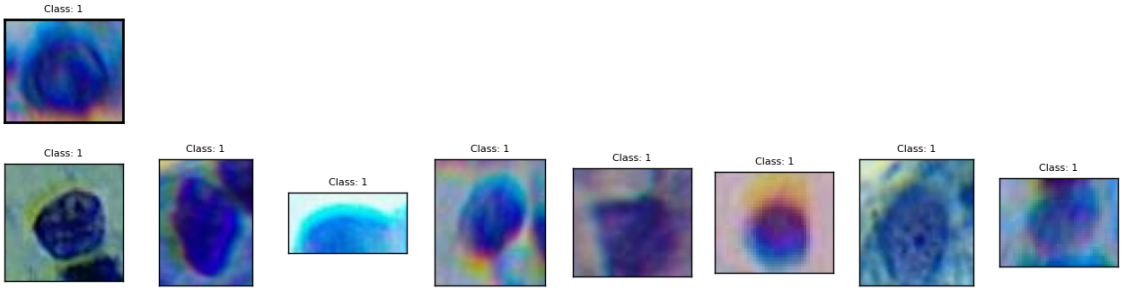

Figure F.4: Results of VAE and Cosine distance for example cell 2

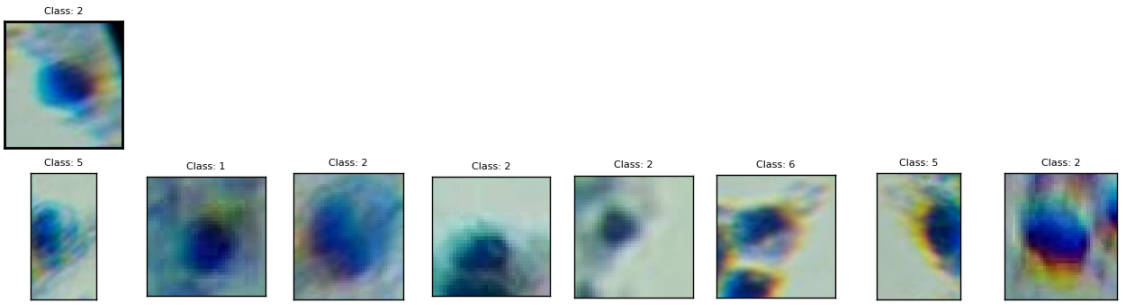

Figure F.5: Results of PCA and Euclidean distance for example cell 3

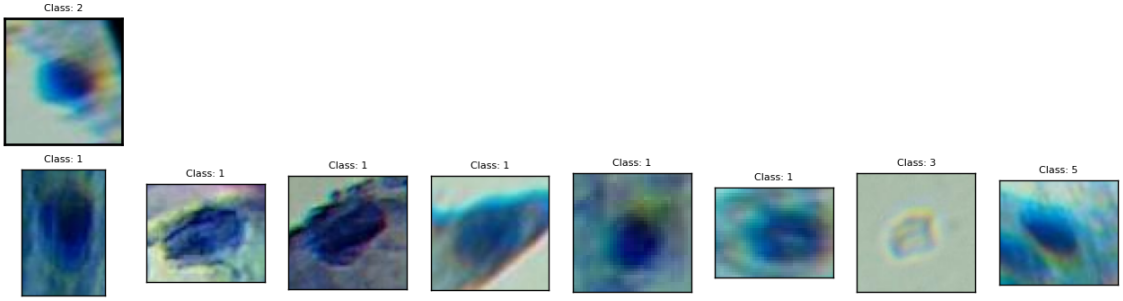

Figure F.6: Results of VAE and Cosine distance for example cell 3