

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Combining machine learning and deep learning approaches to detect cervical cancer in cytology images

Eduardo Luís Pinheiro da Silva



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Luís Filipe Teixeira

Second Supervisor: Ana Sampaio

Third Supervisor: Maria Vasconcelos

July 21, 2021

Combining machine learning and deep learning approaches to detect cervical cancer in cytology images

Eduardo Luís Pinheiro da Silva

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Rui Camacho

External Examiner: Pedro Vieira

Supervisor: Luís Filipe Teixeira

July 21, 2021

Abstract

Cervical cancer is the fourth most common cancer in women. When diagnosed early on, it is one of the most successfully treatable types of cancer. As such, screening tests are very effective as a prevention measure. These tests involve the analysis of microscopic fields of cytology samples, which, when performed manually, is a very demanding task, requiring highly specialized laboratory technologists (cytotechs). Due to this, there has been a great interest in automating the overall screening process. Most of these computer-aided diagnosis systems subject the images from each sample to a set of steps, more notably focus and adequacy assessment, region of interest identification and respective classification. This work is focused on the last two stages, more specifically, the detection of abnormal regions and the classification of their abnormality level. The main approaches can be divided into two types: deep learning architectures and conventional machine learning models, both presenting their own set of advantages and disadvantages. This work explores the combination of both of these approaches in hybrid pipelines to minimize the drawbacks of each one whilst taking advantage of the best they have to offer, ultimately contributing to a decision support system for cervical cancer diagnosis. More specifically, a deep learning approach is proposed for the detection of the regions of interest and respective bounding box generation based on the RetinaNet architecture, followed by the extraction of multiple features from each region and their classification through two SVM models. Additionally, the nuclei in each detected region are also segmented through an iterative thresholding algorithm, providing a set of nuclei-specific features, whose impact in the classification result is studied, motivated by the relevance of the nuclei properties in the reasoning of clinical experts. Besides the individual assessment of each module and the evaluation of the entire system, a comparative analysis of different algorithms is also performed, aiming to support future research of similar solutions. The system achieved a precision, recall and F1 score of 0.20, 0.40 and 0.27, respectively, indicating that additional strategies need to be explored to overcome its main limitations.

Keywords

cervical cancer, machine learning, deep learning, detection, classification

Categories

- CCS → Computing Methodologies → Artificial Intelligence → Computer Vision → Computer Vision Problems → Object Detection
- CCS → Computing Methodologies → Machine Learning → Learning Paradigms → Supervised Learning → Supervised Learning by Classification

Acknowledgements

I would like to express my sincerest gratitude to all of those who supported me throughout this journey, not only for the past few months but over the course of these five years as well, particularly:

My advisor and teacher, professor Luís Teixeira, for his guidance through my master's, who always strived to provide the best education possible, even in the face of the present circumstances we endure.

My co-advisor, Ana Sampaio, whose continuous support was invaluable to this work. I truly am thankful for all her help and direction throughout the past year and all the knowledge shared during that time.

My family, especially my mom, for her unconditional love and motivation, my dad, for his unfaltering wisdom and for pushing me to give my absolute best, and my sister, for always believing in me and "pulling" me back to reality during the most challenging times.

Lastly, I would also like to thank Maria Vasconcelos and everyone else at Fraunhofer Portugal for taking me in and guiding me through my final months as a student, as well as for the availability and professionalism demonstrated during this period.

This work was done under the scope of “CLARE: Computer-Aided Cervical Cancer Screening”, project with reference POCI-01-0145-FEDER028857 and financially supported by FEDER through Operational Competitiveness Program – COMPETE 2020 and by National Funds through Foundation for Science and Technology FCT/MCTES.

Eduardo Luís Pinheiro da Silva

*“Uncertainty is an uncomfortable position.
But certainty is an absurd one.”*

Voltaire

Contents

1	Introduction	1
1.1	Document Structure	1
1.2	Context	2
1.2.1	Cervical Cancer	2
1.2.2	Cervical Cancer Screening	2
1.2.2.1	Pap Smear Test	2
1.2.2.2	Computer-Aided Diagnosis Systems	2
1.3	Objectives	3
2	Related Work	5
2.1	Region of Interest Detection	5
2.1.1	Conventional Machine Learning	5
2.1.2	Deep Learning	6
2.2	Classification	10
2.2.1	Deep Learning	10
2.2.2	Conventional Machine Learning	12
2.3	Hybrid Approaches	12
2.4	Summary	15
3	Proposed Method	21
3.1	Datasets	21
3.1.1	Data Preparation	24
3.2	Experimental Pipeline	27
3.2.1	Pre-processing	28
3.2.2	Abnormal Region Detection	29
3.2.3	Nuclei Detection	33
3.2.3.1	Pre-processing	33
3.2.3.2	Segmentation	34
3.2.3.3	Post-processing	37
3.2.4	Feature Extraction	37
3.2.4.1	Colour Features	38
3.2.4.2	Texture Features	39
3.2.4.3	Geometrical Features	40
3.2.5	Classification	41
4	Experimental Setup and Results	45
4.1	Abnormal Region Detection	45
4.1.1	Evaluation Metrics	45

4.1.2	RetinaNet with ResNet50 backbone	46
4.1.3	Summary	48
4.2	Classification	50
4.2.1	Evaluation Metrics	50
4.2.2	Random Forest	50
4.2.3	SVM	52
4.2.4	Summary	53
4.3	Complete System	55
5	Conclusions	59
5.1	Limitations and Future Work	60
A	Abnormal Region Detection	63
A.1	Initial Experiments	63
B	Nuclei Detection	67
B.1	Alternative Approaches	67
C	Classification	69
C.1	Feature Importance	69
	References	73

List of Figures

2.1	Example of detected objects by Zhout et al. [16]. The red rectangles represent the ground-truth bounding boxes, while the blue ones represent the detection network’s results.	8
2.2	Example of nucleus centre detection by Prum et al. [27]. (a) Input image, (b) binary image, (c) detected nucleus edges, (d) detected nucleus center point, (e) cropped image.	13
2.3	Classification decision tree structure by Gautam et al. [32].	15
3.1	Source image example acquired through a μ SmartScope.	22
3.2	Lesion type examples.	22
3.3	Inadequate annotation example.	24
3.4	Overall training subset’s class distribution before balancing.	25
3.5	Horizontal translation example.	26
3.6	Similar examples from the HSIL and SCC classes.	26
3.7	Overall training subset’s final class distribution after balancing.	27
3.8	Dataset’s empty-annotated image distribution after empty image down-sampling.	27
3.9	Proposed system’s architecture.	28
3.10	Split regions with (a) moderate and (b) insufficient information for detection.	29
3.11	Sum of squared distances (scales and aspect ratios) and average absolute difference per cluster (aspect ratios) for the 320×320 patches.	31
3.12	ROI before (a) and after (b) automatic brightness and contrast adjustment.	33
3.13	Contrast adjustment through histogram stretching.	34
3.14	ROI before (a) and after (b) binarization through the iterative thresholding algorithm (the nucleus is represented in white).	36
3.15	Iterative thresholding results with (b, d) and without (a, c) additional solidity constant.	36
3.16	Example of a segmented nucleus.	37
3.17	Visual representation of the CIELch colour space [43].	38
3.18	Example of different colours between classes.	38
3.19	Example of similar colours between classes.	39
3.20	Example of rescaled and padded image (a) and its HOG representation (b).	40
3.21	Classification Tree.	42
4.1	Intersection over union (IoU) representation [47].	45
4.2	RetinaNet with ResNet50 backbone R1 experiments (green and grey for folds one and two, respectively) and test set evaluation (red) metrics.	47
4.3	Misidentified normal region with enlarged nucleus.	49

4.4	Classification loss for the RetinaNet with ResNet50 backbone R1 experiments (green and grey for folds 1 and 2, respectively), train (blue) and test set evaluation (red).	49
4.5	Random forests' test set confusion matrices.	52
4.6	SVM's test set confusion matrices.	55
A.1	RetinaNet with MobileNetV1 backbone first experiment (a) mAP@0.5IoU and (b) AR@10.	64
A.2	Excessive overlapping bounding boxes when using an IoU threshold of 0.6.	65
B.1	Contrast adjustment techniques.	68
C.1	Random forests' top five most important features.	70
C.2	Nuclei mean energy (chroma channel) box plot (overall train set).	70
C.3	Random forests' most important texture features box plots (overall train set).	70
C.4	Bounding box area box plot (overall train set).	71

List of Tables

2.1	Summary table with highlighted works for the ROI detection problem. Prec (Precision), Rec (Recall), F1 (F1-measure).	9
2.2	Summary table with highlighted works on cervical cell classification. When a large number of experiments was conducted, performance is only given for the best case. Acc (Accuracy), Prec (Precision), Sens (Sensitivity), Sp (Specificity). .	18
2.3	Summary table with highlighted works on hybrid approaches. When a large number of experiments was conducted, performance is only given for the best case. Acc (Accuracy), Prec (Precision), Rec (Recall), Sens (Sensitivity), Sp (Specificity).	19
3.1	HFF dataset class distribution.	23
3.2	Nuclei dataset class distribution.	23
3.3	Deep learning hyper-parameter random search grid.	30
3.4	Random forest random search grid, with n being the total number of samples in the training dataset and f the total number of features.	43
3.5	SVM grid search hyper-parameter values, with f and $variance_N$ being the total number of features and variance of the fitted data, respectively.	44
4.1	RetinaNet with ResNet50 backbone experiments. Experiment (Exp), Learning Rate (LR), Warm Up (WU), Detections (Detect).	46
4.2	RetinaNet with ResNet50 backbone experiments' 3-fold cross-validation results. Experiment (Exp), Average (Avg), Standard Deviation (Standard Dev).	47
4.3	Random forest's random searches best parameters, with n and f being the total number of samples and features, respectively. Experiment (Exp), Feats (Features), Min Samples Leaf (MSL), Min Samples Split (MSS).	51
4.4	Random forest's random searches cross-validation results. Experiment (Exp). . .	51
4.5	Random forests' results on the test set.	51
4.6	SVM's grid searches best parameters, with f and $variance_N$ being the total number of features and variance of the fitted data, respectively. Experiment (Exp), Features (Feats).	53
4.7	SVM's grid searches cross-validation results. Experiment (Exp).	54
4.8	SVMs' results on the test set. Hyper-parameter (HP) combination (combo). . . .	54
4.9	System test results per image hierarchy level.	56
4.10	Sample labels predicted by the system (test subset) and respective ground truth values.	56
4.11	Detection module test results per image hierarchy level.	57
4.12	System test results per image hierarchy level (when considering examples only correctly identified by the detection module).	57

A.1 RetinaNet with MobileNetV1 backbone experiments performed on the first fold's validation set. Experiment (Exp), Warm up (WU), Learning Rate (LR), Detections (Detect), Constant (const). 64

Abbreviations

AR	Average Recall
ASC-H	Atypical Squamous Cells, Cannot Rule Out High Grade Squamous Intraepithelial Lesion
ASC-US	Atypical Squamous Cells of Undetermined Significance
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
GLCM	Grey-Level Co-Occurrence Matrix
HOG	Histograms of Oriented Gradients
HSIL	High Grade Squamous Intraepithelial Lesion
IoU	Intersection over Union
LSIL	Low Grade Squamous Intraepithelial Lesion
mAP	Mean Average Precision
NMS	Non-Max-Suppression
ROI	Region of Interest
SCC	Squamous Cell Carcinoma

Chapter 1

Introduction

The high incidence rate of cervical cancer in women [1] has propelled research of effective screening methods significantly over the last century. Due to this task's laborious nature when performed manually, automated systems have gained a particular interest in recent decades. However, these systems still face some challenges, especially related to the algorithmic approaches employed. These can be based on deep learning architectures, known for their good performance but also their need for large amounts of training data, or other more conventional machine learning techniques, commonly associated with highly interpretable models but more modest results. In order to overcome some of these problems, this document explores the combination of both types of algorithms in hybrid pipelines for the detection and classification of abnormal regions in liquid-based cytology images. While hybrid approaches have been investigated before in similar settings, most of the studies falling in this category do not explicitly employ both types of algorithms for the detection and classification stages, using them instead for related tasks, such as feature extraction. In this manner, the present work provides a fresh perspective in terms of the architecture of computer-aided diagnosis (CAD) systems, ultimately contributing to the artificial intelligence component of the CLARE project [2].

1.1 Document Structure

In the following section, some context is given regarding cervical cancer and respective screening, focusing on the pap smear test and computer-based systems aimed at supporting the diagnosis process. In Chapter 2, a literary review of the state-of-the-art methods for the region of interest (ROI) identification and classification in similar settings is conducted, exploring deep learning and conventional machine learning methods for both stages, as well as hybrid approaches. Following this, the proposed solution for the ROI detection, feature extraction and classification tasks is presented in Chapter 3, along with a strategy for the segmentation of the nuclei in these regions. The datasets used for training and evaluation purposes are also detailed in this section. In Chapter

4, the complete system's results are discussed in detail, as well as those of each individual module. Finally, in Chapter 5, some conclusions are drawn up together with some remarks about future work and possible system improvements.

1.2 Context

1.2.1 Cervical Cancer

Cervical cancer is the fourth most common type of cancer in women with 99% of the reported cases being linked to infections with high-risk human papillomaviruses [1]. As such, it comes as no surprise that one of the most efficient prevention methods is vaccination for this type of virus. However, this is not always a possibility, especially in lower resource countries. Due to this, screening methods emerge as alternative (secondary) prevention measures. The importance of screenings is also reinforced by the fact that cervical cancer is one of the most treatable types of cancer **when detected early on** [1], having a 92% 5-year survival rate when in an early invasive stage [3]. Treatment for detected precancerous cells is usually simpler though, with options ranging from simple monitoring to laser therapy and cryosurgery or, in more serious cases, removal surgery [4].

1.2.2 Cervical Cancer Screening

The main screening methods for cervical cancer are HPV testing and cervical cytology. This section will focus on the latter, specifically on the Papanicolaou (or Pap) test and its variations. Furthermore, a brief context on CAD systems directed towards these tests is also given.

1.2.2.1 Pap Smear Test

The origin of the classical pap smear test is attributed to George N. Papanicolaou, being first suggested in 1928 but only beginning to be adopted during the 1940s [5]. It consists on the microscopical analysis of cells scraped from the squamocolumnar junction in the cervix [6] for malignant or pre-malignant signs. While highly successful as a prevention method, the quality of the samples obtained through this process can vary greatly, making them especially hard to analyze automatically. Due to this, several alternative techniques have been developed over the years, such as liquid-based cytology (LBC) preparations. In LBC, the samples collected are suspended in a liquid and then go through several treatment processes before being fixed in a glass slide and stained [6], resulting in a thin layer of cellular material easier to interpret visually.

1.2.2.2 Computer-Aided Diagnosis Systems

Initial cervical cytology CAD systems, such as the Cytoanalyzer, had trouble differentiating a sample's cells from other irrelevant artefacts, resulting in many false positives [6]. This problem, combined with the high cost of automated microscopes and sufficiently powerful computers,

prevented them from being commercialized [6]. Since then, advances in computing technologies and sample preparation techniques have made it possible to overcome these problems to a certain extent, enabling some systems to gain the approval of the United States Food and Drug Administration (FDA) and enter the market. Some noteworthy devices are PAPNET, FocalPoint (previously known as AutoPap 300) and the ThinPrep Imaging System, with the last two still being commercially available today.

Typically, the automatic screening process followed by these systems can be divided into different stages, namely focus and adequacy assessment, ROI identification and classification [7]. The goal of the first two stages is to obtain the optimal focus level(s) for a sample and evaluate its adequacy for further processing, respectively. In the ROI identification stage, as the name suggests, certain regions of interest of the image are identified, such as cells and their constituents. This process is usually done through segmentation or bounding box regression. Finally, in the classification stage, the regions previously identified are labelled according to their abnormality level. This problem can be of a binary or multi-class nature, depending on the approach followed.

The algorithms employed in the last two stages can be roughly divided into two groups: conventional machine learning combined with computer vision techniques, and deep learning approaches, with both types having their own set of advantages and disadvantages. For example, deep learning architectures usually achieve greater performance due to their ability to extract deep hierarchical features. However, these networks require large amounts of training data, something which is not always possible to obtain. On the other hand, simpler machine learning algorithms surpass deep learning in this aspect; however, their results are usually not as satisfactory. It is then expected that by combining both types of algorithms in hybrid pipelines, the shortcomings of each type are mitigated or even cancelled out by the other, thus originating a more robust system in the end.

1.3 Objectives

This work aims at exploring the combination of deep learning and conventional machine learning algorithms for the detection and classification of abnormal regions in cervical cytology images. More specifically, it intends to study the application of different algorithms from both approaches to the detection and classification tasks and perform a comparative analysis of their performance, ultimately evaluating their integration in a single hybrid pipeline. In addition to this, the relevance of features extracted from the nuclei present in a region is also assessed to evaluate whether to pursue further research into strategies involving them or focus mainly on the complete regions. Through all of this, the present work also intends to support future research into hybrid approaches applied to similar scenarios.

Chapter 2

Related Work

After the general adoption of the Pap test, the automatic detection and classification of abnormal regions in cervical cytology images began to be researched. With the technological advances in computer hardware and slide preparation techniques over the years, numerous approaches to these problems have been proposed. More recently, deep learning architectures have gained considerable popularity due to their outstanding performance in many applications. Despite this, methods based on conventional machine learning algorithms continue to be studied, mainly due to their simpler nature and higher model interpretability.

In this section, various state-of-the-art solutions are examined for the detection and classification tasks with an emphasis on deep learning and conventional machine learning algorithms. Afterwards, a review of hybrid pipelines combining both types of algorithms is also conducted, followed by a summary of the discussed works.

2.1 Region of Interest Detection

2.1.1 Conventional Machine Learning

When it comes to detecting and isolating regions of interest in cytology images, the approaches more frequently encountered are based on conventional computer vision techniques aiming to segment an image's cells, more specifically their nucleus and cytoplasm, since the ratio between these is one of the most discriminative features for the identification of cervical lesions [7]. The Overlapping Cervical Cytology Image Segmentation challenges, held in conjunction with the Institute of Electrical and Electronics Engineers (IEEE) International Symposium on Biomedical Imaging (ISBI), and more commonly known as the ISBI challenges, focus precisely on this issue, with the added difficulty of performing this type of segmentation on overlapping cells.

The submission ranked first in the 2014 challenge by Ushizima, Bianchi and Carneiro [8] used a graph-based algorithm to identify cellular clumps (aggregates of partially overlapping cells),

merging regions based on pixel adjacency and intensity similarity. The nuclei of these clumps were then identified with a local thresholding algorithm, which were then used as the seeds for a region growing process to identify the cytoplasm regions. It achieved a precision and recall of 0.959 and 0.895, respectively, in the nuclei detection problem and a dice coefficient (DC), true-positive (TP) rate and false-positive (FP) rate of 0.872, 0.841 and 0.002, respectively, on the cytoplasm segmentation problem. However, the cytoplasm segmentation's visual appearance is unrealistic due to the final Voronoi image partitioning employed [9].

As for the winners of the 2015 version of the challenge, Phoulady et al. [10] take a different approach in which the cell's nuclei are detected through iterative thresholding and the cell clumps segmented through binary thresholding. The threshold value used for the latter is obtained through a two-component (background and foreground) Gaussian mixture model learned on pixel intensities, more specifically, it is equivalent to the background component's quantile function at 0.1. The images are then divided in a grid-like manner and the cytoplasm segmentation is obtained through several calculations and morphological operations based on the standard deviation and average edge strength computed for each grid square, properties which were themselves based on the Sobel operator. The results were obtained through 2-fold cross-validation and showed a 0.847 DC, 0.236 false-negative (FN) rate, 0.859 TP rate and 0.001 FP rate.

Kuko and Pourhomayoun [11] propose another more classical approach as well through the use of a sliding window strategy to analyse Pap smear slides, extracting the cells through computer vision and image manipulation techniques. More specifically, grey-scale conversion and subsequent binary thresholding is performed to identify each cell's contours, drawing the respective bounding boxes afterwards. However, one major drawback with this approach is that many other artefacts, such as cell debris, are also captured. After this step, segmentation of each cell in four components (nucleus, cytoplasm, cell membrane and slide background) is achieved by applying a mini-batch K-means algorithm to each image's pixel RGB values vector. This step also allows the generation of 33 morphological features that are used afterwards for classification purposes. Although adequate for this type of task, this segmentation strategy suffers from over and under segmentation problems, potentially lowering the extracted features' accuracy.

2.1.2 Deep Learning

Despite the major significance of the features that can be extracted from an adequately segmented image, this task is not particularly easy to accomplish. Owing partially to this, deep learning techniques have emerged as alternatives for the detection and classification of abnormal regions, especially in images containing large numbers of cells or other artefacts.

As an example, Du, Li and Li [12] proposed a Faster R-CNN based system for the detection and classification of cervical exfoliated cells, with the ultimate goal of "segmenting overlapping regions into multiple single cells". The analysis of these regions is still one of the most challenging aspects in CAD systems due to the difficulty in singling out the composing cells. In this work,

several convolutional neural networks were compared as backbones for the Faster R-CNN, namely VGG16, ResNet50 and ResNet101. Experimentation was conducted using transfer learning with the pre-trained weights from ImageNet on a private dataset comprised of 680 LBC cervical exfoliated cell samples, with a total of 18,957 individual cells. Data augmentation was performed on the training set to compensate for class imbalances. Results showed that the mean of the predicted bounding boxes that overlapped the ground truth boxes was 0.8259 in the network with a ResNet101 as its backbone, which was better than in the VGG16 and ResNet50 cases, but also time-consuming.

Li et al. [13] and Mahmood et al. [14] also proposed deep learning systems to detect mitotic cells in histopathology images. In the latter case, a Faster R-CNN was used with a Resnet50 as its backbone while also being trained on the ImageNet database and on the ICPR 2012 and ICPR 2014 public datasets. The performance was improved at this stage by fixing the anchor scale and limiting the number of anchor boxes. However, the results produced by the Faster R-CNN at this point still contained a large number of false positives. To mitigate this, the regions proposed undergo post-processing operations based on statistical local binary pattern (LBP) histograms of oriented gradients (HOG) and other colour features. The candidate cells' final classification is then accomplished through the score-level fusion of two separate convolutional neural networks (CNN), namely another ResNet50 and a DenseNet201, trained over the patches of the detected mitotic cells. The system reportedly achieved a state-of-the-art precision, recall and F1 score of 0.876, 0.841 and 0.858, respectively, on the ICPR 2012 dataset. It was recognised that the ResNet50's skip connections proved very helpful since there are significant variations in the size of mitotic cells. Through these connections, this type of information is not lost, being re-used in deeper layers.

One-stage object detection architectures have also been applied to similar scenarios. The main difference between these algorithms and two-stage ones, such as the Faster R-CNN, is that the latter have a dedicated region proposal network (RPN) which generates regions of interest in a first phase, performing the bounding box regression and object classification of these regions in a following stage. One-stage detectors, such as RetinaNet and single shot detector (SSD), do not have a dedicated RPN, learning the class probabilities and bounding box coordinates in the image as a whole. Usually, one-stage models are faster than two-stage methods but have lower accuracy scores [15], making them more suited for real-time applications. Despite this, some studies demonstrated that one-stage detectors can still achieve comparable or even better results than two-stage architectures.

Zhou et al. [16], for example, compared different methods for detecting abnormal cells in cervical pathology images, namely a Faster R-CNN, SSD, Grid R-CNN, fully convolutional one-stage object detector (FCOS) and RetinaNet. A private annotated set of 10,619 images was used to evaluate their performance (an example can be seen in Figure 2.1), with the RetinaNet displaying the highest average precision (AP) of 0.715, closely followed by the Faster R-CNN with an AP of

0.706, while also having the second-lowest runtime of 0.128 seconds, only bested by the FCOS system. Similar results whilst using RetinaNet were also reported by Marzahl et al. [17] for the detection of hemosiderophages in cytology slides, achieving a mean average precision (mAP) of 0.66.

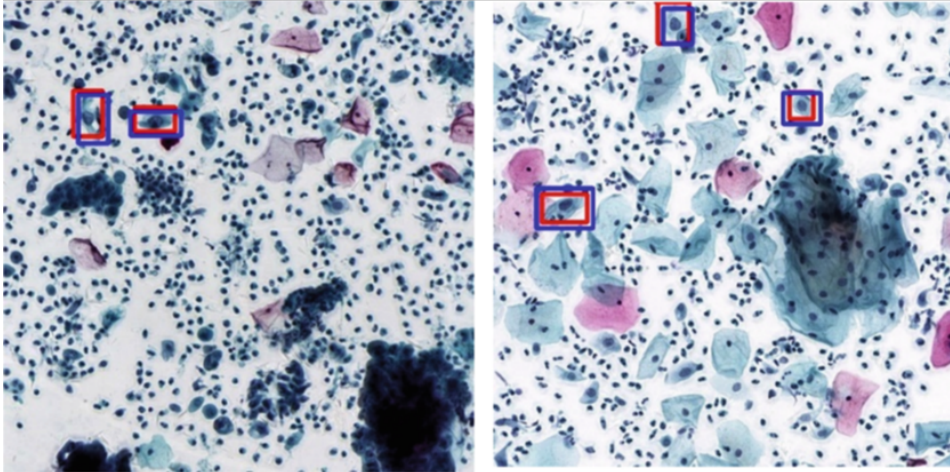


Figure 2.1: Example of detected objects by Zhout et al. [16]. The red rectangles represent the ground-truth bounding boxes, while the blue ones represent the detection network's results.

Grabel et al. [18] also utilized RetinaNet in an innovative way by using circular anchors, instead of the usual box-shape, to detect hematopoietic cells, which are mostly circular as well. It improved results slightly, from an AP of 0.952 while using only the RetinaNet with area-based non-maximum suppression (ANMS) to an AP of 0.956 when adding the circular anchors. The number of false positives decreased more considerably, from 2560 to 1598. It is worth noting that the AP value was significantly higher than in the previously discussed studies due to the apparent substantial variation between hematopoietic cells and regular ones.

In the work of Yi et al. [19], a one-shot detector is used as well, in this case an SSD, to identify and segment neural cell instances. The detector's performance is improved by using a fusion module for the features of the third, fourth and fifth convolutional layers, which are then fed to an attention mechanism used to help the network focus on the relevant regions before a classification being produced and respective bounding box generated. The AP was higher than the best of the detectors explored by Zhou et al. [16] (RetinaNet), with an AP@0.5 value of 0.7955, which might be explained by the seemingly easier nature of neural cell detection, mainly due to the relatively clean background as opposed to some instances of cervical cytology samples. Furthermore, Zhou et al. [16] evaluated their results on 10619 annotated pathology images, while Yi et al. [19] used a collection of only 386, 129 and 129 neural cell images for training, validation and testing purposes, respectively, while also making use of data augmentation techniques and transfer learning from the ImageNet dataset (as opposed to Zhou et al. [16]).

Paper / Authors	Detection Technique	Datasets	Performance
Ushizima, Bianchi and Carneiro [8]	Nuclei detection through superpixel representation. Cytoplasm detection through nuclear narrow-band seeding + graph-based region growing + Voronoi diagrams.	ISBI 2014	(Nucleus Detection) Prec: 0.959; Rec: 0.895; Prec (Pixel): 0.968; Rec (Pixel): 0.871; Dice (Pixel): 0.914;
Phoulady et al. [10]	Nuclei detection through iterative thresholding.	ISBI 2015	DC: 0.847; FN: 0.236; TP (Pixel) 0.859; FP (Pixel): 0.001
Kuko and Pourhomayoun [11]	Sliding window + thresholding + contour detection	Private	-
Du, Li and Li [12]	Faster R-CNN w/ VGG16 / ResNet50 / ResNet101 backbone	Private	(ResNet101) Mean of predicted bounding boxes overlapping ground truth: 0.8259
Li et al. [13]	Faster R-CNN	ICPR 2012, ICPR 2014	(ICPR 2012) Prec: 0.854; Rec: 0.812; F1: 0.832
Mahmood et al. [14]	Faster R-CNN w/ ResNet50 backbone + Post processing (LBP, HOG) + Score-level fusion of ResNet50 and DenseNet201	ICPR 2012, ICPR 2014	(ICPR 2012) Prec: 0.876; Rec: 0.841; F1: 0.858 (ICPR 2014) Prec: 0.848; Rec: 0.583; F1: 0.691
Zhou et al. [16]	Faster R-CNN, SSD, Grid R-CNN, FCOS, RetinaNet	Private	(Faster R-CNN) AP: 0.706; (SSD) AP: 0.661; (Grid R-CNN) AP: 0.689; (FCOS) AP: 0.697; (RetinaNet) AP: 0.715
Marzahl et al. [17]	Modified RetinaNet w/ ResNet18 backbone	Private	mAP@50: 0.66
Grabel et al. [18]	ResNet50 w/ ANMS and Circular Anchors	Private	AP: 0.956; F1: 0.925
Yi et al. [19]	SSD w/ ResNet50 backbone	Private	AP@50: 0.7955; AP@0.7: 0.3592;

Table 2.1: Summary table with highlighted works for the ROI detection problem. Prec (Precision), Rec (Recall), F1 (F1-measure).

2.2 Classification

2.2.1 Deep Learning

The classification of cervical cells has had, arguably, a bigger focus than their detection, partially because besides the classification-only studies conducted, research on the detection of abnormal cells is frequently paired with their respective classification afterwards. Other factors that also contribute to this are the simpler nature of this task compared to the previous one and the greater availability of annotated data.

For example, the previously seen Faster R-CNN system proposed by Du, Li and Li [12] also performs the classification of the identified cells in five classes, namely normal epithelium, abnormal epithelium, lymphocyte, neutrophil and garbage cell, achieving a sensitivity, specificity, mAP and accuracy of 93%, 90%, 66.98% and 91.61%, respectively, whilst using a ResNet101 as the backbone.

A classification stage is also present in the already mentioned work of Zhou et al. [16]. After the detection phase, fixed-size patches are extracted from each detected region, whose feature maps are then obtained through an encoder network. These are next multiplied by the detection network's confidence level, obtained from before, and fused afterwards to obtain a final classification of the whole image. Results showed that the use of the detection network's confidence in the second stage improved the classification accuracy by 7%, with a final value of 92% and an F1-score of 87%.

In the also previously discussed work of Kuko and Pourhomayoun [11], a binary classification of the extracted cells is performed (normal or abnormal). For this purpose, two ensembles were compared, one composed of eight random forests, each trained on a different image rotation (performed as a data augmentation step), and another composed of five convolutional neural networks, each trained on a different cell cluster. The deep learning ensemble achieved the best accuracy and specificity of 91.63% and 87.43%, respectively, albeit with slightly lower sensitivity than the random forest ensemble (95.47% compared to 96.33%).

Zhang et al. [20] obtained better results with their custom deep learning network composed of five convolutional layers, three pooling layers and three fully connected layers. Tested on the Herlev and H&E stained manual LBC (HEMLBC) datasets, it achieved a sensitivity, specificity, accuracy and area under the curve (AUC) of 98.2%, 98.3%, 98.3% and 0.998, respectively, on the first one and a sensitivity, specificity and accuracy of 98.3%, 99.0% and 98.6%, respectively, on the latter. The system used transfer learning with weights trained on the ImageNet dataset and data augmentation techniques, operating on nucleus-centred patches extracted from each image. Note that this patch extraction was also performed in test images. In these cases, the final classification was obtained by averaging the scores of each patch's prediction. Despite the promising results, this data augmentation step increased the classification time of a single patch to 3.5 seconds (compared

to just 0.035 seconds without it) with only a 1% increase in accuracy, which is not nearly quick enough in a clinical setting.

Kwon et al. [21] also used a custom network for the classification of cervical cells, composed of four convolutional layers followed by a dropout and two fully connected layers. The system was evaluated on a private dataset in both a binary and multiclass-type problem. The first was a simple normal/abnormal classification, while the second considered the NILM (negative for intraepithelial lesion or malignancy), low grade and high grade classes. The low grade class was comprised by regions of type ASC-US (atypical squamous cells of undetermined significance) and LSIL (low grade squamous intraepithelial lesion), while the high grade class was composed of ASC-H (atypical squamous cells, cannot rule out high grade squamous intraepithelial lesion) and HSIL (high grade squamous intraepithelial lesion) regions, due to the shared morphological features between each group's types. However, it only achieved an accuracy of 84.5% and 76.1% on the two and three-class problem, respectively.

Ghoneim, Muhammad and Hossain [22] investigated the classification of cervical cancer cells using deep learning methods as well. In their work, the VGG16 and CaffeNet networks are compared as feature extractors, as well as a shallow CNN. Each is then combined with either an extreme learning machine (ELM), multi-layer perceptron (MLP) or auto-encoder (AE)-based classifier. The performance of each combination was evaluated using the Herlev dataset, and in both the two and seven-class problems ¹, the CaffeNet-ELM system achieved greater or equal accuracy than the others (99.7% and 98.2%, respectively). The shallow model also performed well, achieving an accuracy of 99.5% and 97.5% when combined with the ELM-based classifier in the two and seven-class problems, respectively, showing that if there are time constraints, it too can be used without a significant decrease in accuracy.

The process of feature extraction and classification of cervical cells is also analysed by Mat-Isa, Mashor and Othman [23], which propose a system based on a modified version of the seed-based region-growing (SBRG) algorithm for the former problem and a hybrid MLP (HMLP) for the latter. In the first task, a customised version of the SBRG algorithm presented in [24], in which a moving K-means algorithm is used to automatically determine the threshold values of the nucleus and cytoplasm regions of a cell, is further modified to extract relevant features. As for the HMLP, the main difference from a regular MLP is the presence of direct connections between input and output nodes. To handle these, a modified recursive prediction error (MRPE) algorithm was introduced as a learning algorithm. However, this came with the limitation that it could only be applied to HMLP networks with only one output node. As such, to efficiently classify the cells in normal, LSIL or HSIL, two HMLP were assembled in a cascade system (H^2MLP). The first performed the distinction between normal and abnormal cells, feeding the second with the abnormal ones, which would then be classified as LSIL or HSIL. The system was compared with three other artificial neural networks (ANN), namely a radial basis function (RBF) and an MLP network, and

¹In the Herlev dataset, the two-class problem refers to the distinction between normal and abnormal cells, and the seven-class problem differentiates between three types of normal cells and four types of abnormal ones.

a system composed of three separate HMLP networks to obtain the three-class classification. For this purpose, a dataset was used composed of 550 Pap smears (211 normal, 143 LSIL and 196 HSIL) from Kota Bharu Hospital and University Sains Malaysia Hospital. Results showed that the H^2MLP system outperformed the other networks independently of the number of hidden nodes used, obtaining the best accuracy, specificity and sensitivity of 97.50%, 100.00% and 96.67%, respectively, with only four hidden nodes.

2.2.2 Conventional Machine Learning

In the work of Su et al. [25], a more conventional approach is used to tackle the classification task. In the proposed system, the cells in a slide are first segmented through adaptive thresholding to extract a selection of 28 features from each one, more specifically 20 morphological and eight texture features. These are then fed to a C4.5 decision tree which classifies the cells into lymphocytes, epithelial, neutrophils or garbage cells. The ones classified as epithelial are then sub-classified into normal or abnormal by a logistic regression (LR) algorithm. The system's evaluation was performed on a private dataset with 120 LBC slides composed of 20,000 cells, with 2,500 of those being abnormal epithelial cells. It achieved an accuracy and precision of 95.805% and 95.642%, respectively, both higher than when using the algorithms separately or the K-Nearest Neighbour (K-NN) algorithm in a one-stage scenario. Furthermore, a Naive Bayes (NB) classifier was also tested as a replacement for each stage's algorithm, but the performance was not as good in either stage. Sharma et al. [26] also experimented with K-NN as a sole classifier of the (binary) abnormality level of cervical cells, with results showing that the algorithm had a worse performance than other approaches (accuracy of 82.9%), despite the previous cell segmentation.

Lastly, Prum, Handayani and Boursier [27] proposed a support vector machine (SVM)-based system to classify cervical cells while using a HOG descriptor for feature extraction. As a data normalization step, the nuclei of the cells were previously identified through thresholding and edge detection techniques, followed by the centring and cropping of the images around it, as can be seen in Figure 2.2. Testing was conducted on the Herlev dataset, and while a moderate accuracy of 88.83% was achieved in the more realistic configuration of the binary classification problem (normal versus abnormal), the accuracy on the seven class problem was much lower (around 45.50%), possibly due to the similarity between the abnormal cells' nucleus shape and texture, making them difficult to differentiate using this method.

2.3 Hybrid Approaches

The explicit use of both deep learning and conventional machine learning to detect and classify cervical cells or similar is somewhat scarce, so, in this section, a brief analysis of past research is conducted where both types of algorithms are used for related tasks, such as in one of the experiments of the already seen work of Kuko and Pourhomayoun [11].

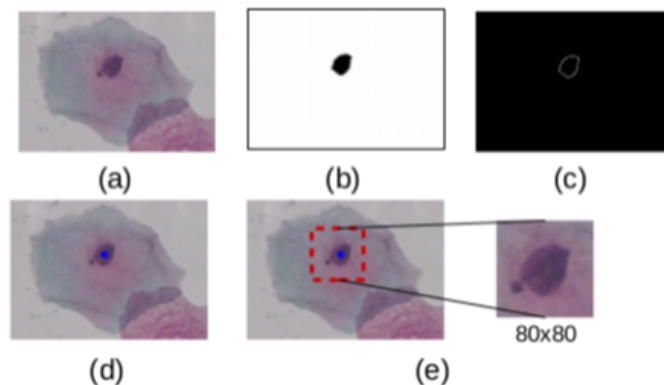


Figure 2.2: Example of nucleus centre detection by Prum et al. [27]. (a) Input image, (b) binary image, (c) detected nucleus edges, (d) detected nucleus center point, (e) cropped image.

For instance, Jia, Li and Zhang [28] proposed a novel system to classify cervical cells based on the combination of a CNN and an SVM. Two methods for feature extraction were employed, one aimed at extracting textural, morphological and chromatic features of the cell, referred to as strong features in the article, and another intended to retrieve the more abstract features. The strong features were extracted using various methods and filters, namely grey-level co-occurrence matrix (GLCM), Fourier and Gabor transformations, and Markov random field. These were subsequently standardised and dimensionally reduced. As for the abstract features, these were retrieved using a modified LeNet-5 CNN, more specifically from its last pooling layer, being normalised afterwards. Both strong and abstract features were then standardised and fused, with weights associated with each set of features. The resulting feature vector had a dimension of 5260, so the principal component analysis (PCA) method was used to reduce it. Finally, an SVM was used for the final classification. The system was tested on the Herlev dataset as well as on a private dataset from Guangdong Province People's Hospital. Results for the former showed an accuracy, sensitivity and specificity of 99.3%, 98.9% and 99.4%, respectively, for the two-class problem, and of 93.8%, 93.7% and 93.7%, respectively, for the seven-class problem. Furthermore, two other models were also tested, one based purely on the CNN for both feature extraction and classification, and another using only the features extracted from the CNN for SVM classification, but the proposed model outperformed them. In addition to this, other shallower layers from the CNN were tested for feature extraction as well but showed worse results.

In a more investigative manner, Khamparia et al. [29] developed a framework for the classification of cervical cells in which several CNN and machine learning algorithms were tested as feature extraction mechanisms and classifiers, respectively. The examined CNN were the InceptionV3, SqueezeNet, VGG19 and ResNet50 models, all pre-trained on the ImageNet dataset. The features extracted by each of the networks were then fed to one of the following algorithms for the final binary (normal/abnormal) classification: K-NN, Naive Bayes, logistic regression, random forest and SVM. The system was tested on the Herlev dataset and results showed that the ResNet50

combined with the random forest achieved the best accuracy of 97.89%. In fact, the random forest classifier had the greatest accuracy of all the other classifiers when combined with the SqueezeNet and VGG19 as well, only being surpassed by the SVM paired with the InceptionV3 network.

Wang et al. [30] also use a CNN for feature extraction and a random forest classifier for mitosis detection. However, the system is not as straightforward as in the work of Khamparia et al. [29], being assembled in a cascade manner instead. In the first stage, the three-layer network and a random forest model produce independent classifications, using several handcrafted features for the latter extracted through classical computer vision techniques, such as thresholding and local non-maximum suppression (NMS). The combined probabilities of these predictions are then calculated, and if they are within a particular couple of thresholds, a second random forest ensemble performs the final classification using a combination of the handcrafted features with the CNN extracted ones. The results showed that the cascade structure benefited the system, having a better performance (F-measure of 0.7345) than when using the CNN or random forest individually (F-measure of 0.5730 and 0.6864, respectively).

Sarwar, Sharma and Gupta [31] also proposed a hybrid "ensemble of ensemble methods" for the classification of cervical cells in which the output of several algorithms, some of them ensembles themselves, was used to produce a final prediction equal to the mode of the individual class outputs. The algorithms considered were Naive Bayes, ANN, decision table, filtered classifier, random committee, partial decision tree algorithm (PART), ensemble of nested dichotomies (END), decorate, J48 graft, rotation forest, bagging, random forest, multiclass classifier, radial basis function network and random subset space. The results, evaluated on the Herlev dataset, showed that the hybrid ensemble achieved the highest accuracy in the two and seven-class problems (98.5700% and 78.8571%, respectively) compared to the individual use of each algorithm. The second-best performing ones were the random forest ensemble in the two-class problem and the END in the seven-class problem, with the respective accuracy levels of 97.7132% and 72.2857%. While this suggests that the combined use of different algorithms is beneficial due to the capacity of mutually overriding each other's mistakes, such an extensive collection also makes the model harder to explain.

Finally, Gautam et al. [32] performed several innovative experiments for the detection, segmentation and classification of cervical cells, or rather, their nuclei. The detection phase is the simplest, making use of a global threshold to identify the cells' nuclei after applying a median filter and contrast-limited adaptive histogram equalisation (CLAHE). Several features are extracted from the bounding boxes generated around the nuclei, which are then used for the final classification, both with and without previous nucleus segmentation. The feature extraction is performed by an AlexNet, more specifically from the first, third and fifth convolutional layers. The features from each layer were used for classification separately from each other in different tests to evaluate which type of features are the most beneficial². These features were then passed to a decision

²The first convolutional layers in a CNN learn the more basic features of an image, while the deeper layers learn higher-level data.

tree-like structure for classification, which can be observed in Figure 2.3, where in the first split, a distinction was made between normal and abnormal cells, with the descendant nodes differentiating between different abnormality classes. The classification at each split was achieved through two fully connected layers to which the features extracted from the AlexNet were fed. These tests were performed on the Herlev dataset and another one consisting of 80 multi-cell images collected from an oncology centre by Aindra Systems Pvt. Ltd., Bangalore, India. Results showed that classification performance was the greatest when using the features extracted from the first layer of the AlexNet (which goes against the results of Jia, Li and Zhang [28] in this aspect) and without nucleus segmentation, achieving an accuracy of 99.3% and 93.75% on the two and seven-class problems of the Herlev dataset.

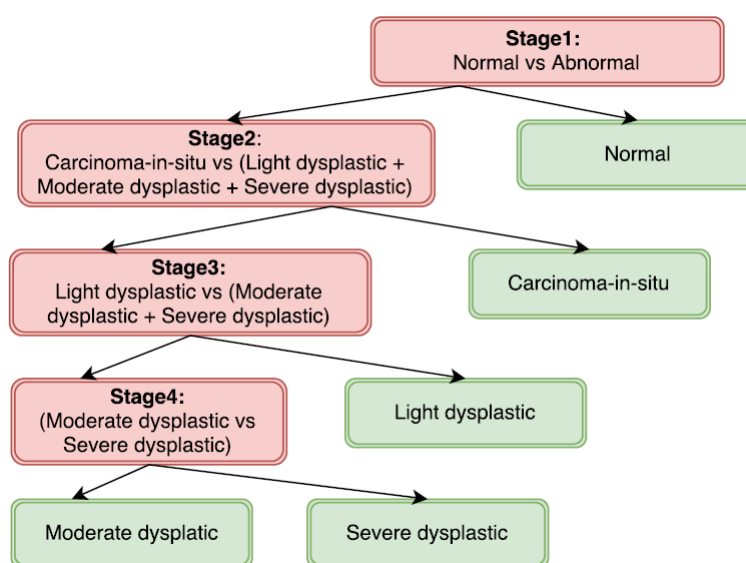


Figure 2.3: Classification decision tree structure by Gautam et al. [32].

2.4 Summary

As seen in Section 2.1, deep learning approaches have shown a greater potential over other non-segmentation-based techniques for detecting cervical cancer in cytology images, mainly due to the generally better performance when analysing new samples, both in terms of produced results and computational power required [33], among other factors, as observed in the work of Kuko and Pourhomayoun [11].

From the previously discussed studies on deep learning architectures, both two and one-stage detectors achieved good results; however, it is hard to establish a direct comparison between each work due to the different private datasets used for evaluation purposes in most of them, as well as distinct performance measures or image types between some of the studies, as can be seen in Table 2.1. Even so, the Faster R-CNN and RetinaNet architectures can be identified as some of

the most promising approaches, with Li et al. [13] showing good precision and recall levels with their Faster R-CNN-based architecture, and Zhou et al. [16] achieving their best AP levels with a RetinaNet. The last case is curious, since when compared to two-stage detectors, one-stage architectures usually have a worse accuracy, favouring detection speed instead [15]. Nevertheless, the RetinaNet has shown comparable (or even better in some cases) results to state-of-the-art two-stage detectors, including in other cytology environments, such as in the works of Marzahl et al. [17] and Grabel et al. [18]. In terms of backbones for the first architecture, the ResNet50 seems to strike a good balance between accuracy and training time/complexity, as seen in the work of Du, Li and Li [12]. In addition to this, its use of shortcut connections also proves helpful for feature extraction purposes, as observed by Mahmood et al. [14]. Furthermore, the best performance and fastest detection speed obtained by Yi et al. [19] was with a ResNet50 backbone as well.

In contrast to the ROI detection task, the works discussed in Section 2.2 for the classification problem, which are summarized in Table 2.2, have a greater variety in the employed methods. The use of nearly the same evaluation metrics also makes their comparison easier, even if the majority still use private datasets. Several studies stand out, such as the systems proposed by Mat-Isa, Mashor and Othman [23], and Su et al. [25], which demonstrated that combining several classifiers in an ensemble manner can be beneficial. Prum, Handayani and Boursier [27] arrived at the also interesting conclusion that textural information from a cell's nucleus is acceptably discriminative for normal and abnormal cells, but not nearly enough for specific abnormality classes. It is also worth mentioning the experiments conducted by Ghoneim, Muhammad and Hossain [22], which obtained the best accuracy (evaluated on the Herlev dataset) of all the highlighted works with a CaffeNet and ELM as a feature extractor and classifier, respectively, as well as a low percentage of false negatives and false positives.

Lastly, regarding hybrid approaches (listed in Table 2.3), even though most of the works that can be found do not use a combination of deep learning and conventional machine learning techniques for the explicit detection and classification of cervical regions in cytology images, some helpful insights can still be gathered. Most notably, the use of abstract cellular features extracted from deep learning networks can successfully be used with classical machine learning classifiers with very competitive results, as seen in the work of Khamparia et al. [29]. This work is also particularly relevant due to the various "mix-and-match" experiments performed, which showed that the random forest classifier achieved the best accuracy in all the tests conducted when paired with the ResNet50, being the top classifier when coupled with any of the tested networks except for the InceptionV3, where the SVM had the best accuracy score. In the other experiments, both the SVM and K-NN algorithms showed accuracy levels fairly close to the random forest, with an average difference between each and the ensemble of 0.95% and 0.33%, respectively. The system proposed by Jia, Li and Zhang [28] was similar, but incorporated the addition of "strong" features extracted using conventional computer vision techniques, as well as a different deep learning

feature extractor ³. The results were also evaluated on the Herlev dataset and showed a slight increase in accuracy when compared to the system proposed by Khamparia et al. [29], possibly demonstrating that the combination of both types of features can be beneficial.

³While an SVM was employed in the system of Jia, Li and Zhang [28], and the best model of Khamparia et al. [29] had a random forest as a classifier, an SVM was also tested by the latter.

Paper / Authors	Classification Technique	Datasets	Performance
Du, Li and Li [12]	Faster R-CNN w/ VGG16 / ResNet50 / ResNet101 backbone	Private	(VGG16) Acc: 0.8916; Sens: 0.90; Sp: 0.8912; mAP: 0.6116; (ResNet50) Acc: 0.9077; Sens: 0.91; Sp: 0.89; mAP: 0.6519; (ResNet101) Acc: 0.9161; Sens: 0.93; Sp: 0.90; mAP: 0.6698
Zhou et al. [16]	Patch Encoder Module (six convolutional + two max pooling layers) + Fusion Module (global average pooling (GAP) + fully connected (FC) layer)	Private	Acc: 0.92; F1: 0.87
Kuko and Pourhomayoun [11]	Random forest ensembles, CNN ensemble	Private	(Random forest ensembles) Acc: 0.9037; Sens: 0.9633; Sp: 0.8359; (CNN ensemble) Acc: 0.9163; Sens: 0.9547; Sp: 0.8743
Zhang et al. [20]	DeepPap (five convolutional + three pooling + three FC layers)	Herlev, HEMLBC	(Herlev) Acc: 0.983; Sens: 0.982; Sp: 0.983; AUC: 0.998; (HEMLBC) Acc: 0.986; Sens: 0.983; Sp: 0.99
Kwon et al. [21]	CNN (four convolutional + one dropout + two FC)	Private	(two-class) Acc: 0.845; Sens: 0.791; Spec: 0.895; (three-class) Acc: 0.761;
Ghoneim, Muhammad and Hossain [22]	VGG16 / CaffeNet / shallow CNN + ELM / MLP / AE	Herlev	(Two-class, CaffeNet-ELM) Acc: 0.997; FN: 0.003; FP: 0.002; (Seven-class, CaffeNet-ELM) Acc: 0.982
Mat-Isa, Mashor and Othman [23]	SBRG + hybrid MLP ensemble	Private	Acc: 0.9750; Sens: 0.9667; Sp: 1.00; FP: 0.03; FN: 0.0133
Su et al. [25]	Adaptive Thresholding + C4.5 decision tree + LR	Private	Acc: 0.95805; Prec: 0.95642
Sharma et al. [26]	K-NN	Private	Acc: 0.829
Prum, Handayani and Boursier [27]	SVM	Herlev	(Realistic, two-class) Acc: 0.8883; (Realistic, seven-class) Acc: 0.4550

Table 2.2: Summary table with highlighted works on cervical cell classification. When a large number of experiments was conducted, performance is only given for the best case. Acc (Accuracy), Prec (Precision), Sens (Sensitivity), Sp (Specificity).

Paper / Authors	Methods	Datasets	Performance
Jia, Li and Zhang [28]	(Feature Extraction) GLCM + Fourier and Gabor transformations + Markov random field + LeNet-5 CNN (Classification) SVM	Herlev, Private	(Herlev, two-class) Acc: 0.993; Sens: 0.989; Spec: 0.994; (Herlev, seven-class) Acc: 0.938; Sens: 0.937; Sp: 0.937
Khamparia et al. [29]	(Feature Extraction) InceptionV3, SqueezeNet, VGG19, ResNet50 (Classification) K-NN, NB, LR, random forest, SVM	Herlev	(ResNet50, random forest) Acc: 0.9789
Wang et al. [30]	(Feature Extraction) CNN / Computer Vision techniques; (Classification) Random forest	ICPR2012, AMIDA13	(ICPR2012) Prec: 0.84; Rec: 0.65; F1: 0.7345; (AMIDA13) F1: 0.319
Sarwar, Sharma and Gupta [31]	(Classification) NB + ANN + decision table + filtered classifier + random committee + PART + END + decorate + J48 graft + rotation forest + bagging + random forest + multiclass classifier + radial basis function network + random subset space	Herlev	Sens: 0.95964; Spec: 0.98425; FP: 0.0157; FN: 0.04035; (two-class) Acc: 0.9857; Prec: 0.99074; (seven-class) Acc: 0.788571
Gautam et al. [32]	(Feature Extraction) AlexNet (Classification) Ensemble of FC layers	Herlev, Private	(Herlev, two-class) Acc: 0.993; (Herlev, seven-class) Acc: 0.9375

Table 2.3: Summary table with highlighted works on hybrid approaches. When a large number of experiments was conducted, performance is only given for the best case. Acc (Accuracy), Prec (Precision), Rec (Recall), Sens (Sensitivity), Sp (Specificity).

Chapter 3

Proposed Method

The literature reviewed in Chapter 2 provided several insights on the tasks undertaken by the present work. For instance, as discussed in Section 2.4, deep learning methods have demonstrated an increased potential for the detection of specific regions in cytology (or similar) images. As such, we opted for a deep learning approach for this stage and a simpler machine learning one for the classification task in order to explore their combination in a single hybrid pipeline. Additionally, an extra nuclei segmentation step was added to the proposed system to extract an increased number of features from these structures. These are usually a key characteristic upon which the specialists base their final decision, making the study of their impact on the final classification performance a topic of interest. Each module in this pipeline is described in Section 3.2, preceded by an analysis of the datasets used in Section 3.1.

3.1 Datasets

Two datasets were used to support this work’s development, both consisting of images from fields of liquid-based slides, more specifically of the ThinPrep kind, with a resolution of 1920×2560 pixels, acquired through a μ SmartScope [34] - a portable device based on a smartphone for the automatic acquisition of microscopic images (an example can be seen in Figure 3.1).

The first dataset comprises images obtained from Hospital Fernando Fonseca (HFF), Amadora, annotated by specialists from the same institution, taken from 21 distinct samples (patients). Even though each sample is composed of approximately 60 to 90 images, only a subset of them contain abnormal cells, hence only 435 annotated images are present. The main annotations here are the bounding boxes surrounding the regions of interest (clumps or individual cells) and the respective class according to The Bethesda System (TBS) [35]. A few examples of these classes can be seen in Figure 3.2, as well as their distribution in Table 3.1. Note that the total number of images does not correspond to the actual number of annotated images since they were counted separately

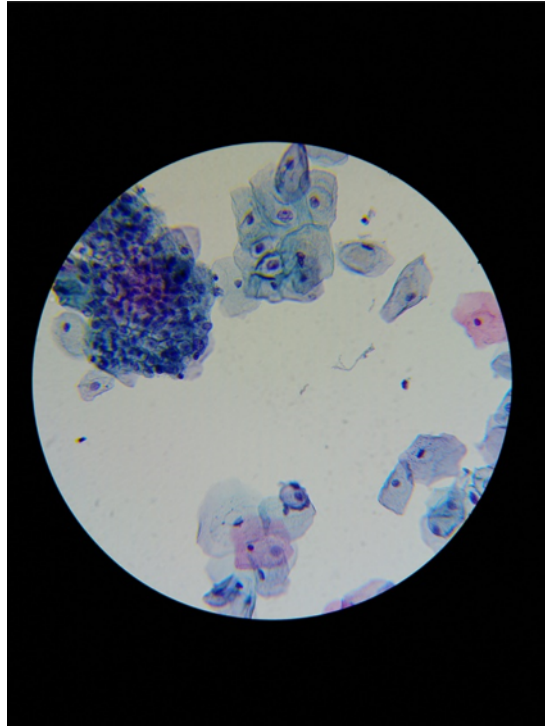


Figure 3.1: Source image example acquired through a μ SmartScope.

for each class and an image may contain, although rarely, objects of more than one class, being counted multiple times in that case.

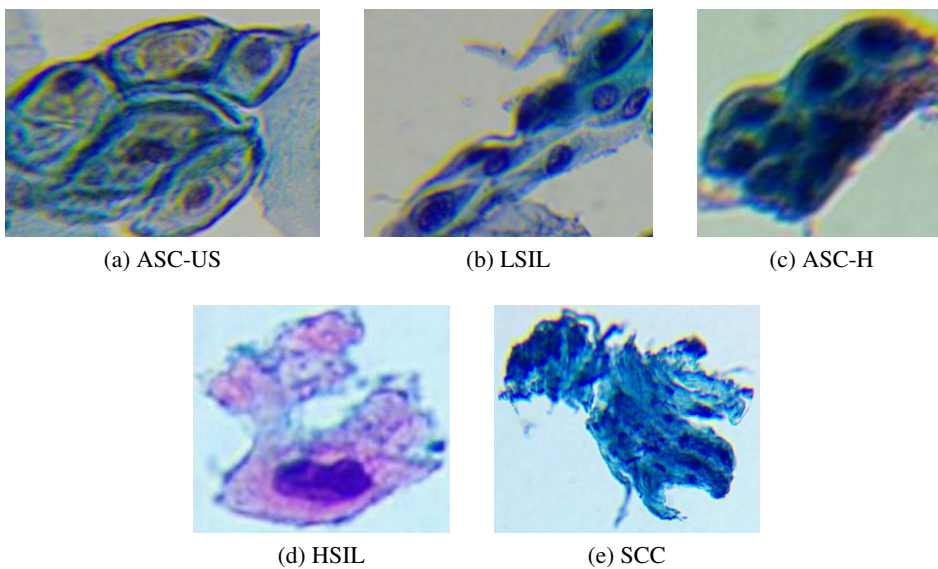


Figure 3.2: Lesion type examples.

The second dataset consists of two subsets of samples acquired from HFF and Instituto Português de Oncologia (IPO), Porto, annotated by specialists from the latter institution, amounting to 559

Class	No. of objects	No. of annotated images
ASC-US	462	292
ASC-H	105	84
LSIL	93	58
HSIL	192	102
SCC	12	8
Total	864	435

Table 3.1: HFF dataset class distribution.

images. The main difference in comparison to the HFF dataset is that the annotations here pertain to the nuclei structures instead of the cell(s) as a whole. It contains the bounding boxes of each nucleus and its classification (the class distribution can be seen in Table 3.2). It is worth noting that not every image from each sample was annotated, mainly to diminish this process effort and increase example variability. In practical terms, this means that only a small portion of the images acquired from HFF have their nuclei annotated. This, in combination with the discrepancy between each dataset's classes, severely limits the possible approaches that can be employed while using nuclei data for the final classification of a region. As such, to enable the usage of the nuclei information in combination with the abnormal regions' data, the development of the pipeline was based on the assumption that a nucleus belongs to the same class as the region it is inserted in.

Class	No. of objects
Glandular nucleus	640
Squamous nucleus	16832
Inflammatory cell	11003
Artefact	140
Indistinguishable object (glandular or squamous nucleus)	632
Indistinguishable object (inflammatory cell, glandular or squamous nucleus)	872
Total	30102

Table 3.2: Nuclei dataset class distribution.

An important characteristic of these datasets to keep in mind, particularly regarding the regions dataset, is the quality of their annotations. Initially, many of them were somewhat inconsistent, with bounding boxes often not capturing the entirety of the structures within (such as the cells)

or including large, unnecessary portions of the background, as can be seen in Figure 3.3. While earlier experiments were performed with this version of the dataset, a correction process was applied to the annotations midway through the system's development to mitigate these situations. Unfortunately, due to time constraints, it was not possible to repeat these experiments for the new version of the dataset.

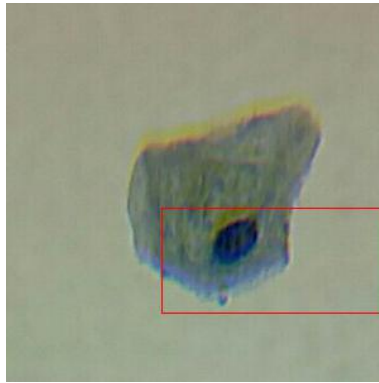


Figure 3.3: Inadequate annotation example.

Another critical aspect to take note of is the highly subjective nature of the annotation process. Different specialists (or even the same one on different occasions) can very well annotate different regions in the same image or assign different labels to the same one, something that can severely limit the performance of both the detection and classification algorithms. Moreover, there are regions with abnormal characteristics whose class is not clearly identifiable. This can be due to several factors, such as the unideal disposition of the cells in the image, insufficient quality of that particular area or the presence of artefacts linked to the acquisition process. In these cases, the specialist cannot accurately propose a label for the region, and as such, they are not annotated, negatively impacting the system's performance, especially that of the detection module.

3.1.1 Data Preparation

The images in each dataset are divided according to the following criteria: first, they are split into an overall training and a test subset at the sample level, meaning that images from the same sample/patient are kept in the same subset. This is done to ensure that the test cases were not too similar to the images observed during training. This split also takes into account the overall diagnosis class in order to balance these out.

From this point, the overall training subset is further divided into several training and validation "sub-subsets" by applying a stratified 5-fold cross-validation procedure to the images of each sample. This is done separately for each sample (for balancing purposes), in which a subset of images is selected for validation and the remaining ones for training. The overall training and validation subsets are then composed of all the images of the same type of "sub-subset" of each sample.

However, due to the long training times of the neural networks, only three folds are selected for the final cross-validation procedure, therefore increasing the number of experiments that were carried out in the limited time frame. The reasoning behind this process and not simply performing a 3-fold cross-validation procedure is linked to the limited amount of available data. Since a 3-fold cross-validation would result in allocating fewer images for training purposes, we opted for this solution instead.

Despite the sample level balancing attempts, there was still a class imbalance when considering the number of annotated images pertaining to each one, as can be observed in Figure 3.4, due to the (naturally imbalanced) distribution of the objects in the dataset.

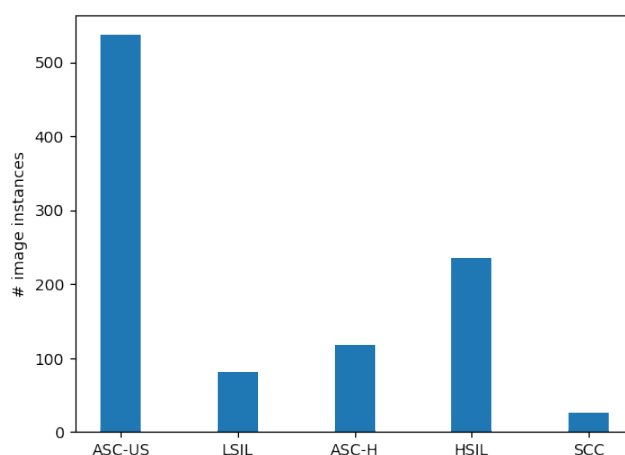


Figure 3.4: Overall training subset's class distribution before balancing.

This situation is undesirable since by not providing enough examples of the under-represented classes, the ability of the algorithm to correctly classify these decreases while also creating a bias towards the over-represented ones. To mitigate this, a balancing operation is carried out in the training subset, using data augmentation techniques to artificially increase the number of examples of the under-represented classes. This was done by applying a calculated number of transformations to each patch, namely rotations of 270, 180 and 90 degrees, vertical and horizontal flipping, blurring and sharpening, in this priority order. These first five operations were selected as they provide modest levels of variation without distorting the size or shape of the regions, which is of paramount importance since these characteristics are highly discriminative, particularly when related to the nuclei of the cells. The last two operations do not modify the patches to the same extent but are nonetheless relevant to mimic the varied focus levels of the structures naturally found in the images.

If, however, the number of transformations needed for a given patch is greater than the number of variations granted by the mentioned operations, random translations are applied to them at the image level, as illustrated in Figure 3.5. This process is achieved by shifting the patch coordinates in relation to the whole image by a random value in the X or Y axis, as many times as required. The

translation’s magnitude is restricted by three factors: the regions of interest present in the patch, the boundaries of the whole image and the regions of interest of other patches. In this manner, it is ensured that its own ROIs are not cropped and that no other ROI from another patch is included in the new image. If a translation is impossible due to these restrictions or all the available ones have already been performed and there is still a need for further transformations, copies of the original patch are performed instead. This situation, however, rarely occurs.

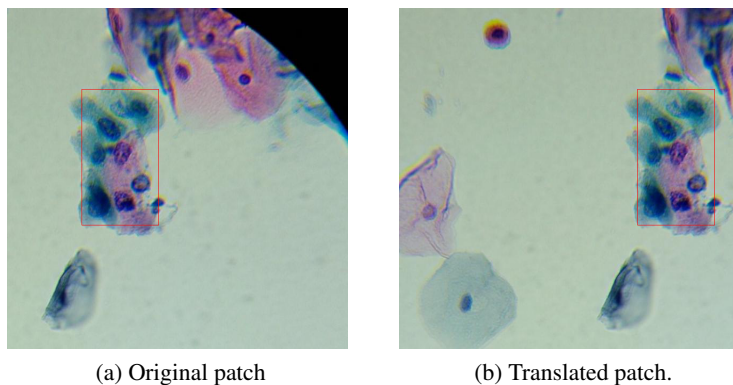


Figure 3.5: Horizontal translation example.

Even so, considering the extremely low number of examples of the SCC class, we opted to merge it with the HSIL class, not only to decrease the overall number of possible copies needed but also due to their similarity, as illustrated in Figure 3.6. Such similarity was also attested by the annotating specialists, which were in doubt themselves when distinguishing some regions from both of these classes.

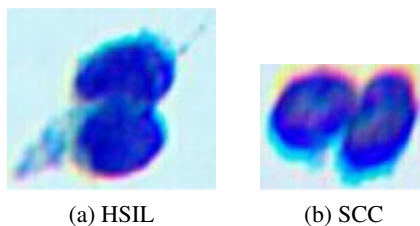


Figure 3.6: Similar examples from the HSIL and SCC classes.

These transformations were validated by visual inspection of some examples to ensure the adequacy of the images remained within acceptable levels. The final class distribution for the overall training subset can be seen in Figure 3.7.

An alternative strategy was also tested at this point in which the balancing was performed exclusively with translations, augmenting all images afterwards through the other transformations. However, this was ultimately abandoned since the translations did not introduce sufficient variety, more so in the case of the 320×320 pixel patches due to the smaller shift range.

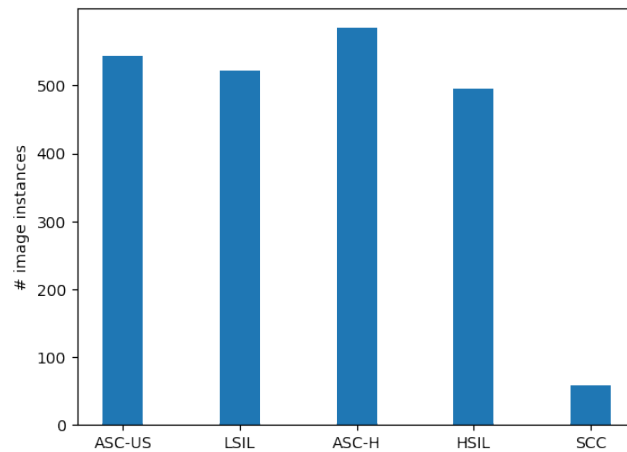


Figure 3.7: Overall training subset's final class distribution after balancing.

Despite achieving a balance between the classes, the number of "empty" patches, i.e., patches without annotations, was still much higher than annotated ones. To avoid the predominance of training instances with no object information, a random down-sampling was conducted on the empty images of this subset (individually for each sample). It was also ensured that at least one empty patch was included for each image since it is important to have negative examples from varied microscopic fields in the dataset to increase the robustness of the detection algorithms. The empty-annotated patch distribution can be seen in Figure 3.8 for each fold and subset.

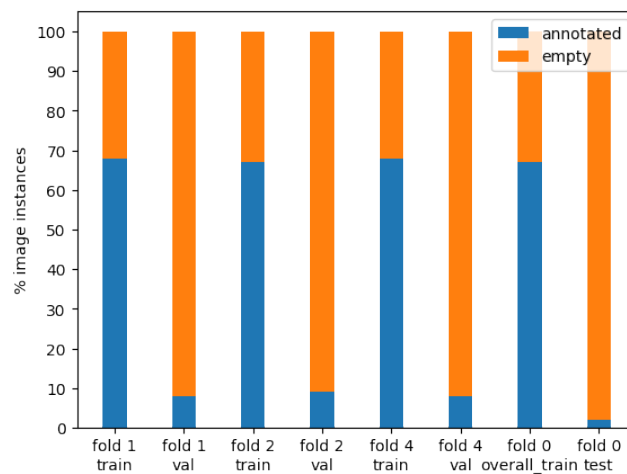


Figure 3.8: Dataset's empty-annotated image distribution after empty image down-sampling.

3.2 Experimental Pipeline

The architecture of the main pipeline can be seen in Figure 3.9. Its input comes in the form of multiple images organized in a hierarchical structure in accordance with their respective source

sample. The first module performs a series of pre-processing operations on these images, most notably their partitioning in smaller-sized patches of 320×320 pixels. These are then fed to a RetinaNet with a ResNet50 backbone for the detection of abnormal regions, which also performs a simpler classification of these regions as low or high grade lesions. Next, for each abnormal region with a detection score higher than 50%, the nuclei within it are segmented through an iterative thresholding algorithm. The segmentation results are used in the feature extraction module to extract 29 geometrical, colour and texture features from the nuclei structures, along with 840 more features obtained from the whole abnormal region. Two SVMs then use these to produce a final classification of ASC-US or LSIL for the low grade regions and of ASC-H or HSIL-SCC for the high grade ones. Finally, the overall classification of a patch, source image and sample is attributed to the most severe classification of a respective detected abnormal region, patch or source image. Each of these modules is described in detail in the following sections, as well as the related experiences performed in each one.

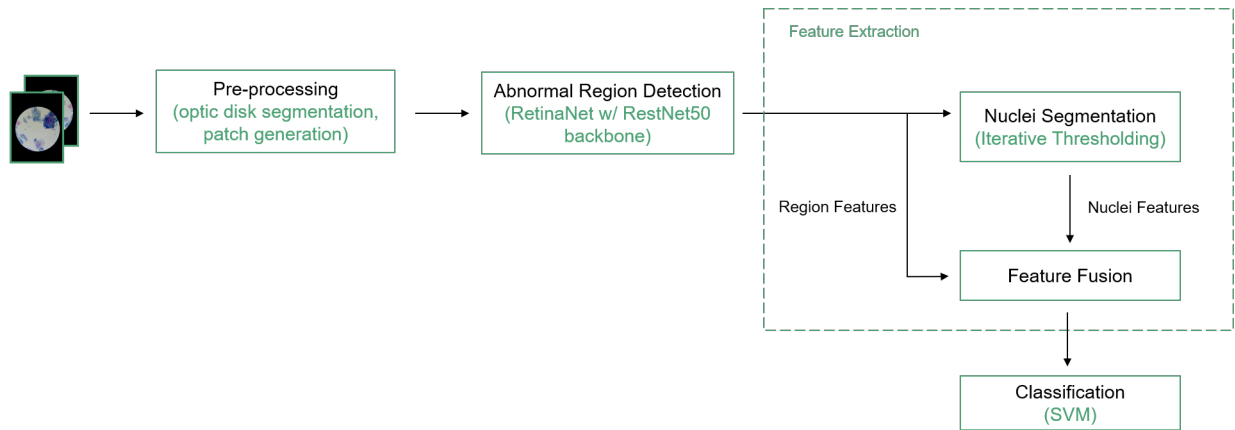


Figure 3.9: Proposed system's architecture.

3.2.1 Pre-processing

Due to the characteristics of the datasets used, several pre-processing steps need to be taken before performing further operations.

For instance, due to their smartphone-based acquisition, a large dark area is present around the optic disk. Since this does not contain any relevant information, it is segmented through the use of the circle Hough transform and then cropped according to the smallest rectangle that encloses the optical disk region. However, the remaining dark areas on the corners of the image hindered some of the nuclei segmentation algorithms further ahead in the pipeline since a basic feature of them is their low overall intensity. As such, the area around the optic disk is also binarized and inverted to create a white background.

The images' high resolution is another aspect to consider since it makes their direct analysis unfeasible due to the large memory requirements. As such, each image is split into several adjacent

patches with a fixed resolution. Initial experiments used patches of 640×640 pixels; however, they still proved limiting when it came to the maximum batch size that could be used by the neural networks further ahead in the pipeline. For this reason, we opted to use patches of 320×320 pixels due to their lower memory requirements.

These patches are initially extracted to yield the maximum amount of adjacent patches within the image’s boundaries, centred in the image. At this stage, the **training** and **validation** patches are adjusted depending on the size of the ROIs within. If a region’s bounding box is larger than the patch and over half of its area is inside of said patch (and provided that the bounding box dimensions do not exceed the double of the patch’s own dimensions), then the dimensions of the patch are adjusted to encompass the whole ROI (and then resized to the originally intended dimensions of 320×320 pixels). Otherwise, the annotation is only kept if at least 10% of its area is included in the patch. Similar strategies are applied to medium and small-sized bounding boxes (with dimensions over and under half the size of the patch, respectively). It is worth noting though that the 10% minimum area threshold for keeping an annotation is not perfect, and while it is sufficient to include relevant information in some patches (such as the presence of a nucleus, as in Figure 3.10a), there are also situations where this is not the case (Figure 3.10b).

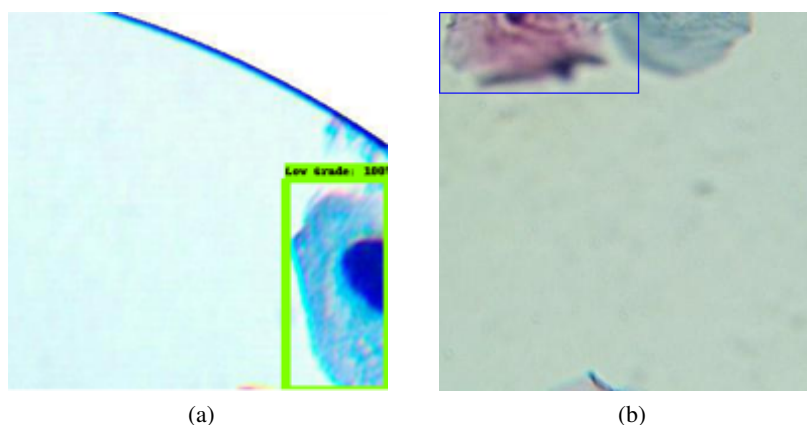


Figure 3.10: Split regions with (a) moderate and (b) insufficient information for detection.

3.2.2 Abnormal Region Detection

We opted to follow a deep-learning-based approach for this stage due to the more promising results of these techniques, as mentioned at the beginning of this chapter. As for the respective development framework, Tensorflow’s Object Detection API [36] was chosen due to its simpler set-up when compared to other similar tools, such as PyTorch [37], as well as its great variety in pre-trained models [38]. From this selection, the RetinaNet architecture with the ResNet50 backbone was singled out due to its good reported performance both in the dataset in which it was pre-trained on (COCO 2017 dataset [39]) and in the literature.

In order to optimize the network’s performance, a random hyper-parameter search process was conducted. Yet, some smaller-scale experiments were performed, detailed in Section A.1, with carefully specified settings, in order to reduce the hyper-parameter search space considered.

The hyper-parameters optimized during this stage were the following:

- **Batch size** - The number of training samples seen before updating the weights of the network.
- **Learning rate** - In the performed experiments, it had a cosine decay, meaning it decreased over the number of steps completed according to a cosine function.
- **Warm up learning rate** - When active, it indicates a starting value for the learning rate from which it increases over a given number of steps until it reaches the actual learning rate value.
- **Score threshold** - Minimum classification confidence score for each proposal.
- **Max detections per class** - The maximum number of detected objects to retain per class.
- **Max total detections** - The maximum number of detected objects to retain across all classes.

The explored parameter values are displayed in Table 3.3. The warm up learning rate values are not shown since they depend on the base learning rate used (whose values are spaced out in a logarithmic scale of base 10). In the cases when a warm up was active (chosen at random), this value was randomly chosen from the remaining, lower values present in the base learning rate list.¹

Hyper-Parameter	Possible Values
Score Threshold	[0, 0.3]
Max detections per class	{8, 10, 12, 14, 16}
Max total detections	[Max detections per class, 20]
Learning Rate	{ 10^{-6} , 3×10^{-6} , 10^{-5} , 3×10^{-5} , 10^{-4} }
Batch Size	{8, 16}

Table 3.3: Deep learning hyper-parameter random search grid.

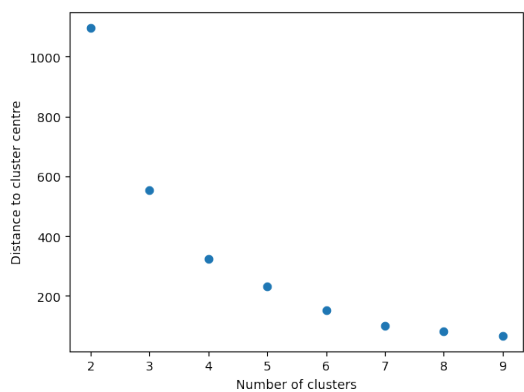
In addition to these, three other hyper-parameters were optimized separately:

- **FPN min and max levels** - The minimum and maximum levels used by the feature pyramid network (FPN); these determine which octaves of the FPN are used for the proposal of object candidates.

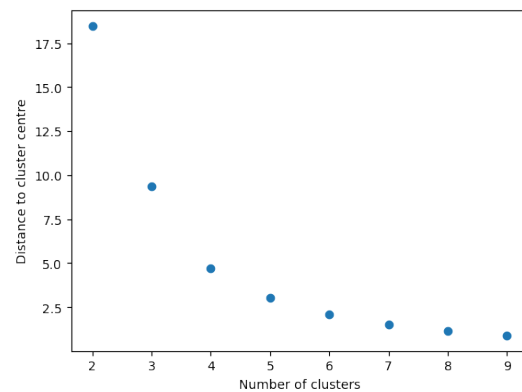
¹If the base learning rate had the lowest value, the warm up learning rate was assigned to 1/10th of the former.

- **Scales per octave** - The number of anchor scales in each level of the FPN.
- **Anchor scale** - The base anchor scale for the FPN from which the scales for each level are calculated.
- **Aspect ratios** - The aspect ratios for the anchors, which define the template shapes used for object proposal.

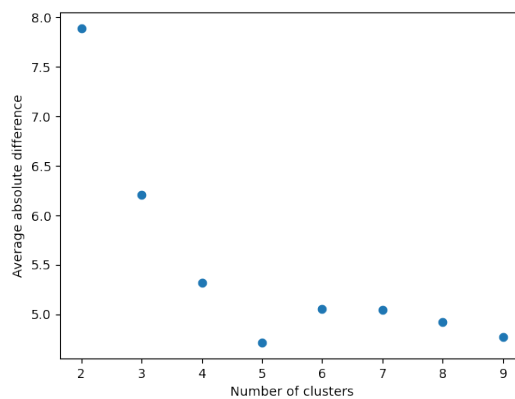
The values for the last two parameters were based on the clustering of the ground-truth bounding boxes scales and aspect ratios for the overall training subset. This clustering was performed using the K-Means algorithm, and their sum of squared distances to the respective cluster centre can be seen in Figures 3.11a and 3.11b for the aspect ratios and scales, respectively.



(a) Aspect ratios' sum of squared distances per number of clusters



(b) Scales' sum of squared distances per number of clusters



(c) Aspect ratios' average absolute difference per number of clusters

Figure 3.11: Sum of squared distances (scales and aspect ratios) and average absolute difference per cluster (aspect ratios) for the 320×320 patches.

In the case of the scales, since Tensorflow's Object Detection API does not allow to specify these values directly when it comes to an FPN, the base scale parameter had to be tuned to reflect the

chosen scales. More specifically, the scale with index i (out of spo scales per octave) associated with octave level l is given by Equation 3.1, with b being the base anchor scale.

$$scale_{ij} = b \times 2^{\frac{i}{spo}} \times \frac{1}{2^l} \quad (3.1)$$

Based on this, it was calculated the base scale value that minimized the average minimum absolute differences between each scale returned by the K-Means algorithm and the scales computed through Equation 3.1, with the help of the generalized reduced gradient (GRG) nonlinear solver. In this case, five scales were chosen, from FPN level three to seven and with one scale per octave, in order to preserve the same number of features maps as in the original implementation of the RetinaNet [40].

As for the aspect ratios, their selection is based not only on the sum of squared distances to the respective cluster centre but also on the average absolute difference between the different values at each K , since these should cover a sufficiently diverse set of object shapes. From Figure 3.11c, it can be observed that the average absolute difference is the lowest when having five clusters, and so this value was discarded even though it presented a good trade-off between the number of clusters and the respective sum of squared distances. The other two cluster numbers immediately before and after this value also achieve a good balance between these factors, and while the sum of squares is naturally lower at the higher K value of six than four, we opted to use the latter since it would also be less taxing on the already limiting total amount of memory available.

A maximum IoU threshold of 0.4 was also set since abnormal regions do not have a high degree of overlap. This parameter is used by non-max-suppression (NMS) proposal filtering and what it entails is that bounding-boxes that have an IoU with other accepted detections higher than the defined threshold are filtered out.

During these experiments, the Adam optimizer was used together with the smooth L1 loss for the localization of the regions, and the focal loss for their classification in two groups:

- Low grade lesion - Composed of the LSIL and ASC-US classes.
- High grade lesion - Composed of the HSIL-SCC and ASC-H classes.

This class composition is related to the severity level of the labels in each group, with the elements of the first and second classes being less and more severe, respectively. This decision was made since there are notable differences between each group, and as such, the network's learning ability might have been hampered if it was only presented with one, all-encompassing class. It also comes with the added benefit of easing the final classification task ahead.

Finally, it also worth noting that even though the network was pre-trained with 640×640 images, the experiments were performed with 320×320 patches to enable the usage of larger training batches.

3.2.3 Nuclei Detection

The most predominant characteristic in the nuclei present in an image is their dark pigmentation, being often the structures with the lowest intensity values. As such, the techniques explored in this work for the detection and segmentation of the nuclei in an abnormal region are based on this property. Moreover, the pre-processing steps taken in this stage were aimed at accentuating this intensity disparity between the nuclei and the other structures.

3.2.3.1 Pre-processing

The first pre-processing operation in this module automatically adjusts the brightness and contrast of the image through the commonly used formula $g(i, j) = \alpha \times f(i, j) + \beta$, where α and β can be viewed as the gain and bias parameters, respectively. They are automatically calculated by defining the output range from 0 to 255. Additionally, the histogram is cut to only include intensities with a frequency of at least 1%. An example of this adjustment can be seen in Figure 3.12.

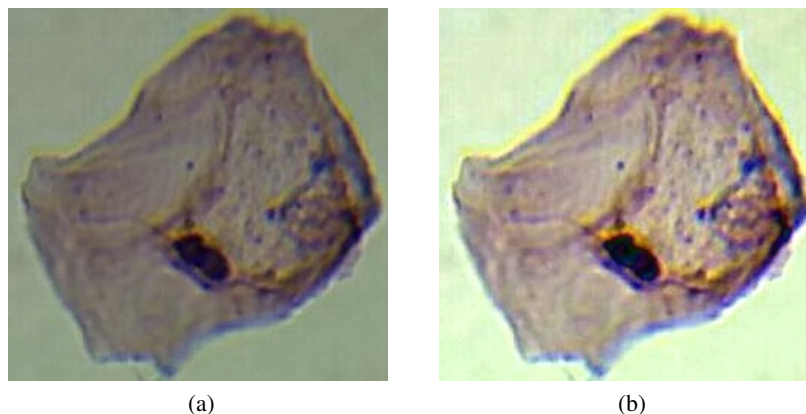


Figure 3.12: ROI before (a) and after (b) automatic brightness and contrast adjustment.

Next, the image is directly converted from the RGB colour space to grey-scale since many contrast enhancement techniques are directed towards single-channel images and, as mentioned at the beginning of Section 3.2.3, the methods explored for the segmentation of the nuclei are also based on their (lower) intensity. Other alternative strategies were tested as well, such as using the individual RGB channels or the value channel from the HSV colour space; however, none of these offered a contrast level greater than the standard grey-scale conversion.

Afterwards, a bilateral smoothing filter is applied to remove any potential noise present in the image while preserving its edges. At this point, the image's contrast is enhanced through histogram stretching by adjusting its intensities so that the minimum and maximum values are 0 and 255, respectively. This was deemed to be the best method for the task at hand since it successfully increased the contrast of the image, making the nuclei stand out even more, and did not produce any artefacts (Figure 3.13). This last characteristic (or lack thereof) was the main reason other

methods were discarded, namely histogram equalization or CLAHE (more details on this can be found in Appendix B).

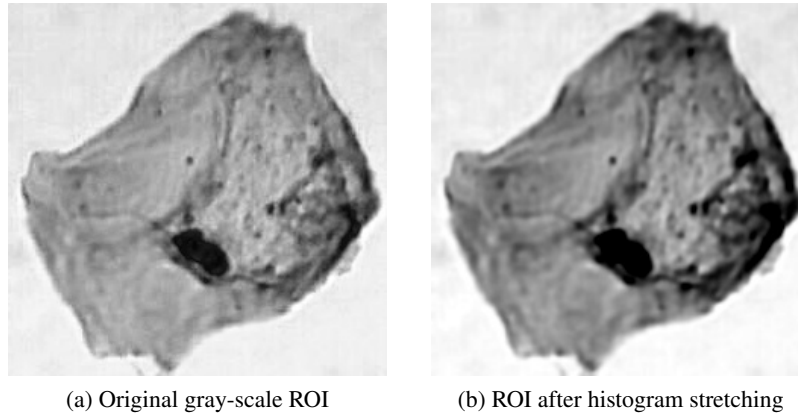


Figure 3.13: Contrast adjustment through histogram stretching.

3.2.3.2 Segmentation

The segmentation phase is focused on generating a binary image with the nuclei present in the mask (and as little noise as possible), followed by a contour detection operation to identify potential nuclei.

Due to its success in the ISBI challenge, the work of Phoulady et al. [41] was used as the basis for the final segmentation strategy. It consists on applying global thresholds of increasing value to the image in an iterative manner (Algorithm 1). At each threshold, each new region must fulfil some criteria before being considered for acceptance. In the original implementation, these consist of a minimum area and solidity. In this context, the solidity of a region refers to the measure of its concavity and is calculated by dividing its area by the area of its convex hull. Since the nuclei dataset used only possesses the bounding boxes of each nucleus (instead of a segmentation mask), the minimum area value was replaced by a minimum bounding box area equivalent to 80% of the minimum bounding box area found in the training subset. Furthermore, a maximum bounding box threshold was added (also based on the corresponding value in the training subset) to avoid picking up large agglomerates of cellular material that might have had a high convexity.

If these criteria are fulfilled, it is then assessed if the segmented nucleus region intersects previously accepted regions. In those cases, the new region is only accepted if it has a greater or equal solidity than each of the intersected regions. If no intersections are detected, then the region is automatically accepted.

Finally, a last filter is applied to all candidate regions in order to discard possible artefacts. This consists on only accepting the ones with a minimum intensity difference between the mean intensity of the region itself and the mean intensity of the exterior area near its boundary. The latter is

calculated by dilating the region with a 5×5 kernel and only considering the area not present in the original region. An example of the final binarization produced by the algorithm is displayed in Figure 3.14.

Algorithm 1 Iterative Thresholding Segmentation Algorithm

```

1: Input: Grey-scale image  $I$ , min solidity  $S$ , min area  $a_1$ , max area  $a_2$ , lower threshold  $t_1$ , higher
   threshold  $t_2$ , step  $s$ , min boundary intensity  $d$ 
2: Output: Binary mask  $N$  with the segmented nuclei
3: for  $t \leftarrow t_1$  to  $t_2$  step  $s$  do
4:    $B \leftarrow I \leq t$ 
5:   for all regions  $r$  in  $B$  do
6:      $area \leftarrow \text{AREA}(\text{bounding box of } r)$ 
7:     if  $area < a_1 \vee area > a_2 \vee \text{SOLIDITY}(r) < S$  then
8:       Skip
9:     end if
10:    if  $r \cap N \neq \emptyset$  then
11:       $reject \leftarrow False$ 
12:      for accepted region  $R$  in  $N$  do
13:        if  $\text{SOLIDITY}(r) < \text{SOLIDITY}(R)$  then
14:           $reject \leftarrow True$ 
15:          Break
16:        end if
17:      end for
18:      if  $reject = True$  then
19:        Skip
20:      end if
21:    end if
22:     $N \leftarrow N \cup r$ 
23:  end for
24: end for
25: for all regions  $r$  in  $N$  do
26:    $boundary = \text{DILATE}(r) \oplus r$ 
27:   if  $\text{MEAN INTENSITY}(boundary) < d$  then
28:     remove  $r$ 
29:   end if
30: end for

```

Some different strategies were also explored for this task (detailed in Appendix B); however, most came with serious drawbacks attached to them. The adopted strategy achieved the most consistent results out of all the ones tested, capturing even nuclei with different intensities with relative ease. However, it also comes with some flaws, namely when dealing with agglomerates of several nuclei. In some of these cases, only one of the nuclei (or part of one) is segmented by the algorithm, possibly because the solidity of the smaller region is larger than the solidity of the whole cluster, as can be observed in Figure 3.15a. One way to mitigate this issue is to add a constant to the new region's solidity (Figure 3.15b). However, this comes with the downside of sometimes picking up

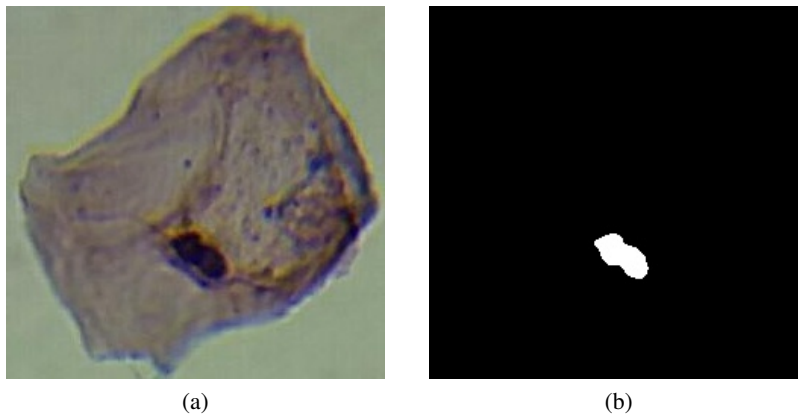


Figure 3.14: ROI before (a) and after (b) binarization through the iterative thresholding algorithm (the nucleus is represented in white).

parts of the nucleus-surrounding region as well, as can be seen in Figure 3.15d. Therefore, this was not performed in the final version of the pipeline. Decreasing the step size or the minimum solidity can sometimes also help in increasing the area of the agglomerate caught. The downside, in this case, is that the number of artefacts caught by mistake also increases.

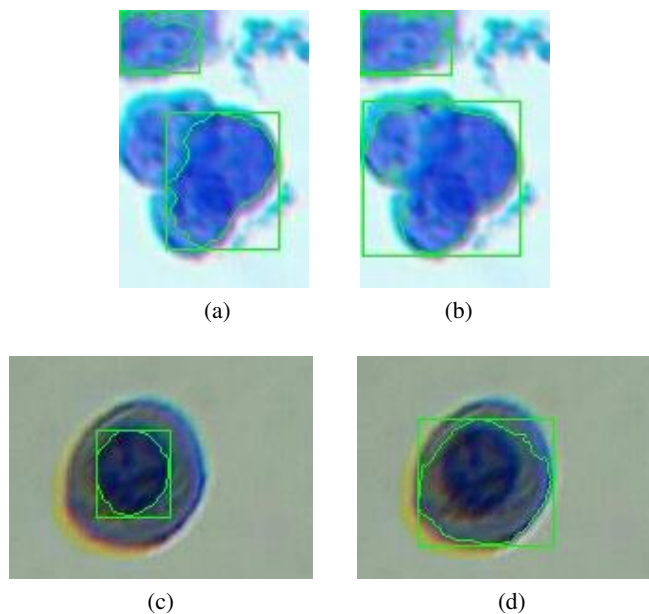


Figure 3.15: Iterative thresholding results with (b, d) and without (a, c) additional solidity constant.

Despite its limitations, this method had the most success overall, and as such, it was chosen as the main segmentation algorithm.

3.2.3.3 Post-processing

As for post-processing operations, the regions returned by the iterative thresholding algorithm are filtered according to two properties: the aspect ratio of the bounding boxes and their eccentricity. Both of these are put in place to eliminate thin streaks of cellular material that might have been caught by accident. More specifically, the bounding boxes of the detected regions must have a minimum and maximum aspect ratio and a maximum eccentricity. In the case of the former two, since it is related to the bounding boxes, both limits are equal to their counterparts in the training subset. The maximum eccentricity though was determined by observation of the outputted results. This measure is determined by first fitting an ellipse around the region and then calculated through Equation 3.2, with a and b being the major and minor axis lengths of the ellipse, respectively. The extreme cases of this property are 0 and 1 and indicate that the shape is a circle or a line segment, respectively, and as such, a maximum value of 0.8 was stipulated.

$$eccentricity = \sqrt{1 - \frac{b^2}{a^2}} \quad (3.2)$$

An example of the final output of this module can be seen in Figure 3.16.



Figure 3.16: Example of a segmented nucleus.

3.2.4 Feature Extraction

After identifying an abnormal region, several features are extracted from it. These can be grouped into three categories: colour, texture and geometrical or shape-related features. Depending on each feature, these are extracted either from the entire region, from each nucleus contained within or from both. Regarding the nucleus-only features, since the number of nuclei in a region is variable, these are fused by taking the average of each one. In the cases when no nuclei are detected, they are replaced with the value -1.

3.2.4.1 Colour Features

The extraction of the colour features is performed in the CIELch colour space (Figure 3.17), which is perceptually uniform, unlike other more commonly used representations, such as RGB and HSV. This means that the perceptual difference between two colours can be represented by the difference between two points in this space [42]. It is composed of three channels: lightness (L), chroma (C) or saturation and hue (H), and in this case, the features of this group are extracted from the chroma and hue channels, **both for the region and the nuclei** contained within it, mainly because visual inspection showed that there can be significant differences between images of different classes when it comes to their perceived colour, as illustrated in Figure 3.18.

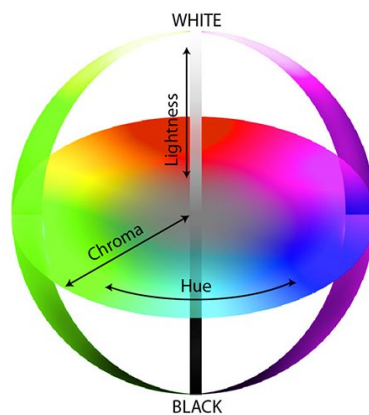


Figure 3.17: Visual representation of the CIELch colour space [43].

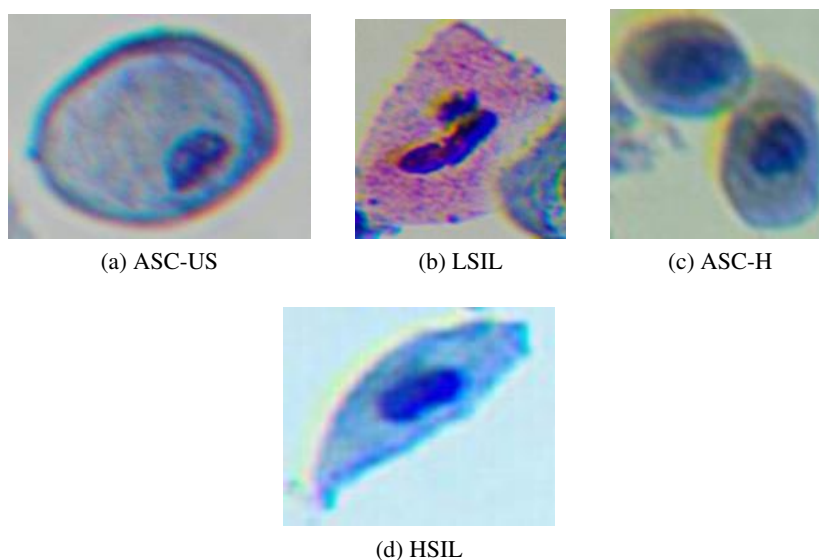


Figure 3.18: Example of different colours between classes.

These features are the following:

- Energy - The overall intensity in each channel, calculated by: $energy = \frac{1}{N} \times \sum_{i=1}^N pixel(i)^2$, with N being the total number of pixels in the considered region.
- Mean - The average color in the region, calculated by: $mean = \frac{1}{N} \times \sum_{i=1}^N pixel(i)$
- Standard deviation - How much the values in each channel deviate from the mean, calculated by: $stdev = \sqrt{\frac{1}{N} \times \sum_{i=1}^N |(pixel(i) - mean)|^2}$
- Skewness - The asymmetry of the color distribution, calculated as the Fisher-Pearson coefficient of skewness by: $skew = \frac{m_3}{m_2^{3/2}}$, where $m_i = \frac{1}{N} \times \sum_{n=1}^N (pixel(n) - mean)^i$
- Range - The difference between the maximum and minimum intensities in a channel.

3.2.4.2 Texture Features

While different abnormality classes can have distinct colour features, this is not always the case, as exemplified in Figure 3.19. Owing to this, other types of features are necessary to further differentiate between each abnormality group's (low and high grade) classes. For instance, the texture appearance of a region, particularly those of the nuclei, is an important factor when analyzing an image since normal ones usually have a uniform chromatin distribution, while some abnormal nuclei have a more heterogeneous appearance.

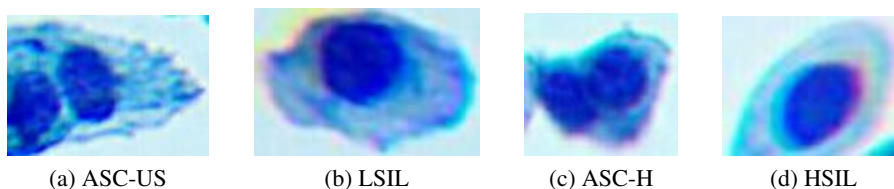


Figure 3.19: Example of similar colours between classes.

In order to produce a compact summary of the texture of **the region and the nuclei**, a set of properties is computed for both based on their gray-level co-occurrence matrices (GLCM). A GLCM is a histogram representation of the co-occurring gray-scale values at a given offset and angle in an image [44]. In this case, four GLCMs are computed for each image or nucleus with an offset of 3 and angles ranging from 0 to $\frac{3}{4}\pi$ radians in increments of $\frac{1}{4}\pi$ radians. Then the means of the following properties are calculated ²:

- Dissimilarity: $\sum_{i,j=0}^{levels-1} P_{i,j} |i - j|$
- Homogeneity: $\sum_{i,j=0}^{levels-1} \frac{P_{i,j}}{1+(i-j)^2}$
- Energy: $\sqrt{\sum_{i,j=0}^{levels-1} P_{i,j}^2}$

² $P_{i,j}$ represents the number of times gray-level j occurs at a certain distance and angle from gray-level i

- Correlation: $\sum_{i,j=0}^{levels-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right]$

In addition to this, the histogram of the local binary patterns (LBP) is computed for each **region** as well. An LBP describes textural information by analysing the points surrounding a central one and tests whether they have a greater or lesser value. In this case, a radius of three pixels from the central point was used in conjunction with a rotation-invariant version of the algorithm. Since the output of this method depends on the dimensions of the image, these are resized beforehand to a size of 120 by 120 pixels, which is approximately the average dimensions of the abnormal region bounding boxes present in the overall training subset, resulting in a total of 26 bins per region. Additionally, in order to keep the aspect ratio of the original image, padding is applied when necessary, as exemplified in Figure 3.20a.

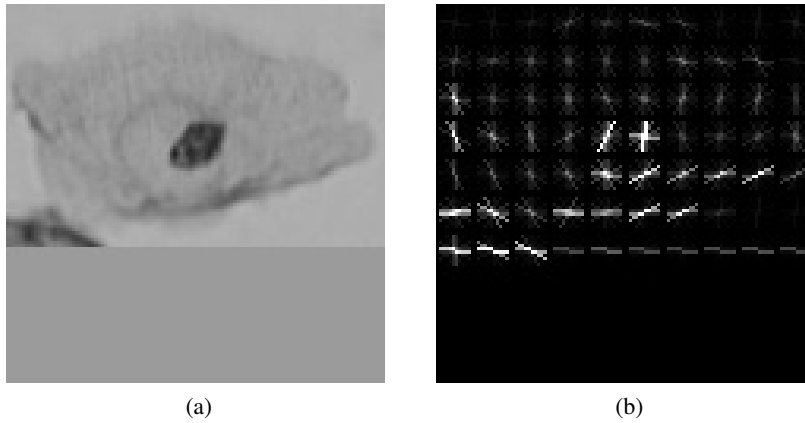


Figure 3.20: Example of rescaled and padded image (a) and its HOG representation (b).

3.2.4.3 Geometrical Features

The shape and size of the cell's nucleus are other key factors when determining if a particular region is indicative of a cervical lesion or not. For instance, abnormal cells tend to have larger nuclei with irregular structures, as opposed to the usual circular shape found in healthy cells [45]. Owing to this, the following features are extracted for each **nucleus**:

- Area - Number of pixels of the nucleus region.
- Bounding box area - The area of the minimum nucleus-enclosing straight rectangle.
- Convex hull area - The area of the minimum convex shape containing the nucleus.
- Maximum and minimum diameters - The maximum and minimum length of a line connecting two border points of the nucleus whilst passing through its centre.
- Equivalent Diameter - The diameter of the circle with the same area as the nucleus, calculated by: $equivalent\ diameter = \sqrt{\frac{4 \times area}{\pi}}$

- Solidity - The measure of the nucleus' concavity, calculated by: $solidity = \frac{area}{convex\ hull\ area}$
- Extent - The ratio between the area of the nucleus and the area of its bounding rectangle: $extent = \frac{area}{bounding\ box\ area}$.
- Minimum enclosing circle area ratio - The ratio between the area of the nucleus and the area of its minimum enclosing circle: $minimum\ enclosing\ circle\ area\ ratio = \frac{area}{\pi \times radius_{min\ enclosing\ circle}^2}$
- Elliptical symmetry - Measures how well the shape of the nucleus can be represented by an ellipse. This is computed by fitting an ellipse around the nucleus and calculating the ratio between the area of the non-overlapping region (between the ellipse and the nucleus) and the total area of both structures, as so: $elliptical\ symmetry = \frac{area_{non-overlap\ region}}{total\ area}$
A perfect elliptical and symmetric nucleus will have an elliptical symmetry of zero, increasing towards one with the nucleus' shape asymmetry and irregularity.
- Compactness Index - Measures how compact the nucleus is, calculated through: $compactness = \frac{equivalent\ diameter}{max\ diameter}$
- Principal axis ratio - The ratio between the minimum and maximum diameter: $principal\ axis\ ratio = \frac{min\ diameter}{max\ diameter}$.
- Bounding box aspect ratio - The ratio between the width and height of the nucleus' bounding box: $bbox\ aspect\ ratio = \frac{bbox\ width}{bbox\ height}$
- Eccentricity - Specifies the eccentricity of an ellipse fitted around the nucleus. It is calculated through Equation 3.2, where a and b are the major and minor axis lengths of the ellipse, respectively.
- Irregularity Index - The difference between the maximum and minimum diameters: $irregularity\ index = max\ diameter - min\ diameter$

It is worth noting that features depending on the perimeter of the nucleus were not considered because they were highly dependent on the segmentation result, whose robustness was below desirable in some cases.

When it comes to the entire **region**, the histogram of gradients (HOG) is computed instead, after resizing the image in the same manner as for the calculation of the LBP histogram, resulting in an additional 800 features. The representation of an image through its HOG can be seen in Figure 3.20b.

3.2.5 Classification

The final classification of a detected region is performed by one of two separate classifiers, depending on the class outputted by the detection network, as illustrated in Figure 3.21. The low grade lesion classifier distinguishes between ASC-US and LSIL instances, whereas the high grade model classifies the regions as ASC-H and HSIL-SCC examples.

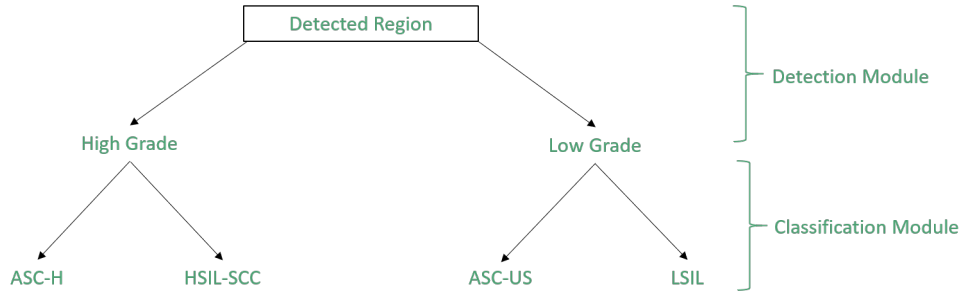


Figure 3.21: Classification Tree.

The dataset used in this stage's experiments followed the same distribution as in the detection phase (same training, validation and test subsets); however, since there are no ground-truth segmentation masks for the nuclei, the features of this type used for training were extracted from nuclei detected on the overall training subset through the process proposed in Section 3.2.3.

With this in mind, based on the reviewed literature, two model types were tested for this stage: the random forest ensemble and the SVM. Both types underwent a process of hyper-parameter optimization, namely a grid search for the SVM and a random search for the random forest. The cause for the difference between each process derives from the higher training times in the case of the random forests due to the large number of trees considered, making a grid search unfeasible in terms of total search time.

To assess the influence of the properties of the nuclei structures in the developed models, two types of feature sets were used in the experiments: one with only features extracted from the detected region as a whole - the baseline set -, and another with added nuclei features. Both types of experiments were performed through a 3-fold cross-validation procedure.

It is also worth noting that the features used by the random forest algorithms remained unchanged while the ones provided to the SVM were normalized according to Equation 3.3, where \bar{x} and s represent the average and standard deviation for that feature, respectively. This was done since SVMs are based around the distance between the data points from different classes, thus not being scale-invariant.

$$x_{normalized} = \frac{x - \bar{x}}{s} \quad (3.3)$$

In the case of the random forest, the hyper-parameters optimized were the following (their possible values can be seen in Table 3.4):

- **Estimator No.** - The number of trees in the forest.
- **Max depth** - The maximum tree depth.

- **Min samples split** - The minimum number of samples required to split a node.
- **Min samples leaf** - The minimum number of samples required to form a leaf node.
- **Max features** - The maximum number of features to consider when searching for the best split.
- **Criterion** - The split quality measuring function.

Hyper-Parameter	Possible Values
Estimator No.	{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000}
Max depth	{3, 8, 13, 18, 23, 29, 34, 39, 44, 50, ∞ }
Min samples split	{2, 0.05 n , 0.1 n , 0.15 n , 0.2 n , 0.25 n , 0.3 n , 0.35 n , 0.4 n , 0.45 n , 0.5 n }
Min samples leaf	{1, 0.05 n , 0.1 n , 0.15 n , 0.2 n , 0.25 n , 0.3 n , 0.35 n , 0.4 n , 0.45 n , 0.5 n }
Max features	{0.5 f , 0.75 f , \sqrt{f} , $\log_2 f$, f }
Criterion	Gini impurity, information gain

Table 3.4: Random forest random search grid, with n being the total number of samples in the training dataset and f the total number of features.

As for the SVM, the optimized hyper-parameters were:

- **Kernel** - The kernel function used by the algorithm.
- **C** - The regularization parameter.
- **Gamma** - Coefficient for the radial basis function (RBF), polynomial and sigmoid kernels.
- **Degree** - Degree of the polynomial kernel function.

As previously mentioned, the optimization process was performed through grid searches, more specifically one for each classification algorithm and each kernel type, with the nuclei features or not. The tested values for each parameter can be seen in Table 3.5.

Finally, the output of this stage (and the pipeline) is the label not only of the detected regions but of the respective patches, source images and samples as well. This is done in a hierarchical fashion in which the label of a parent element, i.e., composed by multiple sub-elements, is equal to the highest severity label found in one of the respective sub-elements, with ASC-US being the less severe (excluding the cases where no abnormal regions are detected), followed by the classes LSIL, ASC-H and HSIL-SCC, in this order. In this manner, it is possible to perform an evaluation closer to the methods used by human specialists.

Hyper-Parameter	Possible Values
Kernel	RBF, polynomial, sigmoid, linear
C	{0.00001, 0.0001, 0.001, 0.1, 1, 10, 100, 1000, 10000}
Gamma	{0.00001, 0.0001, 0.001, 0.1, 1, 10, 100, 1000, 10000, $\frac{1}{f \times \text{variance}_N}$, $\frac{1}{f}$ }
Degree	{3, 5}

Table 3.5: SVM grid search hyper-parameter values, with f and variance_N being the total number of features and variance of the fitted data, respectively.

Chapter 4

Experimental Setup and Results

This chapter provides an in-depth performance evaluation of the experiments detailed in Chapter 3, first for the detection and classification modules individually and then for the complete system.

4.1 Abnormal Region Detection

4.1.1 Evaluation Metrics

The system's performance for the detection task is assessed through the COCO competition metrics [46], which include the mean average precision at specific intersection over union (IoU) thresholds (mAP@KIoU) and the average recall at a maximum number of detections per image (AR@K).

In the case of the former, this means that a detected region is considered as a true positive (TP) if its IoU with a ground truth box is greater or equal than K. The IoU is a measure of the overlap between two bounding boxes and is computed by dividing the area of their intersection by the area of their union, as depicted in Figure 4.1.

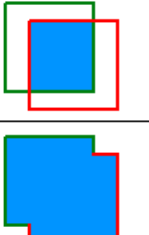
$$IoU = \frac{\text{area of overlap}}{\text{area of union}} = \frac{\text{area of intersection}}{\text{area of union}}$$


Figure 4.1: Intersection over union (IoU) representation [47].

On the other hand, if a detected region does not fulfil this criterion with any ground truth bounding box, it is considered a false positive (FP). Similarly, if a ground truth bounding box does not have

an IoU greater or equal than K with any detected region, a false negative (FN) occurs. Through these values, it is possible to calculate other metrics, such as the precision (equation 4.1) and recall (equation 4.2).

$$Precision = \frac{TP}{TP + FP} \quad (4.1)$$

$$Recall = \frac{TP}{TP + FN} \quad (4.2)$$

The average precision (AP) can then be calculated by averaging the precision values across all recall levels, and the mAP by taking the mean AP over all classes. In the context of this work, it is given increased attention to the mAP at an IoU of 0.5 since abnormal regions do not always respect individual object’s boundaries (either by their very nature or due to lower quality annotations), making it more difficult to accurately pinpoint its borders.

As for the average recall, it is computed by averaging the recall scores across all IoU levels considered. Since the number of abnormal regions in a patch is usually less than 10, it is given more focus to the AR@10 score.

4.1.2 RetinaNet with ResNet50 backbone

All experiments for this stage were executed in a Linux Server featuring one NVIDIA T4 GPU with 16GB of RAM. As mentioned in Section 3.2.2, the hyper-parameter combinations tested were obtained through a random search and their values for each experiment can be seen in Table 4.1. The number of steps performed for each experiment is the equivalent of approximately 300 epochs in total with ten epochs of warm up (when active).

Exp	Batch Size	LR	WU LR	Min Score	Max Class Detect	Max Detect
R1	16	0.00003	0.00001	0	8	10
R2	8	0.00001	0.000001	0.1	12	16
R3	16	0.0001	N.A	0.2	14	18
R4	16	0.000001	N.A	0.1	16	16

Table 4.1: RetinaNet with ResNet50 backbone experiments. Experiment (Exp), Learning Rate (LR), Warm Up (WU), Detections (Detect).

The mean mAP@0.5IoU and AR@10 across the three validation sets are displayed in Table 4.2 for each experiment, alongside their respective standard deviations.

Exp	Avg mAP@0.5IoU	Standard Dev mAP@0.5IoU	Avg AR@10	Standard Dev AR@10
R1	0.15	0.02	0.42	0.01
R2	0.11	0.03	0.24	0.03
R3	0.08	0.02	0.16	0.02
R4	0.1	0.02	0.23	0.01

Table 4.2: RetinaNet with ResNet50 backbone experiments’ 3-fold cross-validation results. Experiment (Exp), Average (Avg), Standard Deviation (Standard Dev).

As can be observed, experiment R1 achieved the best mAP@0.5IoU and recall@10 of 0.15 and 0.42, respectively. The most likely reason for these relatively low scores is probably related to the characteristics and general quality of the dataset, as will be discussed in the following section. Even so, there are other improvements that could help achieve a better performance, namely in terms of the hyper-parameter combinations, as only a limited set of different experiments was carried out due to the long training times of the networks. Increasing the number of training steps could also lead to better performances being achieved, but probably not in a significant manner, as both the mAP and recall plateau early on (Figure 4.2). The specific input image sizes considered might also contribute to the models’ low performances, so other patch sizes between the ones used to fine-tune the network (320×320) and the ones it was pre-trained on (640×640) could also be worth exploring.

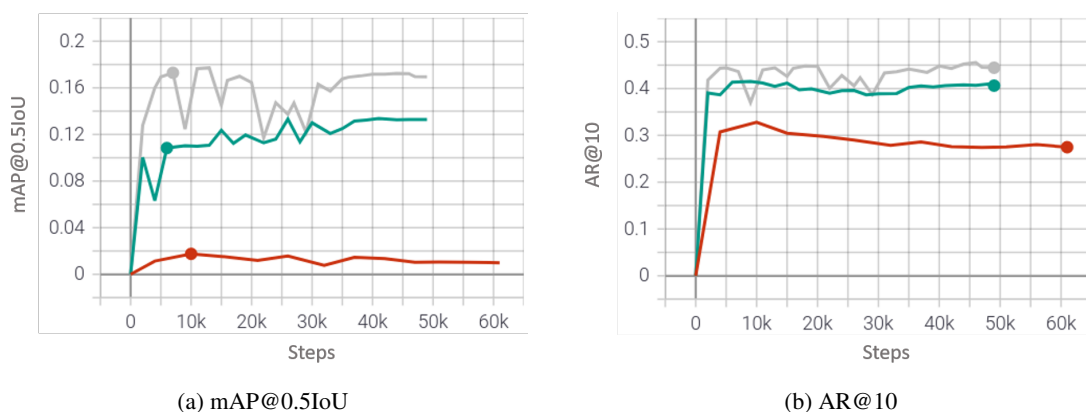


Figure 4.2: RetinaNet with ResNet50 backbone R1 experiments (green and grey for folds one and two, respectively) and test set evaluation (red) metrics.

Nevertheless, since the R1 model achieved the best cross-validation mAP and AR, its hyper-parameter combination was chosen for the final detection model as, due to the long training times of the networks, it was not possible to train every combination on the overall training subset. When evaluated on the test subset, the R1 model (now trained on the overall training subset for 61800

steps, with 2000 of which being performed in the warm up phase) achieved a $\text{mAP}@0.5\text{IoU}$ of 0.01 and an $\text{AR}@10$ of 0.27 (Figure 4.2).

While these results are much lower than the ones obtained in the cross-validation, they can be in part explained by the composition of the test subset. As mentioned in Section 3.2.1, only the train and validation patches are adjusted to encompass the entirety of some of the bounding boxes. In the test set, this is not performed in order to more accurately represent a real scenario. However, this also leads to a greater number of bounding boxes that are partitioned into smaller ones, whereas in the other subsets this would not have happened since the patch would have been enlarged instead. While this in itself is not necessarily a negative aspect, it also leads to an increased number of bounding boxes without any relevant information regarding the abnormal structure since the decision to take into account a fraction of a bounding box is based only on the overlapping area between itself and the respective original annotation. As expected, such areas are then missed by the algorithm, negatively impacting its performance metrics.

4.1.3 Summary

As mentioned in the previous section, the R1 model (whose hyper-parameter settings are displayed in Table 4.1) was chosen as the final system detector of abnormal regions as it obtained the best cross-validation results out of all the other experiments. However, the overall results were fairly low, especially in the test set, where the increased presence of uninformative ground-truth bounding boxes significantly limits the algorithm's performance. A careful analysis of this set revealed that a significant portion of the annotations contained in it (sensibly 25%) fit this description, having little to no information relevant for the tasks at hand, making their identification (and classification) extremely difficult (or near impossible in some cases), even for trained human specialists. Although a stricter overlap threshold could reduce the amount of irrelevant ground truth annotations, experiments conducted with larger threshold values demonstrated that they resulted in the disposal of some truly abnormal regions. Due to this, the 10% overlap threshold was ultimately kept because it provided a reasonable trade-off between the inclusion of all the relevant regions and the exclusion of uninformative annotations. Nonetheless, in the future, a more informed method should be developed, taking into account other properties of the region, such as its texture, for example, to decide whether to keep a bounding box or discard it.

Apart from this, there are other situations which help to explain the low performance. For instance, there are cases where the network identifies certain regions as abnormal even though they are not annotated in that manner, even when sometimes they present characteristics similar to the ones found in abnormal cells, such as enlarged nuclei (Figure 4.3). This could be due to the fact that abnormal regions that do not have a clear label are not annotated in the dataset, as mentioned in Section 3.1. Another aspect that might lead to this is the presence of multiple types of squamous cells that possess different abnormality characteristics, making it difficult to distinguish, for instance, between a normal example of a particular type and an abnormal one from a different type.

An extra step to categorize each cell before outputting a final label could prove useful to mitigate this issue.

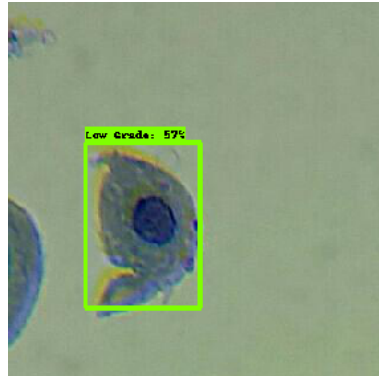


Figure 4.3: Misidentified normal region with enlarged nucleus.

Finally, it is also worth pointing out the instances where an abnormal region was successfully identified but assigned the wrong label, revealing a certain difficulty by the network in differentiating between the two in some situations. Furthermore, the classification loss obtained on the validation and test subsets tends to slightly increase over time, as opposed to the one calculated during training (Figure 4.4), revealing that there might be a certain degree of overfitting, probably due to the relatively small number of examples. Additional evidence of this can be observed in Figure 4.2b, where the AR@10 tends to decrease after around 10000 steps (on the test set).

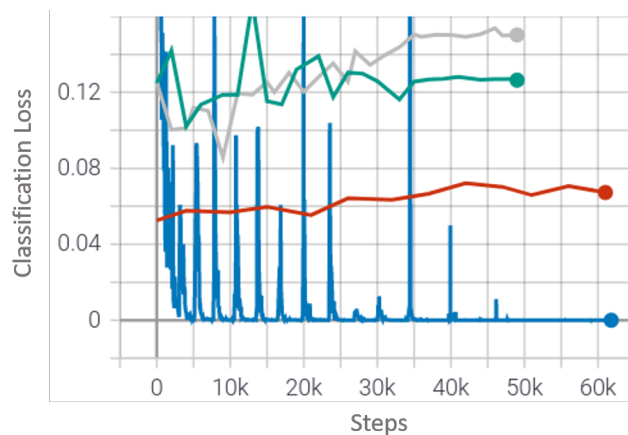


Figure 4.4: Classification loss for the RetinaNet with ResNet50 backbone R1 experiments (green and grey for folds 1 and 2, respectively), train (blue) and test set evaluation (red).

In summary, despite the dataset being the most limiting factor, either due to the presence of inconsistent annotations, uninformative partial regions or the small number of examples, it is reasonable to say that there could still be relative performance gains through other means, such as with other hyper-parameter combinations or by using patches with the same dimensions as the ones in which

the network was pre-trained on, given that there would still be enough computational power to accommodate reasonable batch sizes.

4.2 Classification

4.2.1 Evaluation Metrics

The main metrics used to evaluate the classification performance are the precision (Equation 4.1), recall (Equation 4.2) and the F1 measure, which represents a weighted harmonic mean of the previous two metrics [48] and is computed according to Equation 4.3.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4.3)$$

These three values are **calculated for each class** (pertinent to each type of classifier) and then averaged, i.e., their final value is equal to the unweighted mean between all classes so as to ignore possible imbalances between them. In addition to this, the overall accuracy is also calculated through Equation 4.4.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.4)$$

4.2.2 Random Forest

As mentioned in Section 3.2.5, two types of models were tested for the final classification of an abnormal region: the SVM and the random forest ensemble. Both underwent separate processes of hyper-parameter optimization for each type of classifier (low or high grade) and either with or without using the features extracted from the detected nuclei. In the case of the random forest, multiple random searches were performed (the possible values for each hyper-parameter are displayed in Table 3.4), more specifically, 100 hyper-parameter combinations were tested for each (classifier type)-(feature set) pair, and the best parameter values obtained for each one can be seen in Table 4.3.

It is possible to observe that the low grade classifier had the best performance whilst considering the entire feature set, as opposed to the high grade models. One possible interpretation for this is that a substantial portion of the features extracted share similar values between the low grade classes (LSIL and ASC-US), and as such, the classifier needs to take a larger number of features into account to properly differentiate between them. The cross-validation results for these experiments are displayed in Table 4.4. From its observation, it is clear that the use of nuclei features contributed to a better performance for both types of classifiers, with an average increase of approximately 4% across all metrics.

Exp	Classifier	Nuclei Feats	Max Feats	Max Depth	MSL	MSS	Estimators	Criterion
RF1	Low Grade	No	f	3	$0.05n$	$0.05n$	500	gini
RF2	High Grade	No	$0.5f$	39	$0.3n$	$0.5n$	100	entropy
RF3	Low Grade	Yes	f	13	$0.15n$	$0.15n$	700	gini
RF4	High Grade	Yes	$0.5f$	8	$0.05n$	$0.15n$	600	entropy

Table 4.3: Random forest’s random searches best parameters, with n and f being the total number of samples and features, respectively. Experiment (Exp), Feats (Features), Min Samples Leaf (MSL), Min Samples Split (MSS).

Exp	Accuracy	Precision	Recall	F1
RF1	0.74	0.74	0.73	0.73
RF2	0.80	0.80	0.81	0.79
RF3	0.79	0.79	0.79	<u>0.79</u>
RF4	0.82	0.82	0.84	0.82

Table 4.4: Random forest’s random searches cross-validation results. Experiment (Exp).

As RF3 and RF4 obtained the best results, its hyper-parameter combinations were used to train two new models on the overall train subset. Their performance on the test data can be seen in Table 4.5.

Classifier	Accuracy	Precision	Recall	F1
Low Grade	0.77	0.39	0.49	0.44
High Grade	0.58	0.68	0.59	0.52

Table 4.5: Random forests’ results on the test set.

As evidenced, the performance here was significantly lower than in the cross-validation experiments. It is possible to see why by observing the confusion matrices (Figure 4.5), which show that both classifiers had difficulty in picking up examples belonging to the (respective) most serious class. The low grade model stands out here since it failed to detect any true LSIL region, showing that it was not able to achieve a good generalization regarding this class. And while its accuracy remained close to the one obtained in experiment RF4, this is easily explained by the fact that the number of ASC-US examples is significantly higher than the number of LSIL images, and the vast majority of the former were successfully identified. On the other hand, since just over half

the high grade examples belong to the HSIL-SCC class (and most of them were misclassified), the accuracy for this model is lower, despite achieving better recall and precision scores.

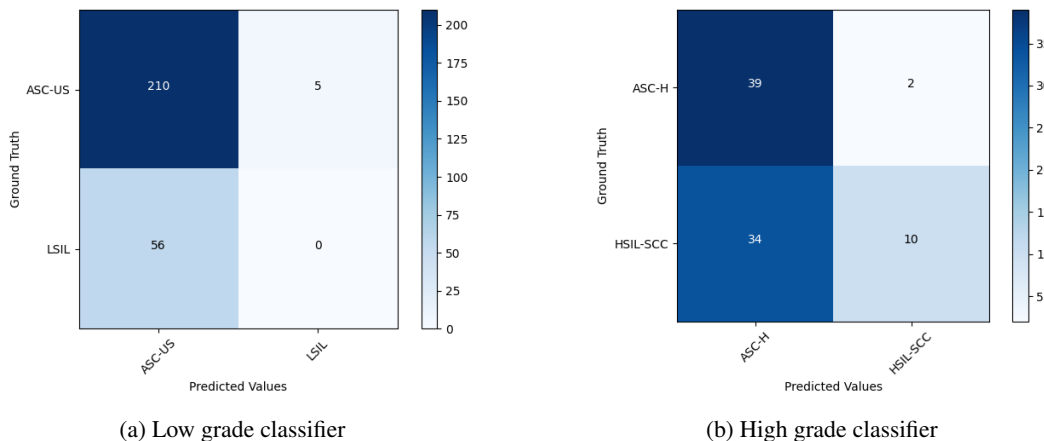


Figure 4.5: Random forests' test set confusion matrices.

Nonetheless, the random forest ensemble makes it possible to extract some insights regarding the importance of each feature. This is briefly discussed in Appendix C.

4.2.3 SVM

Besides the random forest ensemble, the SVM algorithm was also tested for the final classification task, undergoing multiple grid searches to optimise its hyper-parameters. Their possible values are displayed in Table 3.5, and similarly to the random searches performed with the random forests, these experiments were performed for each (classifier type)-(feature set) pair. However, unlike in the case of the other algorithm, multiple grid searches were performed for each of these pairs since not all hyper-parameters are relevant for every kernel type. More specifically, the gamma values were tested for the RBF, polynomial and sigmoid kernels, the degree for the polynomial kernel and the C parameter for all of them. These experiments were conducted in a 3-fold cross-validation manner and the hyper-parameter values that achieved the best performance for each experiment are displayed in Table 4.6, and the respective results in Table 4.7.

As with the random forest experiments, the use of nuclei features here also led to a higher or equal performance in every case, with an average increase of approximately 2% across all metrics, further supporting their relevance to the final classification. Accordingly, the models that achieved the highest results made use of these features, and in the case of the high grade images, two classifiers obtained the same results (SVM4 and SVM12) while using linear and sigmoid kernels; for the low grade examples, the best hyper-parameter combination used an RBF kernel (SVM15). These three models were then trained and tested on the overall train and test subsets, the results of which can be seen in Table 4.8. Similarly to the random forest tests, the results here were also lower than those obtained in the cross-validation. However, they were higher than the scores

Exp	Classifier	Nuclei Feats	Kernel	C	Gamma	Degree
SVM1	Low Grade	No	Linear	0.001	N.A	N.A
SVM2	High Grade	No	Linear	0.001	N.A	N.A
SVM3	Low Grade	Yes	Linear	0.001	N.A	N.A
SVM4	High Grade	Yes	Linear	0.001	N.A	N.A
SVM5	Low Grade	No	Polynomial	0.00001	0.1	3
SVM6	High Grade	No	Polynomial	0.00001	0.1	3
SVM7	Low Grade	Yes	Polynomial	0.00001	0.1	3
SVM8	High Grade	Yes	Polynomial	1.0	$\frac{1}{f}$	3
SVM9	Low Grade	No	Sigmoid	10	0.0001	N.A
SVM10	High Grade	No	Sigmoid	1	$\frac{1}{f}$	N.A
SVM11	Low Grade	Yes	Sigmoid	10	0.0001	N.A
SVM12	High Grade	Yes	Sigmoid	10	0.0001	N.A
SVM13	Low Grade	No	RBF	100	0.00001	N.A
SVM14	High Grade	No	RBF	100	0.00001	N.A
SVM15	Low Grade	Yes	RBF	10	0.0001	N.A
SVM16	High Grade	Yes	RBF	1.0	0.001	N.A

Table 4.6: SVM’s grid searches best parameters, with f and $variance_N$ being the total number of features and variance of the fitted data, respectively. Experiment (Exp), Features (Feats).

obtained by the random forests on the same set. The confusion matrices also show that, despite still misclassifying the majority of LSIL examples, the low grade model still managed to correctly label some of these images, unlike in the case of the random forests (Figure 4.6). As for the high grade classifiers, they both obtained the same scores on the test set as well; however, the model with the sigmoid kernel (SVM12) was deemed the most appropriate since it was the fastest of the two and it does not rely on future examples being linearly separable.

4.2.4 Summary

As demonstrated through the cross-validation experiments performed, the use of nuclei features does indeed increase the performance of both types of algorithms, and as such, they are considered in the final version of the pipeline. The cross-validation results also showed that the random forests performed better than the SVMs during those experiments, with a 10% and 4% difference in F1 between the low and high grade classifiers, respectively. Yet, the results obtained on the test set revealed the opposite, with the low and high grade SVMs having a 14% and 4% F1 increase,

Exp	Accuracy	Precision	Recall	F1
SVM1	0.66	0.66	0.66	0.65
SVM2	0.76	0.77	0.78	0.76
SVM3	0.69	0.68	0.68	0.67
SVM4	0.78	0.78	0.80	0.78
SVM5	0.65	0.62	0.62	0.61
SVM6	0.65	0.67	0.67	0.65
SVM7	0.65	0.64	0.62	0.61
SVM8	0.68	0.71	0.71	0.68
SVM9	0.66	0.66	0.66	0.65
SVM10	0.77	0.78	0.79	0.77
SVM11	0.69	0.68	0.68	0.67
SVM12	0.78	0.78	0.80	0.78
SVM13	0.67	0.66	0.66	0.66
SVM14	0.78	0.78	0.79	0.77
SVM15	0.70	0.70	0.70	<u>0.69</u>
SVM16	0.78	0.79	0.80	0.77

Table 4.7: SVM’s grid searches cross-validation results. Experiment (Exp).

HP Combo	Classifier	Accuracy	Precision	Recall	F1
SVM15	Low Grade	0.80	0.67	0.57	0.58
SVM4	High Grade	0.61	0.74	0.62	0.56
SVM12	High Grade	0.61	0.74	0.62	0.56

Table 4.8: SVMs’ results on the test set. Hyper-parameter (HP) combination (combo).

respectively, over the random forest classifiers. However, these results were lower than the ones obtained during cross-validation (for both algorithms), particularly for the most severe classes in each group. Even though there might be multiple factors responsible for this, such as eventual inconsistencies regarding the nuclei features due to the segmentation method used, it is worth remembering that the most limiting factor is likely the dataset itself, especially the subset used for the final tests of each classifier, as discussed previously in Section 4.1.3. Nevertheless, as the SVM classifiers achieved better results on that subset and thus demonstrated a better generalization

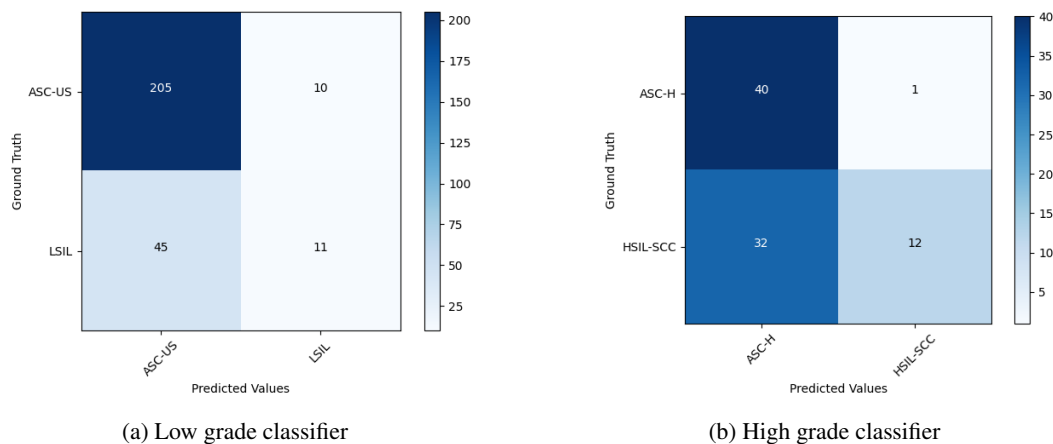


Figure 4.6: SVM's test set confusion matrices.

ability, they were chosen as the final pipeline classifiers.

4.3 Complete System

The complete pipeline was evaluated using the recall, precision and F1 measures at four different levels: regions of interest, patches, source images and samples. This is accomplished by assigning the most severe label detected in a given level to the level immediately above in the same image hierarchy. For example, the label attributed to a patch is equivalent to the most serious one that was detected in all of its abnormal regions. This process is then repeated for each patch and source image belonging to the same sample, where its final classification will be equal to the most severe label found in all of its source images. The reasoning behind this method is to try to mimic (to some extent) the evaluation process performed by the human specialists since their final diagnosis relies on the analysis of multiple images and specific regions within them for each patient.

The ground truth labels for the patches, source images and samples¹ are assigned in a similar manner, with the only difference being that the base region labels from which the other ones are derived are the actual specialist annotations. The ground truth associated with a detected abnormal region on the other hand is equal to the label of the specialist-annotated region with the highest IoU between itself and the detected abnormal region, provided this value exceeds 0.5. Otherwise, it is considered a false positive and assigned a negative (or "normal") ground truth label.

With this in mind, the results obtained on the test subset are present in Table 4.9 for each level. As evidenced, the performance for the abnormal regions is very low. However, there is a noticeable improvement in the higher levels, as the individual errors pertaining to the regions cease to have such a significant impact.

¹The sample ground truth labels are also directly assigned by the specialists, and the output of the method detailed above was validated against these to ensure a match between them.

Level	Precision	Recall	F1
Abnormal Region	0.01	0.01	0.01
Patch	0.20	0.21	0.20
Source Image	0.24	0.23	0.20
Sample	0.20	0.40	0.27

Table 4.9: System test results per image hierarchy level.

However, the performance is still relatively low, even at the sample level. By analysing the individual predictions for each one (Table 4.10), it is possible to observe that the system only correctly classified the high grade examples while attributing the other ones a higher severity label than their actual one. This is linked to the method described above in which a level’s classifications are passed up in the image hierarchy since it only takes one example with a severity higher than the other ones for the following level to be classified with that more severe label. This is the case, for example, of sample 1, in which only two abnormal regions were classified as HSIL-SCC, while the vast majority of the other detected regions were classified as ASC-US (the ground truth label for that sample). And even though this method is similar to the one used by the human specialists, this drawback is not as evident in that case since they usually need multiple instances of a class (or a clear example of one) before attributing a final diagnosis. However, these criteria are not well defined, instead being more of a subjective nature derived from the observation of the images, and as such, it was not possible to replicate them in the present system.

Sample	Prediction	Ground Truth
1	HSIL-SCC	ASC-US
1	LSIL	Normal
3	HSIL-SCC	HSIL-SCC
4	ASC-H	Normal
5	HSIL-SCC	LSIL
6	HSIL-SCC	HSIL-SCC
6	ASC-H	ASC-H

Table 4.10: Sample labels predicted by the system (test subset) and respective ground truth values.

In addition to this, the output of the detection module was evaluated separately in the same fashion as the full system. The results, present in Table 4.11, show that the system’s performance is already quite limited at this stage. This prompted yet another evaluation of the system, but this time only

taking into account the examples correctly classified by the network (as low or high grade) for the full classification afterwards. The results of this experiment are presented in Table 4.12 and were substantially higher than the ones in Table 4.9, especially at the sample level, where **all** examples correctly classified by the RetinaNet were correctly classified by the SVMs as well.

Level	Precision	Recall	F1
Abnormal Region	0.02	0.03	0.02
Patch	0.36	0.38	0.37
Source Image	0.38	0.43	0.38
Sample	0.17	0.33	0.22

Table 4.11: Detection module test results per image hierarchy level.

Level	Precision	Recall	F1
Abnormal Region	0.33	0.69	0.34
Patch	0.44	0.76	0.44
Source Image	0.64	0.57	0.50
Sample	1.0	1.0	1.0

Table 4.12: System test results per image hierarchy level (when considering examples only correctly identified by the detection module).

When combining all of this with the separate module evaluations performed in the previous two sections, it becomes clear that the main performance bottleneck is in the detection phase. As discussed in Section 4.1, there can be multiple reasons for this; however, as in the classification stage, the most limiting factor is probably the dataset itself. From the subjective and sometimes incoherent annotation process to the existence of bounding boxes with very little information due to the annotation splitting process described in Section 3.2.1, a difficult situation is created which hinders not only the training of the algorithms but their proper evaluation as well.

Chapter 5

Conclusions

The early detection of cervical cancer or pre-cancerous lesions is closely related to the successful treatment of the disease. As a consequence, screening methods are very effective as a prevention measure. This is usually done through the analysis of pap smears, which is a laborious task when performed manually, resulting in the research and advent of automatic methods. Motivated by this, this work explored the combination of deep learning and conventional machine learning algorithms for the detection and classification of abnormal regions in cervical cytology slides.

The proposed system made use of a RetinaNet with a ResNet50 backbone for the first task, which, at the same time, classified the detected regions in two severity groups (low or high grade). A multitude of geometrical, colour and texture features were then extracted from each region and from the nuclei within them, which were previously segmented through an iterative thresholding algorithm. These features were then fed to two SVM classifiers, one for each severity group, for the final classification of each region. To obtain a final diagnosis outcome, this classification was then propagated up through the image hierarchy until the sample/patient level according to the most severe class detected, where it achieved a precision, recall and F1 score of 0.20, 0.40 and 0.27, respectively.

In addition to this, other algorithms were investigated as well, such as another RetinaNet architecture with a different backbone (MobileNetV1) for the detection of abnormal regions. However, while the RetinaNet with the ResNet50 backbone obtained better results than the one with the MobileNetV1, it is not possible to directly compare their results as they made use of datasets with different characteristics. The random forest ensemble was also tested for the classification of the detected regions; however, the SVMs had better results on the test set, demonstrating a better generalization ability and thus being chosen as the system's final classifiers. In addition to this, other strategies for the nuclei segmentation task were also explored, such as adaptive thresholding and K-Means clustering, but these too were discarded in favour of the iterative thresholding algorithm as it produced more consistent results.

Despite the modest results obtained, it was demonstrated that deep learning and conventional machine learning algorithms can be combined in a seamless manner for the detection and classification of objects in an image, even in scenarios as complex as cervical cytology. As new related studies further shift towards uniquely deep learning methodologies, their combination with more traditional machine learning algorithms in hybrid pipelines proved to have potential, even if there is still room for improvements.

5.1 Limitations and Future Work

As demonstrated, the system's performance is far from ideal, both in terms of precision and recall. The analysis of the independent modules reveals that the main bottleneck is in the detection stage, where the network struggles to correctly identify the annotated abnormal regions whilst at the same time marking other (normal) ones as such. The abnormality classifiers fared much better but still not well enough to be applied to a real-life situation since it would lead to too many false positives and negatives, as indicated by the precision and recall levels.

It is considered that the most significant limitation of these algorithms is related to the dataset used, as it has many peculiarities that negatively impact their performance. For instance, the annotation process has a certain degree of subjectivity that can not be ignored. This means that a region identified by one specialist can be discarded by another one or classified differently. Unfortunately, this issue is very difficult to solve, as having only one specialist annotate every image is unfeasible due to the volume of work and time needed.

This is also connected with another problem: the lack of annotations regarding regions with abnormal characteristics without a clear label, i. e., which do not clearly fit a specific lesion class of the Bethesda system. It not only decreases the system's sensitivity to these regions but also leads to a higher number of false positives during evaluation if they are picked up by the detection module. One possible way to overcome this without assigning a definitive label to the region would be to create a new "unknown" class composed of these uncertain instances. In this manner, such cases would still be learned by the algorithms and the specialists notified of their occurrence for further examination.

Another aspect to consider is the presence of multiple squamous cell types that possess different abnormality characteristics, making it, at times, easy to mistake a normal cell of a particular type for an abnormal one of a different category. Future approaches should be explored to differentiate between each cell type before evaluating their respective abnormality level.

It is also important to note the uninformative bounding boxes present in the dataset originating from the partitioning of an image into patches. Since this method only relies on the ratio between the area of the smaller bounding box and the area of the original annotation to decide whether to take the former into account, certain situations arise where the bounding box does not contain

any relevant information for its detection and classification, negatively impacting the training and evaluation of the algorithms. In the future, other methods should be investigated to avoid these cases, for example, by analysing the texture of the region, as an empty bounding box will have a more homogeneous texture in comparison with another covered by cellular material.

The proposed nuclei segmentation method could also be improved, especially for the cases when there is a high degree of overlap between them. As it was described, the algorithm does not attempt to identify each individual nucleus in a cluster, which could compromise certain features extracted afterwards. Possible future approaches to tackle this could involve segmenting single nuclei and clusters separately by, for example, utilizing different thresholds for the minimum solidity accepted, and applying further processing and algorithms to the latter cases, such as the watershed method, in order to segment each nucleus in the cluster.

Finally, it would also be worth exploring additional algorithms for the detection and classification tasks, such as the Faster R-CNN, provided there are enough computational resources, and extreme gradient boosting (XGBoost) ensembles, respectively, and compare their performance to that of the ones tested in this work.

Appendix A

Abnormal Region Detection

A.1 Initial Experiments

As mentioned in Section 3.2.2, some smaller-scale experiments were conducted before the random search hyper-parameter optimization of the RetinaNet with the ResNet50 backbone in order to decrease the considered search space. These experiments were performed using 640×640 patches since the most promising models made available by TensorFlow were pre-trained with images of the same size. However, this factor, when coupled with the heavier nature of some of these networks, such as the RetinaNet with a ResNet50 backbone and the Faster R-CNN, resulted in only being able to achieve extremely small batch sizes. Subsequently, in order to increase the range of possible values for this parameter, a lighter backbone was chosen for the RetinaNet architecture, more specifically a MobileNetV1, which allowed for increased batch sizes in the same circumstances. But even so, since the maximum size was still fairly small, this parameter was fixed at that value (7), as smaller sizes might have led to an extremely slow network convergence.

These initial tests were performed on a single fold for time-saving purposes and using only one class composed by the aggregation of every final label. The hyper-parameter combinations can be seen in Table A.1 for each experiment, alongside the mAP@0.5IoU and AR@10 after 10000 steps.

As a first experiment, a relatively small learning rate of 0.00001 was used with a cosine decay. No warm up was performed and no minimum score was imposed, but a maximum IoU threshold of 0.4 was established since abnormal regions do not have a high degree of overlap between them. This threshold is used by NMS proposal filtering and it results in bounding boxes that have an IoU with other previously accepted bounding boxes higher than this threshold being removed. Essentially, it is used to filter out overlapping bounding boxes predicted for the same object location. As it can be observed, this combination achieved a low mAP of 0.11 at 0.5 IoU and an AR@10 of 0.15. However, observation of these metrics over the number of completed steps (Figure A.1) showed that they stabilized early on, at around 1000 steps, indicating that the model had converged, so an increased number of training steps would probably not lead to performance improvements.

Exp	LR	WU LR	Min Score	Max IoU	mAP@0.5IoU	AR@10
M1	0.00001	N.A	0	0.4	0.11	0.15
M2	0.00001 (const.)	N.A	0	0.4	0.13	0.15
M3	0.001 (const.)	N.A	0	0.6	0.03	0.13
M4	0.0001	0.00001	0.2	0.5	0.04	0.06

Table A.1: RetinaNet with MobileNetV1 backbone experiments performed on the first fold’s validation set. Experiment (Exp), Warm up (WU), Learning Rate (LR), Detections (Detect), Constant (const).

Furthermore, an inspection of the detected regions showed that the predicted bounding boxes were positioned around cells and nuclei, similarly to the ground-truth ones; yet, as evidenced by the low recall, it failed to detect many of the abnormal ones while having multiple false positives as well, explaining the low mAP.

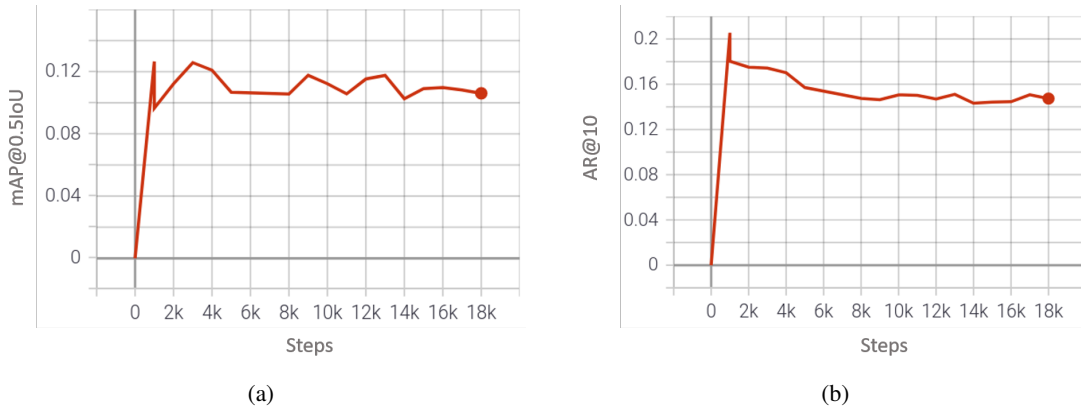


Figure A.1: RetinaNet with MobileNetV1 backbone first experiment (a) mAP@0.5IoU and (b) AR@10.

To assess if this early convergence was due to an excessively steep learning rate decline (which would not give the network enough time to properly learn the more intricate details of the abnormal regions), throughout the second experiment, the learning rate was left at the same initial value. However, this only led to a slight mAP increase of 2%. The following experiment still left the learning rate at a constant value, but it was increased to 0.001. In addition, the maximum IoU was also increased to 0.6 to check if this threshold was too strict. However, this was not verified as the observation of the predicted regions showed many overlapping bounding boxes around the same type of structures as in the other experiments (Figure A.2). When coupling this with the fact that most of these structures were not actually abnormal, it led to a significant decrease in mAP to a value of just 0.03 (and a slightly lower recall as well).

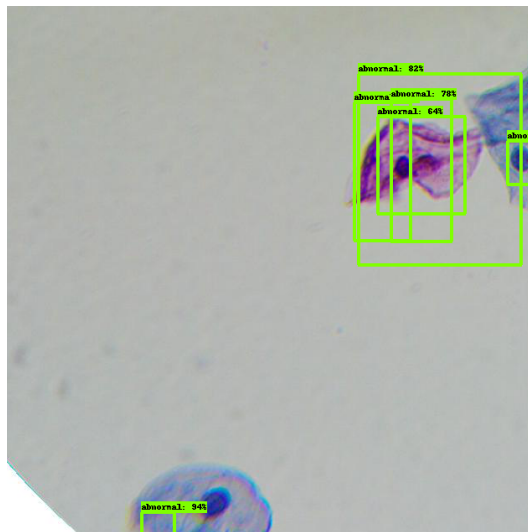


Figure A.2: Excessive overlapping bounding boxes when using an IoU threshold of 0.6.

On the other hand, we also noticed that some abnormal regions that were not identified in the previous experiments were now more likely to be detected. As such, in an effort to balance these effects, in the last experiment performed with this model, the maximum IoU was slightly decreased to 0.5, but a minimum score of 0.2 was set as well. In addition, the learning rate was decreased by an order of magnitude and once again set to have a cosine decay, as well as a warm up phase from a learning rate of 0.00001. While this solved the overlapping bounding boxes issue, the mAP still remained lower than in the first experiments, having a noticeable drop in AR from the previous one as well, probably derived from the imposed score threshold.

Overall, it seemed that higher IoU thresholds lead to worse results, even when other parameters were adjusted to mitigate the negative aspects derived from them, so this parameter was fixed at 0.4 for the remaining experiments. In addition to this, even though the constant learning rate lead to a slight increase in mAP between experiments M1 and M2, further experiments were conducted with a cosine decay in learning rate in order to possibly avoid overfitting, something all the more likely in datasets with a relatively small number of examples such as the one used.

It is worth noting though that this model's results are not directly comparable to the ones obtained by the RetinaNet with the ResNet50 backbone, not only due to the different image sizes but also because after these initial experiments, a new version of the dataset was made available featuring some adjustments to the annotations, aiming at increasing their accuracy and consistency.

Appendix B

Nuclei Detection

B.1 Alternative Approaches

In addition to the main strategy used to detect the nuclei in a region of interest, other approaches were also tested. For instance, as mentioned in Section 3.2.3.1, both histogram equalization and CLAHE were tested as contrast enhancers. However, as can be observed in Figure B.1b, a great deal of artificial noise is generated by these methods in otherwise empty portions of the image. Several attempts were made to solve this issue, such as applying yet another smoothing filter afterwards (which did not have much success) or saving a mask of these background-like areas (characterized by their lighter intensities) before the equalization operation and applying it on top of the contrast-enhanced image (Figure B.1c). The latter method had better results; however, it was still hard to determine a single correct threshold value to create the initial mask. This drawback was the main reason the histogram stretching method was chosen instead, as it did not produce any artificial noise.

As for the nuclei segmentation, strategies such as adaptive thresholding, single and multi-level Otsu thresholding were also examined. These methods were reasonably successful in many scenarios, but in the present case, they would result in many artefacts being caught in the threshold since, when it comes to intensity values alone, it is difficult to differentiate them from actual nuclei.

Apart from these methods, the K-Means clustering algorithm was also tested. In these experiments, the nuclei would correspond to the regions in the cluster with the lowest intensity. It obtained very good results depending only on the number of clusters selected (K). Usually, this value was between 2, in simple images where the nuclei were the only objects present and contrasted heavily against the background, and 6, where there were many artefacts or agglomerates of nuclei / cellular components. However, it was not possible to find a sufficiently robust method to automatically determine this parameter. Furthermore, this approach also had the added problem of failing when there were nuclei of significantly different intensities in the same image since they

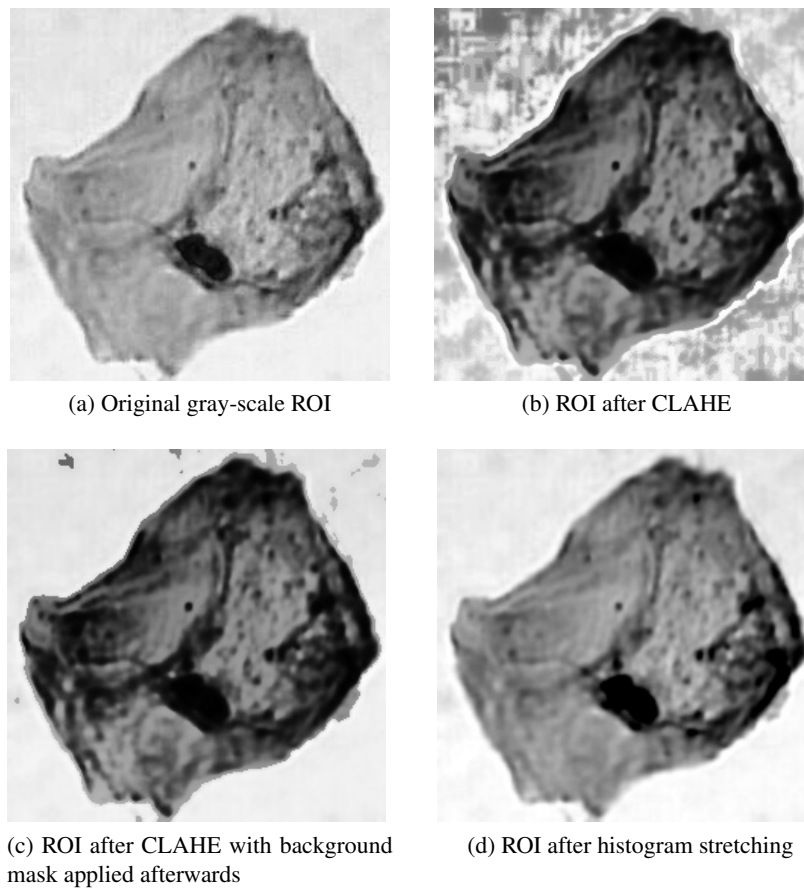


Figure B.1: Contrast adjustment techniques.

would be assigned to different clusters, with only the one having the lowest intensity being considered. Due to this, the iterative thresholding algorithm described in Section 3.2.3.2 was chosen instead as it consistently produced superior results.

Appendix C

Classification

C.1 Feature Importance

Even though the random forest ensembles achieved a worse performance than the SVMs, there are still some insights that can be extracted from them regarding the relevance of each feature. This relevance is computed as the normalized total reduction brought by that feature [49]. As it can be observed in Figure C.1, the top five most relevant features for each classifier are related to the colour of the image. In fact, the two most relevant features for both are related to the chroma value (or saturation) of the detected nuclei, further supporting their importance for the final decision. For example, when looking at the box plot of the mean distribution of the nuclei chroma channel energy (Figure C.2), it is indeed possible to see that the classes in each group (low and high grade) concentrate about 50% of their values in almost non-overlapping ranges. For instance, the LSIL and HSIL-SCC seem to have blander (less saturated) nuclei than their respective group counterparts. However, when looking at the distributions of the top geometrical (Figure C.4) and texture features (Figure C.3), there is no such clear distinction between the classes. This does not mean that these features are not relevant though, just that the classes are not so easily separable through them alone as with the colour-related features. But even the latter do not achieve this separation for all examples, therefore needing to be complemented by the other ones.

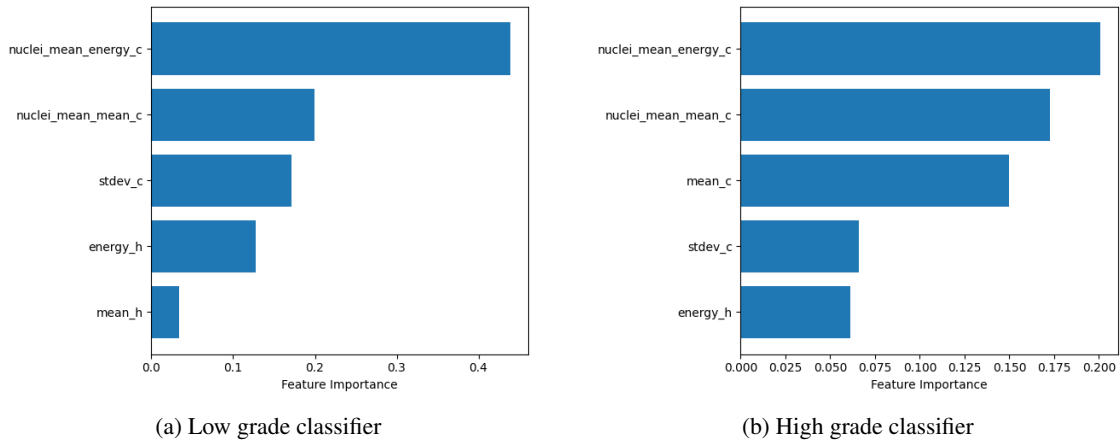


Figure C.1: Random forests' top five most important features.

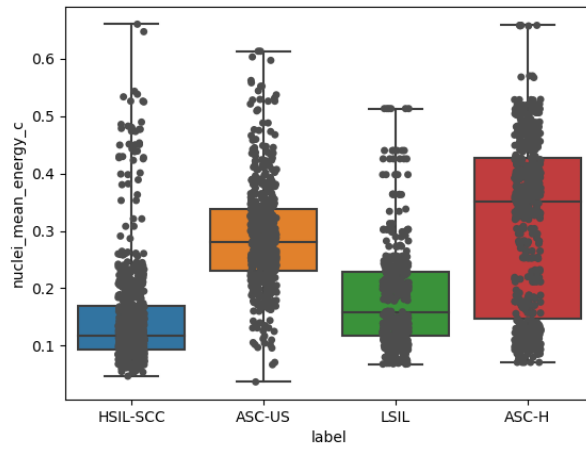
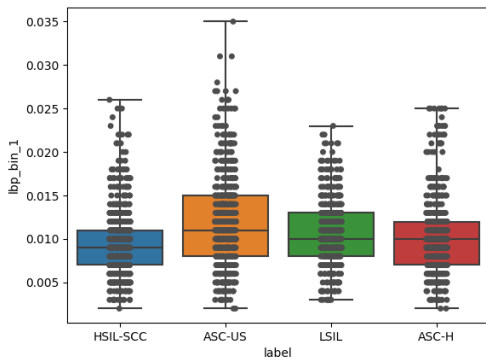
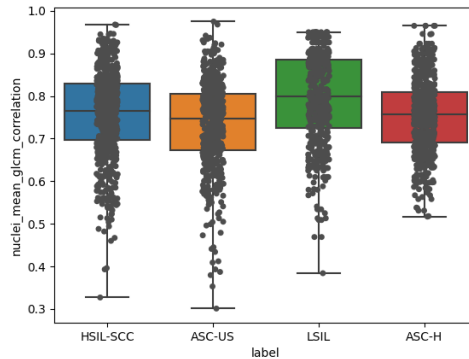


Figure C.2: Nuclei mean energy (chroma channel) box plot (overall train set).



(a) LBP bin #1 (low grade classifier)



(b) Nuclei mean GLCM correlation (high grade classifier)

Figure C.3: Random forests' most important texture features box plots (overall train set).

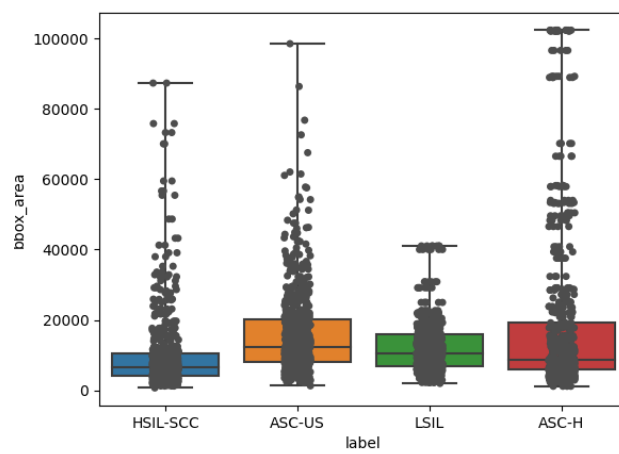


Figure C.4: Bounding box area box plot (overall train set).

References

- [1] World Health Organization. Cervical Cancer. https://www.who.int/health-topics/cervical-cancer#tab=tab_1. Accessed: 2021-06-29.
- [2] CLARE. Fraunhofer Portugal. https://www.aicos.fraunhofer.pt/en/our_work/projects/clare.html. Accessed: 2020-12-05.
- [3] Cancer.Net. Cervical Cancer: Statistics. <https://www.cancer.net/cancer-types/cervical-cancer/statistics>. Accessed: 2020-12-09.
- [4] Fox Chase Cancer Center. Precancerous Cervical Changes: They're Common and Treatable. <https://www.foxchase.org/blog/precancerous-cervical-changes-they%E2%80%99re-common-and-treatable>. Accessed: 2020-12-09.
- [5] Patricia A Shaw. The history of cervical screening i: The pap. test, 2000.
- [6] Ewert Bengtsson and Patrik Malm. Screening for cervical cancer using automated analysis of pap-smears. *Computational and Mathematical Methods in Medicine*, 2014, 2014. General Overview.
- [7] 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.
- [8] Daniela M Ushizima, Andrea G C Bianchi, and Claudia M Carneiro. Segmentation of sub-cellular compartments combining superpixel representation with voronoi diagrams, 2014.
- [9] Zhi Lu, Gustavo Carneiro, Andrew P. Bradley, Daniela Ushizima, Masoud S. Nosrati, Andrea G.C. Bianchi, Claudia M. Carneiro, and Ghassan Hamarneh. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE Journal of Biomedical and Health Informatics*, 21:441–450, 3 2017.
- [10] Hady Ahmady Phoulady, Dmitry B Goldgof, Lawrence O Hall, and Peter R Mouton. An approach for overlapping cell segmentation in multi-layer cervical cell volumes, 2015.
- [11] Mohammed Kuko and Mohammad Pourhomayoun. Single and Clustered Cervical Cell Classification with Ensemble and Deep Learning Methods. *Springer Nature*, 2020.
- [12] Jun Du, Xueyu Li, and Qinghua Li. Detection and Classification of Cervical Exfoliated Cells Based on Faster R-CNN. In *2019 IEEE 11th International Conference on Advanced Infocomm Technology, ICAIT 2019*, pages 52–57. Institute of Electrical and Electronics Engineers Inc., oct 2019.

- [13] Chao Li, Xinggong Wang, Wenyu Liu, and Longin Jan Latecki. DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical Image Analysis*, 45:121–133, apr 2018.
- [14] Tahir Mahmood, Muhammad Arsalan, Muhammad Owais, Min Beom Lee, and Kang Ryoung Park. Artificial Intelligence-Based Mitosis Detection in Breast Cancer Histopathology Images Using Faster R-CNN and Deep CNNs. *Journal of Clinical Medicine*, 2020.
- [15] Petru Soviany and Radu Tudor Ionescu. Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction, 2018.
- [16] Ming Zhou, Lichi Zhang, Xiaping Du, Xi Ouyang, Xin Zhang, Qijia Shen, and Qian Wang. Hierarchical and Robust Pathology Image Reading for High-Throughput Cervical Abnormality Screening. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12436 LNCS, pages 414–422. Springer Science and Business Media Deutschland GmbH, 2020.
- [17] Christian Marzahl, Marc Aubreville, Christof A. Bertram, Jason Stayt, Anne Katherine Jasensky, Florian Bartenschlager, Marco Fragosso-Garcia, Ann K. Barton, Svenja Elsemann, Samir Jabari, Jens Krauth, Prathmesh Madhu, Jörn Voigt, Jenny Hill, Robert Klopffleisch, and Andreas Maier. Deep Learning-Based Quantification of Pulmonary Hemosiderophages in Cytology Slides. *Scientific Reports*, 10(1), dec 2020.
- [18] Philipp Gabel, Ozcan Ozkan, Martina Crysandt, Reinhild Herwartz, Melanie Baumann, Barbara M. Klinkhammer, Peter Boor, Tim H. Brummendorf, and Dorit Merhof. Circular Anchors for the Detection of Hematopoietic Cells Using Retinanet. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2020-April, pages 249–253. IEEE Computer Society, apr 2020.
- [19] Jingru Yi, Pengxiang Wu, Menglin Jiang, Qiaoying Huang, Daniel J. Hoepfner, and Dimitris N. Metaxas. Attentive neural cell instance segmentation. *Medical Image Analysis*, 55:228–240, 7 2019.
- [20] 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, nov 2017.
- [21] Melanie Kwon, Mohammed Kuko, Vanessa Martin, Tae Hun Kim, Sue Ellen Martin, and Mohammad Pourhomayoun. Multi-label classification of single and clustered cervical cells using deep convolutional networks, 2018.
- [22] Ahmed Ghoneim, Ghulam Muhammad, and M. Shamim Hossain. Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems*, 102:643–649, jan 2020.
- [23] Nor Ashidi Mat-Isa, Mohd Yusoff Mashor, and Nor Hayati Othman. An automated cervical pre-cancerous diagnostic system. *Artificial Intelligence in Medicine*, 42(1):1–11, jan 2008.
- [24] Nor Ashidi Mat Isa. Automated Edge Detection Technique for Pap Smear Images Using Moving K-Means Clustering and Modified Seed Based Region Growing Algorithm, 2005.
- [25] Jie Su, Xuan Xu, Yongjun He, and Jinming Song. Automatic Detection of Cervical Cancer Cells by a Two-Level Cascade Classification System. *Analytical Cellular Pathology*, 2016, 2016.

- [26] Meenakshi Sharma, Sanjay Kumar Singh, Prateek Agrawal, and Vishu Madaan. Classification of Clinical Dataset of Cervical Cancer using KNN. *Indian Journal of Science and Technology*, 9(28), jul 2016.
- [27] Sophea Prum, Dini Oktarina Dwi Handayani, and Patrice Boursier. Abnormal Cervical Cell Detection using HOG Descriptor and SVM Classifier. In *Proceedings - 2018 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018*. Institute of Electrical and Electronics Engineers Inc., oct 2018.
- [28] 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, oct 2020.
- [29] Aditya Khamparia, Deepak Gupta, Victor Hugo C. de Albuquerque, Arun Kumar Sangahiah, and Rutvij H. Jhaveri. Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. *Journal of Supercomputing*, 76(11):8590–8608, nov 2020.
- [30] Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. *University at Buffalo School of Medicine and Biomedical Sciences, USA*, 2014.
- [31] Abid Sarwar, Vinod Sharma, and Rajeev Gupta. Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis. *Personalized Medicine Universe*, 4:54–62, 2015.
- [32] Srishti Gautam, Harinarayan K K, Nirmal Jith, Anil K Sao, Arnav Bhavsar, and Adarsh Natarajan. Considerations for a PAP Smear Image Analysis System with CNN Features, 2014.
- [33] Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu. Object Detection with Deep Learning: A Review, 2017.
- [34] Luís Rosado, Paulo Silva, José Faria, João Oliveira, Maria Vasconcelos, Dirk Elias, José Manuel Correia da Costa, and Jaime Cardoso. *μSmartScope: Towards a Fully Automated 3D-Printed Smartphone Microscope with Motorized Stage*, pages 19–44. 07 2018.
- [35] George G. Birdsong and Diane Davis Davey. The Bethesda System for Reporting Cervical Cytology. In *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*, pages 1–28. Springer International Publishing, jan 2015.
- [36] TensorFlow. Object Detection API. https://github.com/tensorflow/models/tree/master/research/object_detection. Accessed: 2021-05-16.
- [37] PyTorch. <https://pytorch.org/>. Accessed: 2021-05-16.
- [38] TensorFlow. Object Detection API Model Zoo. https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md. Accessed: 2021-05-16.
- [39] COCO - Common Objects in Context. <https://cocodataset.org/>. Accessed: 2021-05-16.

- [40] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Focal Loss for Dense Object Detection*, volume 2017-Octob, pages 2999–3007. Institute of Electrical and Electronics Engineers Inc., 12 2017.
- [41] 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, 7 2017.
- [42] K. Wojciechowski, B. Smolka, H. Palus, R.S. Kozera, W. Skarbek, and L. Noakes, editors. *Computer Vision and Graphics*, volume 32. Springer Netherlands, 2006.
- [43] Gravure Ink. Tolerancing color in presswork - CIE L*a*b* and DeltaE. <http://toyoinkthailand.blogspot.com/p/gravure-ink.html>. Accessed: 2021-06-12.
- [44] Scikit-Image. GLCM Texture Features. https://scikit-image.org/docs/dev/auto_examples/features_detection/plot_glcm.html. Accessed: 2021-05-30.
- [45] Golshah Naghdy, Montserrat Ros, Catherine Todd, E Norachmawati, and Montse Ross. Classification cervical cancer using histology images. 2010.
- [46] COCO. Detection Evaluation. <https://cocodataset.org/#detection-eval>. Accessed: 2021-06-11.
- [47] R. Padilla, S. L. Netto, and E. A. B. da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242, 2020.
- [48] Scikit-Image. Metrics and scoring: quantifying the quality of predictions. https://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-f-measure-metrics. Accessed: 2021-02-24.
- [49] Scikit-Learn. Random forest feature importance. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.feature_importances_. Accessed: 2021-06-14.