FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Cervical Cancer Screening: Deep Learning Tools For Automatic Diagnosis Support

Francisca Marques de Almeida Morgado

Mestrado Integrado em Bioengenharia

Supervisor: Prof. Jaime Cardoso

Co-Supervisor: Dr. Maria Vasconcelos

October 1, 2019

Cervical Cancer Screening: Deep Learning Tools For Automatic Diagnosis Support

Francisca Marques de Almeida Morgado

Mestrado Integrado em Bioengenharia

October 1, 2019

Resumo

O cancro do cólo do útero é o quarto cancro mais frequente em mulheres, ocupando o oitavo lugar, a nível global. Apesar dos avanços médicos e científicos, atualmente não existe um tratamento totalmente eficaz para esta doença, especialmente quando diagnosticada em estadios mais avançados. Desta forma, é atribuída elevada importância a programas de prevenção e rastreio, no combate a este cancro.

O processo de rastreio do cancro do cólo do útero é composto por várias fases respeitando a seguinte ordem: teste do vírus do papiloma humano (VPH), exame citológico (ou teste de Papanicolau), exame de colposcopia e biópsia. Várias ferramentas têm sido desenvolvidas para apoiar este processo, tornando-o mais eficaz, prático e acessível. Neste contexto, surgiu o projecto CLARE que visa desenvolver um sistema inovador de suporte à decisão para o rastreio do cancro do cólo do útero. A presente dissertação integra o projecto CLARE e foca-se no desenvolvimento de um sistema de apoio à decisão destinado ao exame de colposcopia, usando ferramentas de *Deep Learning*. A colposcopia é, na prática, uma endoscopia vaginal realizada por um ginecologista que avalia o grau de risco relativo ao desenvolvimento de cancro, sendo que os casos positivos seguem para biópsia. Esta dissertação visa criar uma ferramenta que realiza o mesmo tipo de avaliação baseando-se numa única imagem do cólo do útero e respectivos dados clínicos da paciente.

Para alcançar o modelo que melhor desempenha a tarefa suprarreferida, foram adoptados vários métodos, sendo que o primeiro baseou-se na exploração de algumas opções de segmentação, com o objectivo de analisar a relevância da segmentação da zona de interesse. Posteriormente, foram desenvolvidas e testadas várias abordagens que integravam técnicas de *Transfer Learning* e *Multitask Learning*. Usando os melhores modelos aí obtidos, foi testado um método de regularização com características canônicas, onde o treino dos modelos é orientado de forma a aprenderem as características que são usualmente extraídas de imagens do cólo do útero em abordagens clássicas de *Machine Learning*. Mais tarde, foram desenvolvidos métodos para integrar os dados clínicos nos modelos criados, transformando-os em modelos multimodais. Finalmente, foram implementadas metodologias para superar a limitação do desiquilíbrio entre classes, incluindo algoritmos como SMOTE, *Cluster Centroids*, SMOTEENN, *over-sampling* manual e algorítmos de *ranking*.

Para seleccionar os melhores modelos, foram consideradas quatro categorias, tendo em conta a inclusão ou não dos dados clínicos e a métrica de selecção. A métrica de selecção varia com o objectivo do modelo. Quando se pretende o modelo com melhor desempenho a nível geral, as métricas mais relevantes são *accuracy* e AUC, no entanto, em casos de rastreio, existe interesse em encontrar um modelo que minimize o número de falsos negativos, sendo avaliado pelas métricas *sensitivity* e NPV. Definidas as categorias, o modelo multimodal com o melhor desempenho global atingiu uma AUC de 91.57% e *accuracy* de 88.37% enquanto o modelo unimodal atingiu uma AUC de 73.86% e *accuracy* de 84.86%. Relativamente aos melhores modelos para rastreio, o multimodal atingiu uma *sensitivity* de 95.42% e NPC de 98.62% enquanto o unimodal ficou pelos 49.85% de *sensitivity* e 89.20% de NPV.

Abstract

Cervical cancer is the fourth most frequent cancer in women and the eighth most commonly occurring cancer overall. Despite medical and scientific advances, there is no total effective treatment for this disease, especially when diagnosed in an advanced state. For this reason, prevention and screening programs play a very important role in the fight against cervical cancer.

Cervical cancer screening follows a standard workflow that includes the following steps: HPV test, cytology test or Pap smear, colposcopy, and biopsy. Several tools have been developed to support this workflow, making it more efficient, more practical and more affordable. In this context, the CLARE project emerged with the aim of creating a novel decision support system designed for cervical cancer screening. This dissertation integrates the CLARE project and focuses on developing decision support tools for colposcopy examination, making use of Deep Learning techniques. Colposcopy is a medical exam performed by gynecologists that consists of performing a vaginal endoscopy to predict the risk of cervical cancer. In this dissertation, the decision support system performs the same prediction based on a single cervix image and patient's clinical data.

To find the classification model that better fits the mentioned task, several methods were applied. The first step explored some segmentation options to analyze the relevance of extracting the region of interest in this problem. After concluding that segmentation adds no value for this case, several approaches were tested integrating Transfer Learning and Multitask Learning techniques. The best models were transformed to test the effect of canonical feature regularization, where CNN's training is oriented so the neural networks learn to extract features that are usually extracted from images, in classical Machine Learning approaches. Later, clinical data was introduced in the models, turning them into multimodal algorithms. Finally, to overcome the class imbalance problem, several approaches were implemented, such as SMOTE, Cluster Centroids, SMOTEENN, over-sampling during data augmentation, and ranking algorithms.

In the end, instead of selecting the best model, four models were selected, considering two variables: availability of clinical data, and the preferred metric. When clinical data is available, the best model is a multimodal algorithm, otherwise, is selected a unimodal model. Considering the metric variable, to select the best overall model, AUC and accuracy are the preferred metrics. However, for screening problems, it is interesting to find a model that minimizes the number of false negatives, preferring sensitivity and NPV metrics. The best overall multimodal model achieved 91.57% of AUC and 88.37% of accuracy, the best overall unimodal achieved 73.86% of AUC and 84.86% of accuracy, the best screening multimodal model obtained a sensitivity of 95.42% and an NPV of 98.62%, and the best screening unimodal model achieved a sensitivity of 49.85% and an NPV of 89.20%.

Acknowledgements

As minhas primeiras palavras de agradecimento são dirigidas ao professor Jaime Cardoso por toda a ajuda e orientação ao longo destes meses. A sua disponibilidade e profissionalismo foram incansáveis assim como toda a paciência em lidar com o meu pessimismo. O meu obrigada também à Doutora Maria Vasconcelos pela sua disponibilidade e apoio, sempre que necessários. Ao Ricardo Cruz, agradeço as explicações, o acesso aos servidores que me permitiram realizar esta dissertação e toda a prontidão na resolução de problemas, mesmo à distância.

Agradeço ainda ao NCI/NIH por disponibilizarem a base de dados que permitiu a realização de todo este trabalho.

Esta dissertação foi também o culminar de uma aventura de cinco anos. Muitas foram as pessoas que me acompanharam nesta jornada e que contribuiram para que esta meia década seja sempre lembrada com carinho. A todas elas, deixo aqui o manifesto da minha gratidão por tudo o que me ensiram, pelos sorrisos vividos e pelos que continuam a surgir. Ao 23, a todos os organelos responsáveis pela produção de energia e respectivos complexos, aos duendes irlandeses, ao Manuel da Antónia e a todos os seres que apreciam e retribuem turrinhas, a minha mais sentida palavra de apreço.

A toda a minha família, em especial aos meus pais, mil agradecimnetos nunca seriam suficientes. Vocês são o meu maior bem e a origem de toda a minha força. Obrigada!

"Any sufficiently advanced technology is indistinguishable from magic."

Arthur C. Clarke

Contents

List of Figures

List of Tables

Abbreviations

API	Application Programming Interface
AUC	Area Under the Curve
AW	Acetowhite
CAD	Computer-Aided Diagnosis
CE	Columnar Epithelium
CIN	Cervical Intraepithetial Neoplasia
CNN	Convolutional Neural Network
FIGO	Federation of Gynecology and Obstetrics
FN	False Negative
FP	False Positive
GBDT	Gradient Boosting Decision Tree
GMM	Gaussian Mixture of Models
GMROSE	Gaussian Modulation of Rotation Structuring Element
GT	Graph-based Transduction
HIV	Human Immunodeficiency Virus
HLA	Human Eukocyte Antigen
HPV	Human Papilomavirus
HSI	Hue, Saturation, Lightness
HSV	Hue, Saturation, Value
kNN	k-Nearest Neighbours
LR	Logistic Regression
MLP	Multilayer Perceptron
MTL	Multi-task Learning
NCI	National Cancer Institute
NIH	National Institute of Health
NPV	Negative Predictive Value
PHOG	Pyramid Histogram of Oriented Gradients
PLAB	Pyramid histogram in L*A*B* color space
PLBP	Pyramid histogram of Local Binary Patterns
R-CNN	Region Convolutional Neural Network
RF	Random Forest
RGB	Red, Green, Blue
ROSE	Rotation Structuring Element
SCJ	Squamocolumnar Junction
SLIC	Simple Linear Iterative Clustering
SVM	Support Vector Machine
TN	True Negative
TNF	Tumor Necrosis Factor

TP	True Positive
TZ	Transformation Zone
VGG	Visual Geometry Group

Chapter 1

Introduction

1.1 Motivation

Cervical cancer is one of the four most frequent cancer in women. For 2018, the estimations indicated 570,000 [1] new cases worldwide and 311,365 deaths [2], representing 6.6% of all female cancers. About 90% of the deaths caused by this cancer will have occurred in low and middle-income countries, where cervical cancer is in the top two of the most common cancers in women [3]. In Portugal, where the Human Papillomavirus (HPV) vaccine is integrated in the National Vaccination Program, there are about one thousand new cases every year [4], killing more than 200 women annually [5].

Cervical cancer, as well as most cancers, has better chances to be treated in the early stages but the absence of signs and symptoms in this phase hampers an early diagnosis of this disease. Therefore, prevention programs play an important role to reduce cancer incidence and mortality.

Nowadays, there are vaccination and screening programs for cervical cancer prevention. The vaccinations usually prevent Human Papillomavirus, once it can stimulate cervical intraepithelial neoplasia (CIN) which in turn is the cervical cancer promoter. Screening programs may include HPV tests as the first triage approach, followed by cytology test (or Pap smear) and colposcopy examination.

Cytological screening requires manual smearing and staining performed by an expert, who can choose between conventional or liquid-based cytology. The second one overcomes some problems such as dense regions caused by uneven distribution of the cells and presence of blood or other artifacts; however, the cost is about 5 to 10 times higher than conventional cytology [6]. After cytological examination, patients with positive or inconclusive results must undergo colposcopy examination.

Colposcopy is a procedure that consists of capturing images of the cervix (cervigrams) following a 4-step procedure. In each step, a filter (i.e. green light) or a solution (i.e. acetic acid) is applied on the cervix to enhance one of the regions of interest. Nowadays, there are low-cost and portable devices to perform colposcopy examinations which promotes greater access to cervical

cancer screening in low-income countries. Even so, this procedure should be performed by experienced professionals who can provide an accurate diagnosis, which is a limitation for low-income countries due to their restricted resources, including medical professionals.

1.2 Objectives

To overcome the limitation related to medical resources, some methods have been proposed to help medical staff during a colposcopy examination, avoiding the need for a specialist to perform such an examination.

Some works provided tools to enhance colposcopy image, segment regions of interest, and detect some abnormalities. These tools help professionals to make a decision about the diagnosis, however, the professional should have enough knowledge and training to perform an accurate diagnosis.

The main goal of this dissertation is to develop an automatic system to support cervical cancer screening, providing a prediction about cervical cancer risk of each patient based on her cervigram and clinical data, when available. When performed by specialist, the colposcopy examination has a sensitivity about 61.1% and a specificity of 53.4% [7], therefore, a good automatic systems should surpass this performance, otherwise, it is not useful to support specialists' decision.

1.3 Expected contributions

This work is part of the CLARE (computer-aided cervical cancer screening) project, which emerged from a partnership between INESC TEC (Institute for Systems and Computer Engineering, Technology and Science) and Fraunhofer Portugal. The goal of this project is to provide automatic tools for cervical cancer screening based on cytology and colposcopy images.

It is expected, from this work, to develop an image-based Computer-aided Diagnosis (CAD) system to assist medical professionals on cervical cancer diagnosis, using only colposcopy images and patient's data. The system would integrate deep learning tools and multimodal methodologies, and aims to exceed the results presented by other authors in previous works (more detailed information in chapter 3).

1.4 Document structure

This document is divided into 6 chapters. The first one is the present introduction followed by two chapters with contextual information. Chapter 2 contains an introduction of cervical cancer including a brief description of female reproductive system anatomy, cervix cytology, signs and symptoms of cervical cancer, prevention and treatment, pathophysiology, and screening methods. In chapter 3, there is a review of the state of art, including a background on Machine Leaning and Deep Learning, and an overview on published work related to CAD systems for cytology images. Previous studies provide methodologies for different tasks such as image quality enhancement and

assessment, segmentation of regions of interest, image registration, detection of abnormal regions and patterns, among others, being the most relevant task the classification of cervical cancer risk.

Chapter 4 presents a description of the methodologies applied to find the best regularized CNN to perform cervical cancer screening. The results from these methods are also presented in chapter 4 as well as the respective discussion. In chapter 5, it is introduced the concept of multimodality to include clinical data into the models and several methods for imbalanced learning are explored as well. At the end of the chapter, there is the final results and respective discussion. Finally, in chapter 6, there is a conclusion and some remarks about future work.

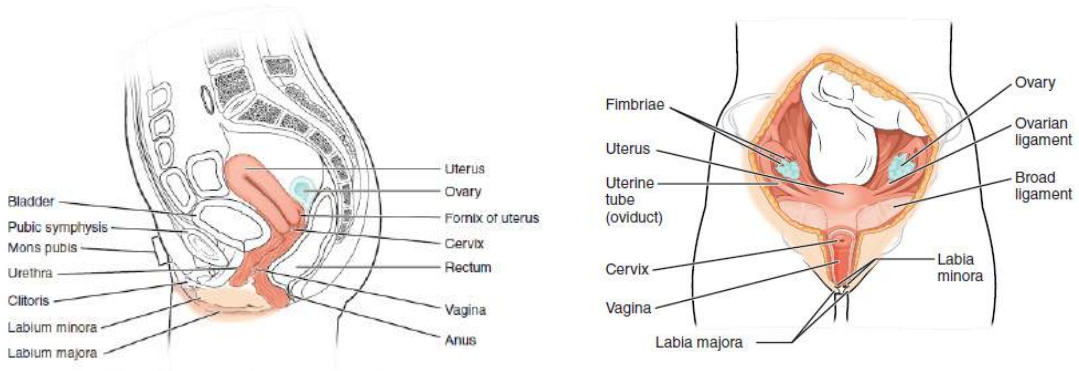
Chapter 2

Cervical Cancer

Cervical cancer is the fourth most common cancer in women and the eighth most common overall. According to statistics from the International Agency for Research on Cancer, there was 569 847 new cases and 311 365 deaths caused by Cervical Cancer, in 2018, worldwide [2]. To learn more about this disease, in this chapter is presented an overview about cervical cancer concerning uterine cervix anatomy and cytology, signs and symptoms, treatment and prevention, pathophysiology, and screening methods.

2.1 Overview of Female Reproductive System Anatomy

Female reproductive system is located inside the pelvic cavity (figure 2.1) and its organs can be split in two main categories: internal genitalia and external genitalia [8]. Internal genitalia includes the organs located inside the pelvis such as the vagina, uterus, cervix, fallopian tubes (uterine tubes or oviducts), and ovaries, while the external genitalia, that is located on the outer part of the pelvis, includes the perineum, mons pubis, clitoris, labium minora, labium majora, urethral meatus, and vestibule [8].



(a) Lateral view.

(b) Anterior view.

Figure 2.1: Lateral and anterior views of human female reproductive system. Adapted from [9].

The ovaries are paired organs responsible for maturing and releasing the female eggs during menstrual cycle [8]. This small and almond-shape organs connect to the uterus via fallopian tubes which are located bilaterally at the upper portion of the uterine cavity. Fallopian tubes, uterine tubes or oviducts have three parts: infundibulum, closest to the ovary and responsible for catching the egg, ampulla, where the fertilization occurs, and isthmus, closest to the uterus [8].

The uterus is an inverted pear-shaped organ located within the pelvis in a position anterior to the rectum and posterior to the bladder [8]. It has two parts: the body of the uterus (or corpus uteri) and the cervix. The body is globe-shaped and it can be anteverted (tilted forward) or retroverted (tilted back). When the uterus is anteverted, the cervix is bent forward and the cervical external orifice is directed to the posterior vaginal wall. When the uterus is retroverted, the cervix enters the vaginal through an anterior approach, being more difficult to detect during a colposcopic examination [10]. The body of the uterus is responsible for holding the fetus during pregnancy and its walls have the function to contract to help fetus to be expelled [8].

The cervix is the portion of the uterus that separates the body of the uterus from the vagina. It has a cylindrical shape working as a canal that allows the entry of semen into the uterine cavity. The cervix has two orifices, the internal os that opens into the endometrial cavity and the external os that opens into the vagina [8]. The internal os is responsible for dilating during labor, allowing fetus delivery. The cervix is divided into ectocervix, which is the portion of the cervix projected into the vagina being visible during a colposcopic examination, and the endocervix that is the canal between internal os and external os.

The vagina is the canal that extends from the cervix to the vulva, connecting internal genitalia and external genitalia. It is located between the rectum and the urinary bladder and lies at an angle of 90° to uterus [8]. Details from external genitalia were not considered in this brief explanation, once it is not the study object.

2.2 Cervix Citology

As explained in the previous section, cervix is the lower portion of the uterus and can be divided into endocervix and ectocervix. Besides the location, these two portions also differ in lining. Endocervix is lined with columnar epithelium (or also referred as glandular epithelium, figure 2.2), while ectocervix is lined with squamous epithelium (figure 2.3) [10].

Columnar and squamous epitheliums meet at the squamocolumnar junction (SCJ), as shown in figure 2.4. Before puberty, SCJ is located very close to the external os (original SCJ) but after the first menstruation, and during the reproductive period, SCJ moves to the ectocervix (new SCJ), staying away from the external os, due to the elongation of the endocervical canal [10]. This transformation leads to the exposure of columnar epithelium to the acidic vaginal environment which cause destruction of this epithelium. To repair the tissue, that portion of columnar epithelium is replaced by newly formed metaplastic squamous epithelium [10]. The epithelium between the original SCJ and the new SCJ is called as the transformation zone (TZ) which is also dynamic,

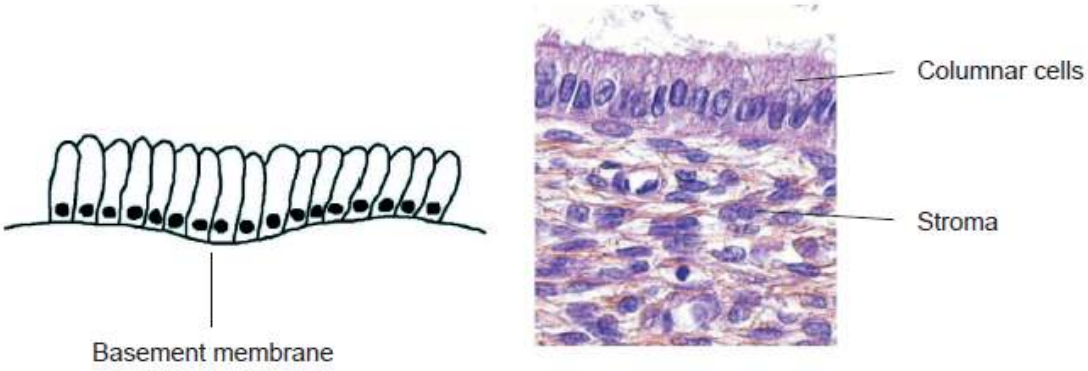


Figure 2.2: Columnar epithelium of the endocervix. Adapted from [10].

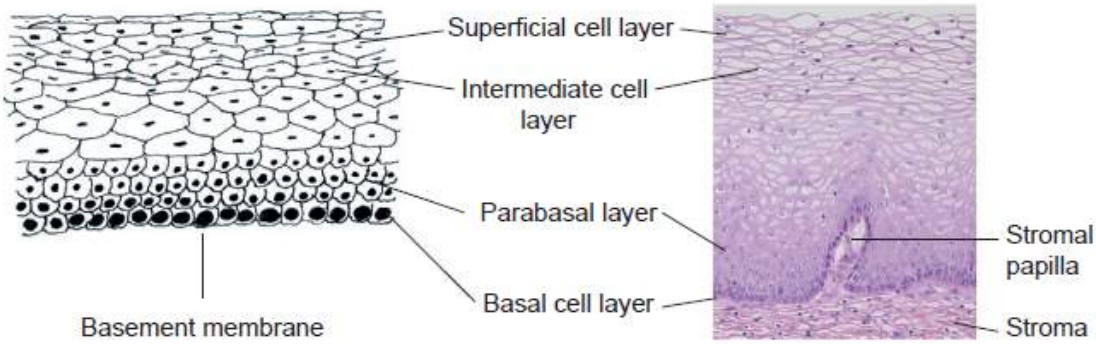


Figure 2.3: Squamous epithelium of the ectocervix. Adapted from [10].

being located at the ectocervix during the reproductive period and moving to an endocervical position after menopause [10].

The identification of the transformation zone is very important during colposcopic examination once squamous cervical cancer, which represents the majority of cervical cancers, begins in the TZ. Glandular cervical cancer also can originate in the TZ or in the columnar epithelium above this zone. Premenopausal women usually present a transformation zone located on the ectocervix. After menopause, the cervix shrinks and, consequently, the TZ moves into the cervical canal, embarrassing TZ visibility. The different transformation zones locations allows to categorize cervix into 3 types:

- type 1 - TZ completely ectocervical and fully visible;
- type 2 - TZ with endocervical component and fully visible;
- type 3 - TZ with endocervical component and not fully visible.

Different types of cervix or TZ lead to different treatment approaches, therefore, cervix type identification prevents ineffectual treatments.

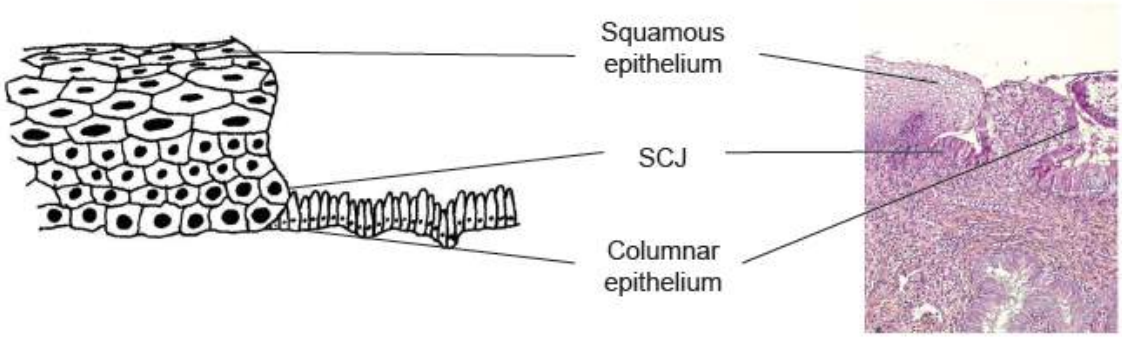


Figure 2.4: Squamocolumnar junction (SCJ). Adapted from [10].

2.3 Signs and Symptoms

Signs and symptoms manifested by the patients depend on the stage of the cancer, nevertheless patients with cervical cancer usually are asymptomatic until an advanced stage. The first noticeable symptom is abnormal vaginal bleeding that may occurs during or after intercourse, between menstrual periods, or after menopause [11]. Other manifested symptoms may include vaginal discomfort, malodorous discharge, dysuria (painful or difficult urination), and pain in lower back or pelvis [11] [12]. In advanced stages, when the cancer spreads into surrounding tissues, the symptoms also include constipation, urinary incontinence, blood in the urine, and swelling of the legs [12].

Cervical cancer is a silent disease, therefore, prevention and screening programs are very important, specially in middle and low income countries, where the levels of mortality caused by cervical cancer are very high [2].

2.4 Prevention and Treatment

Nowadays it is possible to reduce the risk of cervical cancer through vaccination. The 9-valent HPV vaccine decreases the risk of some cancers and precancerous lesions both in men and women, covering nine subtypes of HPV (6, 11, 16, 18, 31, 33, 45, 52, and 58) and preventing cervical squamous intraepithelial lesions [11]. Gardasil is another vaccine used to prevent cervical cancer protecting against 4 subtypes of HPV, including HPV 16 and HPV 18 which are the most carcinogenic types [13]. These vaccines can significantly reduce the risk of cervical cancer, however the immunization can not be guaranteed and every women should attend cervical screening tests, with and without vaccination [13].

The treatment for cervical cancer depends on the stage of the lesion. For early stages, the most common procedure is the removal of the cervix and a part or the entire uterus, radiotherapy, or both. For advanced stages, surgery, radiotherapy and chemotherapy are the most common procedures. The removal technique to be applied also depends on the stage [11]:

- **Stage 0:** It is a carcinoma *in situ* and it can be removed with cryosurgery, laser ablation, loop excision, or with normal surgery;
- **Stage IA1:** For this stage, the common procedures are total hysterectomy, radical hysterectomy, and conization;
- **Stage IA2, IB, or IIA:** External beam radiation with brachytherapy and radical hysterectomy with bilateral pelvic lymphadenectomy, or a combination of both is usually applied for stages IB and IIA. In stage IB1, when the lesion is smaller than 2 cm it can be removed by radical vaginal trachelectomy with pelvic lymph node dissection, being also appropriate for women with lesions of stage IA2 who want to preserve their fertility:
- **Stage IIB, III, or IVA:** The most common procedure is cisplatin-based chemotherapy with radiation;
- **IVB and more advanced stages:** Palliative care, radiation therapy for bleeding and pain control, and systemic chemotherapy.

2.5 Pathophysiology

Cervical cancer only occurs after a woman is infected by human papilomavirus (HPV). When infected, cytology reports may show squamous intraepithelial lesion, however, 90% of HPV infections disappear completely in a few years leaving no sequels [11]. About 5% of HPV infections lead to the development of cervical intraepithelial neoplasia (CIN) of grade 2 or 3. CIN grade 3 is a cancer precursor that leads to invasive cervical cancer with a risk of progression of 20% within 5 years and 40% within 30 years [11].

There are other risk factors that should be considered, such as genetic susceptibility, poor immunity (HIV infection and poor nutritional status), risk behaviors (for example, smoking or imbalanced diet), high number of sexual partners, first intercourse in early age, and lack of access to routine screening [11].

Genetic susceptibility is related with the activation of certain genes. Tumor necrosis factor (TNF) is responsible to initiating cell apoptosis and it can be expressed in different genes [11]. TNFa-8, TNFa-572, TNFa-857, TNFa-863, and TNF G-308A genes increase the incidence of cervical cancer. Also Tp53, a gene involved in apoptosis and gene repair, has been associated with higher incidence of cervical cancer, being responsible for increasing progression rate from HPV infection to cervical cancer. Other genes such as human eukocyte antigen (HLA), hemokine receptor-2 (CCR2), and the Fas gene are also related with genetic susceptibility to cervical cancer. On the other hand, CASP8 gene (also referred as FLICE or MCH5 gene) seems to decrease the risk of this cancer [11].

Concerning HPV subtypes, the most carcinogenic belongs to alpha group 1 (table 2.1), with more than 90% of cervical cancer caused by subtypes 16, 18, 31, 33, 35, 45, 52, and 58. In opposition, subtypes of the alpha group 3 (subtypes 6 and 11) were never been associated to cases

Table 2.1: Human Papillomavirus subtypes associated with cervical cancer. Adapted from [11]

HPV Alpha Group	Subtypes	Evidence for Cervical Cancer
1	16	Most carcinogenic HPV type, known to cause cancer at several sites
	18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59	Sufficient evidence
2A	68	Limited evidence in humans and strong mechanistic evidence
2B	26, 53, 66, 67, 70, 73, 82	Limited evidence in humans
	30, 34, 69, 85, 97	Classified by phylogenetic analogy to HPV types with sufficient or limited evidence in humans
3	6, 11	Inadequate epidemiological evidence and absence of carcinogenic potential in mechanistic studies

of invasive cancer. High-risk HPV integrates the human genome in genes E6 and E7 [11]. This binding leads to resistance to apoptosis, affecting cell growth control which promotes the growing of cells with damaged DNA that can result in malignancy.

The relation between HIV infection and cervical cancer pathogenesis is not completely explained by science, however, it is known that HIV infection suppress immune response, facilitating the infection by HPV which cause more damage than usual. Therefore, cervical cancer is five times more frequent in seropositive women than in seronegative women [11].

2.6 Screening Methods

The screening procedures for cervical cancer follow a set of guidelines that differ according to the country where the screening is performed and patient's age. International Federation of Gynecology and Obstetrics (FIGO) guidelines considered as screening methods for cervical cancer the Papanicolaou (Pap) test, colposcopy, biopsy, conization of cervix, cystoscopy, proctosigmoidoscopy and chest x-ray. Nevertheless, the most common procedures are Papanicolaou test and colposcopic examination [11]. In Portugal, the screening process starts with a HPV test. When the outcome is positive but not for subtypes 16 and 18, the guidelines demand cytology examination (or Pap test). For subtypes 16 and 18, and for positive cytology exams, the next step is colposcopy examination. Nevertheless, if the outcome from colposcopy is not clear, a biopsy and consecutive histological examination are performed. Even when the results are negative, patients with suspicious symptoms or inconclusive outcomes should be followed up, repeating the screening test.

2.6.1 Papanicolaou Test

Papanicolaou test, Pap smear or cervical cytology, is the oldest procedure used to screen cervical cancer and consists on exfoliating cells from the transformation zone to be observed on the microscope. The traditional technique involves transferring the cells directly to the microscope which may include blood and other debris, making the task more difficult [11]. Nowadays, the most common technique is the liquid-base cytology which consists on releasing the cells into a preservative liquid, in order to uniformize the sample and reduce the influence of external artifacts (e.g. blood, debris) [11].

2.6.2 Colposcopy

Colposcopy is a technology that provides a direct observation of the cervix. Nowadays, there are portable and low-cost devices such as EVA COLPO from MobileODT[1], allowing the access of this screening method in remote locations with vulnerable population.

Colposcopy examination should follow a protocol with 4 main steps. In the first step, a normal saline solution should be applied on the cervix to highlight landmarks of the transformation zone. These landmarks can be crypt openings and/or nabothian follicles that define the external boundary of the TZ [14]. A normal crypt opening is represented as a black dot surrounded by a small acetowhite area. If the crypt opening is large and the acetowhite area is denser than normal, it is called a cuffed crypt opening [15]. Also, a cuffed crypt opening may indicate an extension of the neoplasia into the crypt, which only occurs for high-grade lesions [15]. Both squamous and columnar epithelium should be seen in this stage, the squamous epithelium has a pink tone while the columnar epithelium has a dark red tone with grape-like patterns [14].

In the second stage, a green light filter is applied to enhance the blood vessels. Cancer cells have a growth rate higher than usual, so, the tissue around this cells is more vascularized [14].

The third stage is known as Hinselmann test and involves the application of 5% acetic acid solution. The solution enhance squamous and columnar epithelium, facilitating the segmentation of this regions by a human expert [14]. The application of acetic acid solution also enables the observation of precancerous lesions.

The fourth and final stage is the Schiller's test, where Lugol's iodine solution is applied on the cervix. The normal squamous epithelium stain, showing a brown or black tone, while the immature squamous metaplastic epithelium remains with the same color [14]. Iodine does not stain some abnormal patterns such as cervical polyps, improving the discrimination by human eye of normal and abnormal regions in the TZ. During this step, the specialist also pay attention to the velocity of the staining, faster reactions indicate fragile tissues which may be related to neoplasias.

The presence of cervical cancer is manifested by abnormal patterns in the cervix, however, abnormalities' appearance depends on each woman and her age, therefore, cervical cancer is not characterized by one specific appearance, being important to recognize each abnormal feature.

[1] https://www.mobileodt.com/

2.7 Summary

Cervical cancer incidence have been decreasing, however there is still a considerable number of cases and deaths associated to cervical cancer in middle and low-income countries. Therefore, it is important to develop low-cost screening methods to enable the access of health care services among the most vulnerable population. Digital low-cost devices for colposcopy have the potential to overcome this limitation, once they are cheap and portable.

Chapter 3

Literature Review

Several works have adopted different techniques to extract information from digital colposcopy images. The main goal of these studies is to provide tools to help health professionals during a colposcopy examination, regardless of their level of expertise. Previous works developed Computer-Aided Diagnosis (CAD) systems for different tasks that include image quality enhancement and assessment, segmentation of regions of interest, image registration, detection of abnormal regions and/or patterns, classification of TZ type, and classification of the risk of cancer.

In this chapter, we present a background on Machine Learning and Deep Learning tools, the datasets of colposcopy images currently available to support the development of CAD systems, and an overview of the first four tasks aforementioned, including a brief explanation and some methods used in literature. Finally, it is bestowed a state-of-art of CAD systems used as decision support systems for cervical cancer diagnosis.

3.1 Background on Machine Learning and Deep Learning

ML is a category of algorithms that enable computers to learn from data and to improve their performance without being explicitly programmed. This concept includes three types of learning: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

Supervised Learning is used for tasks of classification (when the output variable is a category) and regression (when the output variable is a real number). The algorithms of this ML type are trained using labeled data with the aim of finding generalized patterns that enable them to predict the correct output of new data. In reverse, Unsupervised Learning models are trained with unlabeled data, therefore, the aim is to find similarities between the data points to solve problems of clustering or association. The third type, Reinforcement Learning, corresponds to interactive algorithms that learn with the feedback given by the environment, concerning their performance. Each time the algorithm needs to chose a path, it receives a reward if it chooses the right path, otherwise, it receives a penalty, after this, the algorithm is updated in order to avoid penalties in similar situations.

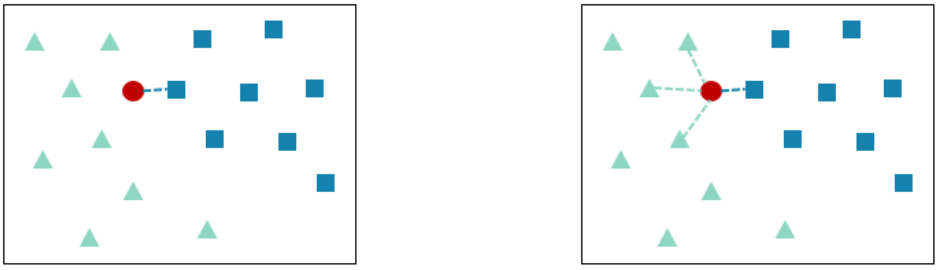
In this dissertation, the goal is to classify the risk of cervical cancer concerning the input data. Once the data is labeled, the models used for this task will be supervised learning algorithms. In this section, it will be presented some models of machine learning used by the authors in the literature and also included in the methodologies of this work. Posteriorly, there is a brief introduction to convolutional Neural Networks and to some models already trained using the ImageNet dataset, a database with more than 14 million images, available online [1].

3.1.1 Machine Learning Classifiers

For small data, it is recommended to use classical machine learning algorithms, yet, those algorithms cannot receive images as input, therefore, it is necessary to extract features from the images and then, use them as the input of the classifiers. Several models were used in literature but, in this dissertation, the classifiers tested were K-Nearest Neighbors, Logistic Regression, Random Forest, and Support Vector Machines, once their performances are usually better than other classifiers. Besides the classical Machine Learning models, it was also tested a Multilayer Perceptron trained with the features extracted from the images. To understand further methodologies, this subsection includes a succinct introduction to the classifiers previously mentioned.

3.1.1.1 K-Nearest Neighbors

kNN, as described in [16], is a method that predicts the classification of unknown samples based on the known classification of its neighbors. This algorithm does not have a training process, *per se*. Instead it organizes the training data in the feature space, and every time it receives an unknown sample, computes the distance between the new sample and all samples in the training set. Afterward, it selects the k nearest samples, where k is the number of neighbors considered in the selection step. Finally, it classifies the sample according to the classification of their selected neighbors. In figure 3.1 there is a practical example of this algorithm.



(a) k = 1 (b) k = 4

Figure 3.1: kNN example. The new sample is represented by the circle. When k=1, the point is assigned to the square class; When the decision rule is k=4, the point is assigned to the triangle class.

[1] http://www.image-net.org/

As represented in figure 3.1, the k parameter is able to affect the decision, therefore it must be selected properly. If k is too large, classes with the bigger population will overwhelm smaller ones, however, when k is too small the overfitting increases.

3.1.1.2 Logistic Regression

Logistic regression is an algorithm also used in classification problems, that computes the probability of the sample **n**, characterized by a set of features $\mathbf{X_n}$, to belong to the class $\mathbf{y_k}$. This algorithm follows a Bernoulli distribution, represented in equation 3.1, where μ is a logistic function represented in equation 3.2, as described in [17].

$$p\left(y_k|X_n\right) = \mu\left(X_n\right)^{y_k}\left(1-\mu\left(X_n\right)\right)^{1-y_k} \tag{3.1}$$

$$\mu\left(x\right) = \frac{1}{1+e^{-\theta^T x}} \tag{3.2}$$

In equation 3.2, θ represents the weight vector which is the parameter that should be optimized during the training of a logistic regression model [17].

3.1.1.3 Random Forest

To understand the concept behind random forest classifiers, it is imperative to understand what is a decision tree. Decision trees are Machine Learning models that classify data considering a net with nodes, branches and leafs, which explain the name of the model. The nodes correspond to questions related with one of the features and correspondent threshold. When there is no need to split the data anymore, a leaf is found. The nodes can be followed by other nodes or by leafs, but the end of each branch should consist of leafs only. In figure 3.2 there is an example of a decision tree used to classify dogs and cats concerning their weight and height, based on a dataset with 5 samples (2 cats and 3 dogs).

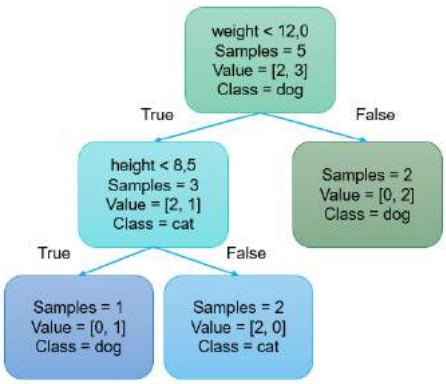


Figure 3.2: Decision tree with two nodes and three leafs, used to classify animals as dogs or cats considering weight and height. Figure adapted from [2].

[2] https://towardsdatascience.com/decision-tree-an-algorithm-that-works-like-the-human-brain-8bc0652f1fc6

Decision trees are very flexible models that easily split and classify each sample of the training data, perfectly, which implies a problem of overfitting. One way to overcome this problem is to use a forest of trees instead of a single tree. In a nutshell, a random forest is a set of decision trees that relies on two key concepts: random sampling of training and random subsets of features. The first concept means that each tree learns from a random set of samples of the dataset and the second means that for each node in each decision tree, only a random subset of features is considered. The final decision is obtained by majority vote or by computing the average of the predictions of each tree [18].

3.1.1.4 Support Vector Machines

The original concept of support vector machines relies on the finding of the optimal hyperplane, in the feature space, that completely splits all samples from two different classes. The data points from each class that are closer to the hyperplane are called the support vectors and those are the only data points that are considered during the process of finding the optimal solution. For the same dataset, there are a lot of candidates for hyperplane, as shown in figure 3.3. To optimize the model, the chosen hyperplane should maximize the margin, i.e. the distance between the hyperplane and the support vectors of each class.

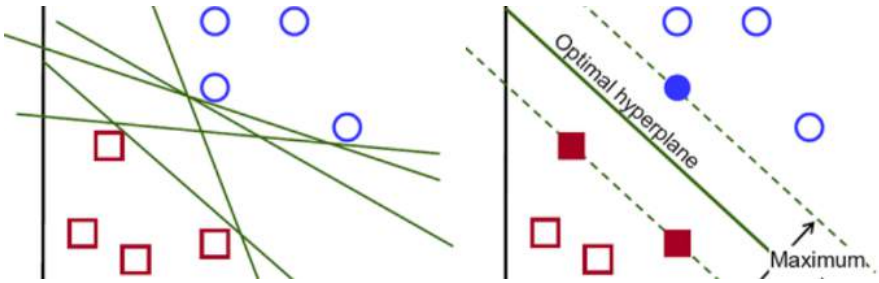


Figure 3.3: On the left, a representation of a feature space with some candidates hyperplanes. On the right, the same feature space with the optimal hyperplane and the support vector represented by filled squares/circles. Figure adapted from [3].

This method requires linearly separable patterns, however, in real datasets, this condition is rarely observed. A method used to overcome this problem is the Kernel trick, which is the application of a Kernel function that map the data into a higher dimensional space [19]. A practical example of the Kernel trick is represented in figure 3.4.

3.1.1.5 Multilayer Perceptron

Multilayer perceptrons (MLP) are one type of neural networks, being closer to the Deep Learning category, depending on the number of layers of the MLP. To introduce the concept of multilayer perceptron, it is important to understand what is a neuron. In a nutshell, a neuron is a linear classifier that combines a weighted sum and an activation function [20], as shown in figure 3.5.

[3] https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f

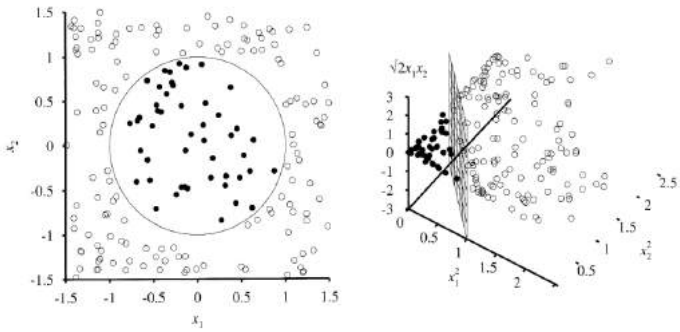


Figure 3.4: Feature space transformation. Adapted from [19].

There is a large set of functions that can be used as activation functions of the neurons, however, the most common is the sigmoid function (equation 3.3), that maps the resulting values into the range (0, 1), and ReLU (equation 3.4), that maps the values between 0 and + ∞. The activation functions can be linear or non-linear, however, if a multilayer perceptron only includes linear functions, the model will be also linear [20].

A multilayer perceptron consists of a network of artificial neurons organized by layers. It starts with the input layer, which is the feature vector, followed by a set of hidden layers, ending at the output layer. Each hidden layer consists of a set of neurons that receive the previous layer as the input and outputs the resulting values from the neurons to the next layer. This concept is represented in the diagram of figure 3.6.

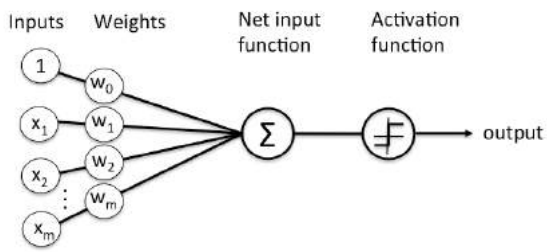


Figure 3.5: Diagram of an artificial neuron. Adapted from [4].

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3.3}$$

$$R(x) = max(0,x) \tag{3.4}$$

Multilayer perceptrons learn in a supervised manner requiring a labeled dataset. During training, the MLP reads the input data and adjust the weights to get the desired output layer, repeating this process several times. Each repetition is called an epoch and the number of epochs is a parameter set by the user. At the end of each epoch, it is computed a loss function and its magnitude define the degree of adjustment that should be applied on the weights [20].

[4] https://skymind.ai/wiki/neural-network

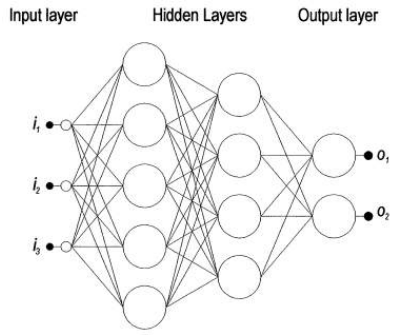


Figure 3.6: Diagram of a multilayer perceptron with two hidden layers. Adapted from [20]

3.1.2 Background on Convolutional Neural Networks

Convolutional neural networks (CNN) are Deep Learning algorithms that can take images as input, avoiding the need for, previously, extracting features from the image. In this subsection, it is presented an overview of CNN, a brief description of some types of layers that commonly integrate these neural networks, and an introduction of CNN architectures already implemented and trained with the ImageNet database.

3.1.2.1 Basics of Convolutional Neural Network

A CNN is a type of neural networks, as multilayer perceptrons, that is fed by input matrices instead of input vectors. To keep the spatial relation of the data, the first layers consists of convolutional operations, receiving matrices as input and returning transformed matrices as output. At the end of the network, the matrices are concatenated in a single vector, so the following layers work as traditional neural network layers. An example of CNN is represented in figure 3.7. For a better understanding, there is a summary description for each type of layers, bellow.

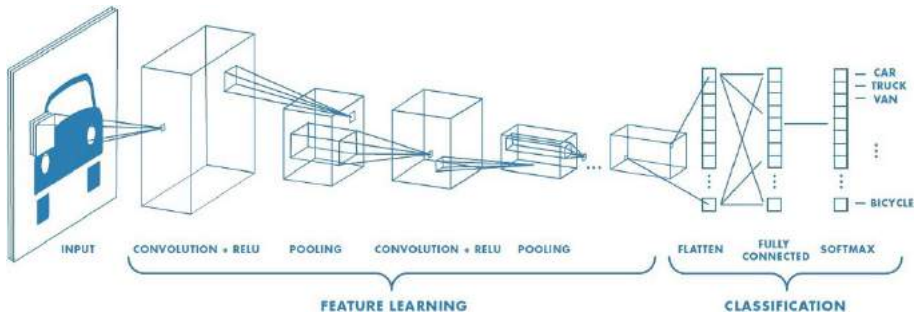


Figure 3.7: Diagram of a convolutional neural network to classify different vehicles. Figure from [5].

Convolutional Layer In a standard neural network layer, the inputs are all connected to every neuron of the following layer. Using this strategy, when the input is an image, would imply

[5] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

thousands or even millions of weights and bias to be computed in each layer; besides, the spatial relation would be lost. In a convolutional layer, it is assumed local connectivity and space stationary, hence, instead of fully connected networks, there is a set of filters/kernels of optional shape that are convolved with the image, i.e. each filter slides across the image computing the dot product, resulting on activation maps. For each filter of shape $W_F \times H_F \times D$, there is W_F*H_F*D weights and 1 bias to be computed, which means that for a layer with K filters the total number of parameters to be learned is $(W_F*H_F*D+1)*K$. The number of activation maps computed in each layer is the same as the filters, K, once a filter only detects a particular localized feature.

Pooling Layer This type of layer is responsible for reducing the size of the feature maps, decreasing the computational cost of the algorithm. There are two types of pooling: Max Pooling and Average Pooling. Both use kernels that slide across the image, but instead of returning the dot product, return the maximum value or the mean of all the values covered by the kernel in each instance of the slide. Between the two pooling, Max Pooling has the best performance, once it can also work as a noise suppressant. In any case, there are no parameters to be learned in this type of layer.

Classification Layers The classification part of a CNN is very similar to a standard neural network. The first step consists of transforming the output of the last convolutional or pooling layer into a vector - **flatten layer**. Following the flattening step, some **fully connected layers**, where each component of the previous layer is connected with every neuron, usually with ReLU activation, are added to the network. For a classification case, the last layer usually is fully connected with **soft-max activation**, which matches each input with the most likely class.

3.1.2.2 Residual Network

Deep networks usually experience a degradation problem, i.e. when the network start converging, the accuracy gets saturated for a while and then degrades abruptly [21]. Adding layers to the network, as an attempt to solve the problem, only led to higher errors. To overcome the degradation problem, He, K. et al [21] proposed the Residual Network architecture that incorporates shortcut connections like the one represented in figure 3.8.

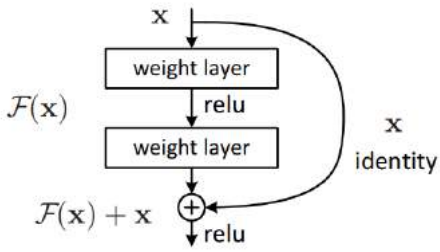


Figure 3.8: Diagram of a residual block. Figure from [21]

Considering that, without shortcut connections, each few stacked layers fit a desired underlying mapping represented by $H(x)$, with shortcuts, those layers fit a new mapping function: $H(x) = F(x) + x$. The shortcut connections only performed identity mapping which does not require extra parameters to be learned or any extra computational complexity [21].

The results in [21] show that ResNet algorithms successfully overcame the degradation problem, once accuracy did not get saturated, and the training error is smaller when compared to plain networks with the same depth.

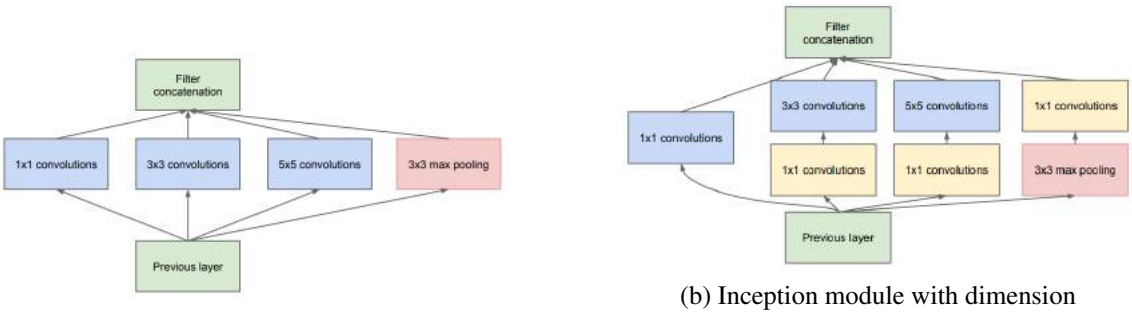
3.1.2.3 VGG

VGG net is a CNN architecture created by Simonyan and Zisserman [22] that won first and second places in the localization and classification tasks in the ImageNet challenge. This architecture has different configurations, the smaller one has 11 weight layers and the deepest has 19 weight layers. All of them are composed by a stack of convolutional layers with filters of size 3×3 interpolated with max pooling layers. The pooling layers follow some of the convolutional layers but not all of them, being five in total. The pooling is performed with a kernel of 2×2 and a stride of 2.

The last max pooling layer is followed by three fully connected layers, two with 4096 neurons, and another with 1000 neurons with a soft max activation, once the ImageNet database contains 1000 different classes [22].

3.1.2.4 Inception V3

Different distributions of the objects on the image require a different analysis. When the information is globally distributed, a larger kernel performs better but a smaller kernel is preferred when the information is locally distributed. Therefore, choosing the size of the kernel can limit the performance of the algorithm. Inception architectures try to overcome this problem by applying different convolutions to the same object and concatenating the results in one single output. The simplest version of inception architectures use modules like the one in figure 3.9a.



(a) Inception module, naive version.

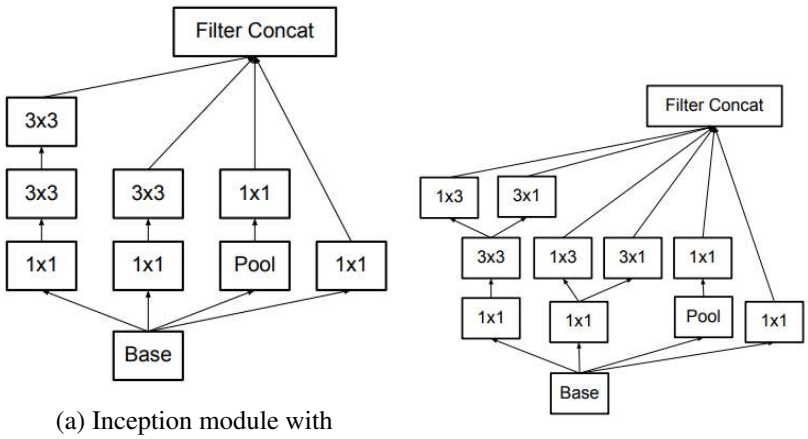
(b) Inception module with dimension reductions.

Figure 3.9: Modules of inception v1. Figures from [23].

To decrease the computational cost, 1×1 convolutions are applied before the 3×3 and 5×5 convolutions, which may seem counterintuitive but 1×1 convolutions are cheaper operations that

reduce the number of input channels, reducing the cost of 3×3 and 5×5 convolutions, which are much more expensive. This update is represented in figure 3.9b.

The authors rethought the inception architecture and created two more versions [24]. In version 2, the modules were updated replacing the 5×5 filters by two consecutive 3×3 filters, as represented in figure 3.10a, to reduce the computational cost. Later, the module was expanded, replacing 3×3 convolutions by two parallel convolutions with filters of shape 1×3 and 3×1, as in figure 3.10b.



(a) Inception module with replacement of each 5×5 filters by 3×3 kernels.

(b) Inception module with expansion of the filter bank outputs.

Figure 3.10: Modules of inception v2. Figures from [24].

The third version of the inception architecture includes the modules of the version 2, and in addition use a RMSProp optimizer and a label smoothing regularization [24].

3.2 Colposcopy Datasets

The decision provided by a CAD system depends on how the system is trained. To achieve a good and complex algorithm, it should be trained with a representative dataset that may include a diversity of abnormalities and acquisition settings, so the algorithm is prepared to evaluate new colposcopy images. Currently, there are four available datasets regarding colposcopy images. The utility of each dataset also depends on the annotations provided by the authors that collected the data. In this section, each dataset will be introduced concerning the content, the propose of the collection, and the annotations available.

3.2.1 Acosta-Mesa H. et al.

In 2009, Acosta-Mesa et al. published a dataset containing 10 videos of colposcopy examination of 10 patients during Hinselmannn stage [25]. This database has no annotations. It can be used to validate image registration methods or segmentation of acetowhite regions, previously segmented

manually, yet it can not be used to train classifiers for cervical cancer diagnosis, once there is no annotations regarding cancer risk.

3.2.2 Fernandes et al.

In a joint collaboration with Fernandes et al. and Hospital Universitario de Caracas, a dataset of 287 images was collected, gathering annotations from 6 experts and images from three modalities (green filter, Hinselmann, and Schiller) [26]. The dataset contains 62 predictive attributes for each image, including segmentation masks of cervix area, external orifice, vaginal walls, speculum, and artifacts, and 7 target variables (one for each patient and a consensus). This database and its notations can be used to train and validate algorithms for image quality assessment and for image segmentation.

3.2.3 NCI/NIH dataset

The American National Cancer Institute (NCI) from National Institutes of Health (NIH) collected a dataset composed by digital colposcopy images (cervigrams) from 10,000 women [27]. A subset of this dataset containing 2,120 cervigrams is available for technical works. Besides the images, the dataset includes information about the patient's age, HPV test, and histology results. Some of this data is missing, including the results from hystologic examination, which should be the ground truth for cancer risk classification problems, resulting in a total of 913 labeled cervigrams. The annotations for histology results correspond to the neoplasia progression level, being categorized as normal, abnormal, CIN2, CIN3, and cancer. In colposcopy examination, each case is classified as high risk or low risk, to convert the categories of the dataset to this two classes, normal and abnormal cases are considered as low risk and CIN2, CIN3, and cancer cases are considered as high risk.

This dataset can be used to train and validate decision support systems for cervical cancer and cervical intraepithelial neoplasia diagnosis, being the dataset used in this dissertation.

3.2.4 Intel & MobileODT

Intel & MobileODT submitted a dataset with about 8000 images for a Kaggle competition. The dataset covers the main colposcopy stages and has annotations about the cervix type regarding TZ location. There are no annotation concerning neoplasia or risk of cancer, but it is considered that every images corresponds to normal cases.

3.3 Computer-Aided Diagnosis systems for digital colposcopy images

3.3.1 Image quality assessment and enhancement

The concept of image quality is commonly related with image definition, focus, distortions, etc, however, in a medical environment, image quality concept goes beyond technical specifications.

When a medical image is recorded for further examination, some aspects should be taken into account such as the visibility of the region of interest, patient's position, the presence of medical artifacts, the presence of blood or other body fluids, among others. Hence, some studies have proposed tools to assess and improve colposcopy image quality.

For image quality assessment, Gu J. and Li W. [28] proposed a methodology to assess images in real time, improving image acquisition. To do so, they implement a framework that includes detection and assessment of the region of interest, contrast dynamic range assessment, blur detection, foreign objects identification, and detection of physical contaminants. The decision model was based on threshold operators and there's no information about the quality of the model. Another approach was proposed by Fernandes et al. [26]. The images were classified as good or bad, concerning image quality. The considered features included area of each anatomical body part, area occupied by artifacts or occluded by specular reflections, maximum area difference between cervix quadrants (4 in total), level of fitness of the cervix to a specific geometric model (convex hull, bounding box, circle and ellipse), distance between external os and the center of the image, and mean and standard deviation for RGB and HSV channels. The decision model was based on an Support Vector Machine (SVM) which is more complex than the previous one.

Regarding quality enhancement, several frameworks were proposed for this task, focusing on specular reflections removal and image normalization. Specular reflection is very common in colposcopy images, however, it degrades pixel information, affecting image interpretation. Different authors tried different approaches such as reflection removal applying the green channel to detect and distinguish large saturated regions from small high contrast areas, the missing information of these glares is filled using Laplace's equation for interpolation and adjusting HSI color space [29]. A different approach suggested by Gordon et al. [30] consists in detecting high brightness regions and low saturation areas and the pixels in the neighborhood with high gradients are considered as candidates of specular reflection pixels. Next, a mixture of two Gaussians is fitted on the saturation map for the HSV color space of the selected pixels, where one of the Gaussians represents a pixel with color information and the other represents white pixels, which are considered as reflection regions. Damaged regions are replaced by "artificial" pixels there are painted according to the surrounding pixels.

For image normalization, some authors proposed simple normalization using equalization [31] while others suggested a color calibration system to be implemented before image acquisition. Despite the better results that normalization provides, normalized systems introduce constraints for already acquired images, or images captured by inexperienced professionals, being an approach with pros and cons.

3.3.2 Semantic image segmentation

Several works developed frameworks to segment cervix from colposcopy artifacts and from vaginal walls. Some of them went further to segment acetowhite regions, squamous and columnar epithelium, and the external os.

The common approach for this semantic segmentation include unsupervised methods such as K-means, Gaussian Mixture of Models (GMM), and Mean shift, concerning a feature space composed by raw color information of Lab and RGB color spaces, color ratios, texture, and spatial information, for instance, the distance between the pixel and the center of the image [27].

The clustering methods aforementioned classify each image as an independent sample, ignoring spatial relation between pixels, therefore, in some frameworks, post-processing is applied to ensure spatial consistency. One of the methodologies, that was used by Gu and Li [28], applied morphological operators to fill small holes. The spatial consistency limitation led some authors to use supervised methods, extracting color and texture features and using SVM models.

The boundaries between the cervix and vaginal walls are very smooth resulting in over-segmentation, especially when some artifacts (i.e. the speculum) were captured. To overcome this limitation, some authors used active contours [32].

To distinguish squamous from columnar epithelium, Lange [29] proposed a framework with a watershed algorithm that identifies the border around the cervix. An iterative watershed algorithm is applied to on a feature space considering the product of green channel and saturation in the HSI color space to split both epithelium. The same method is applied to the red channel feature space to segment the external os.

Fernandes, K. also proposed a framework using deep neural networks and ordinal classification to segment different areas/objects such as the colposcope, vaginal walls, cervix, transformation zone, and external orifice [33]. This approach enabled the segmentation of the entire set of objects achieving a Dice's coefficient of 51,24% in Fernandes et al. dataset and 66,98% in Intel and MobileODT dataset.

3.3.3 Image registration

Image registration is a process that consists in geometrically transforming images of the same object but taken from different perspectives or with different distortions into the same coordinate system. This process is not useful for databases that contain only one image of each patient but it can be applied on the Acosta-Mesa dataset, which contain videos, and in Fernandes et al. dataset to relate images of the three stages of each patient examination.

Medical image registration can be a tricky task. Despite the different viewpoints and modalities (colposcopy stages), there is elastic deformation induced by respiratory motion. To overcome this limitation, three main methodologies where concerned: either rigid or elastic registration, landmark-based, and segmentation based.

Rigid registration combines translation and rotation transformation, considering a static object and a dynamic capture device, while elastic registration considers dynamic objects that undergo distortion. Acosta-Mesa [34] and Garcia-Arteaga [35] used rigid and elastic registration algorithms for colposcopy images on a two-stage approach. First, the phase correlation is applied to remove global translation difference, followed by locally normalized cross-correlation to remove deformation [34].

In more recent works, Acosta-Mesa et al. combined the cross-correlation technique and landmark-based registration [36]. For that purpose, it was considered as landmark anatomical features that show high contrast, such as the external os. In some cases, external os is not a good reference and cervix surface is very smooth, being difficult to track key points on it. Therefore, they propose, as a landmark, the stain introduced by Lugol solution. Other authors used Harris corner detector and similar algorithms to track landmarks over time [37].

3.3.4 Detection of abnormal regions

The presence of abnormal regions may be an indicator of some neoplasia, therefore, the identification of these lesions has high importance in colposcopy examination. These abnormal regions include acetowhite regions, abnormal vascularization, mosaic regions, and punctations.

Several authors proposed frameworks to detect abnormal regions in traditional colposcopy images. In most of the works, when abnormal region segmentation is applied, the main steps include specular reflection removal, cervix segmentation, acetowhite region segmentation, detection of mosaic regions, vasculatures and punctations, and classification [38] [39] [40] [41] [42] [43].

For acetowhite segmentation, Das A. et al. developed a framework that considers cervigram as a GMM and proposed a probabilistic segmentation algorithm based on Expectation-Maximization algorithm [44]. Thereafter, they developed models for classification of abnormal regions as mosaics, vasculatures, and punctations, based on texture features. To segment mosaic regions, images undergo morphological opening, using six structures, combining the six filtered images in a final mask [38]. The same procedure was used for vessels segmentation. For punctations segmentation, it was applied matched filtering using a Gaussian template to map variations of intensity. The intensities distribution was modeled as a mixture of Gaussians, and a variant of the Expectation-Maximization algorithm to assign each pixel to the respective class (background or object) [38]. The Dice metric for this approach was 0.79 for mosaicism, 0.77 for vessels and 0.81 for punctations.

The same approach was, previously, used by Srinivasan Y, using both rotation structuring element (ROSE) and Gaussian modulation of rotation structuring element (GMROSE) for morphological opening [42] and a few examples of the results for vascularization segmentation are represented in figure 3.11. Srinivasan also used filter banks and textons instead of ROSE and GMROSE [43] as a solution for abnormal regions segmentation. Li and Poirson [40] applied a similar morphological transformation for mosaicism and vasculatures, using ROSE, and also for punctation, using a disk as the structuring element, instead.

Gordon S. proposed a methodology for acetowhite segmentation based on watershed segmentation to generate superpixels. After computation of region and edge similarity matrix, superpixels are clustered following a graph-cut criteria [39]. In figure 3.12 there's an example of an original image, cervix segmentation, superpixels generation, and clustering results. Graph-based transduction on superpixels was also applied by Huang [45]. Huang used simple linear iterative clustering (SLIC) as superpixel generator, instead of the watershed, and Graph-based Transduction (GT) for superpixel labeling [45].

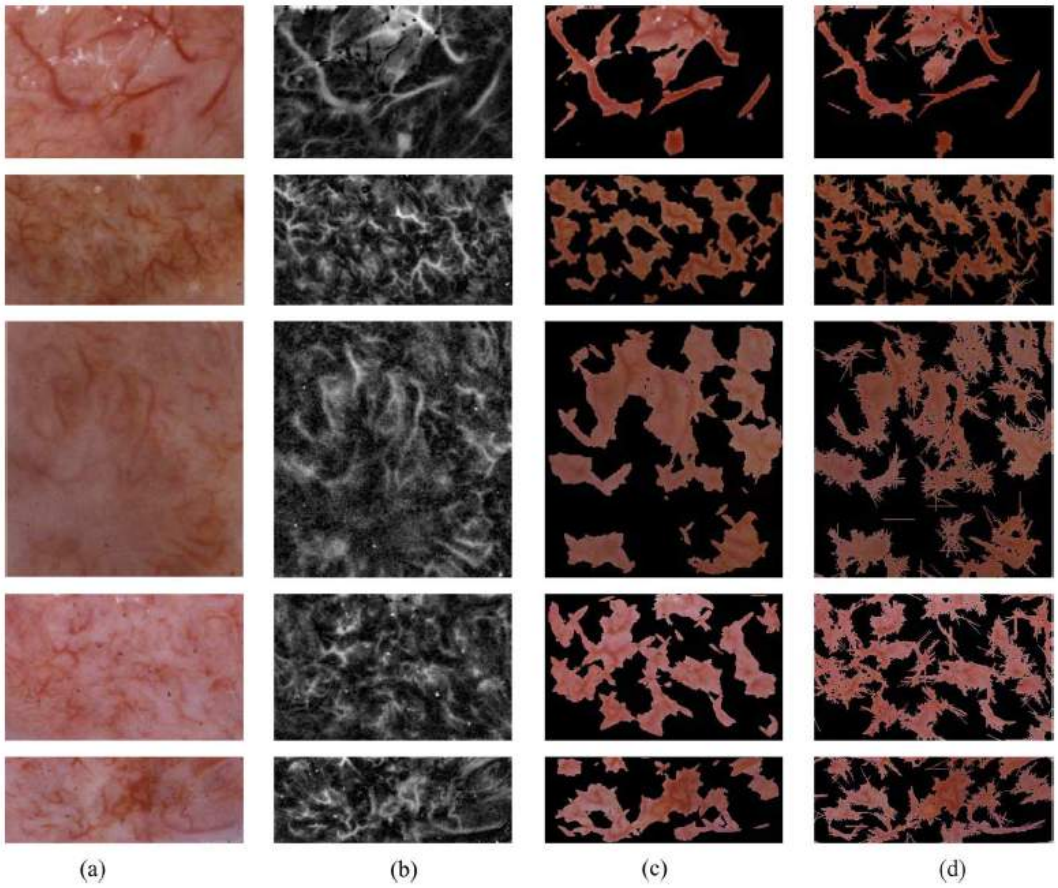


Figure 3.11: Segmentation of vasculature from cervicographic image sections. (a) Original. (b) After morphological opening. (c) Segmented using GMROSE. (d) Segmented using ROSE. Adapted from [42]

3.3.5 Classification of cervical cancer risk

The main goal of colposcopy examination is the detection of cervical intraepithelial neoplasia and the assessment of cervical cancer risk. The tasks previously presented help doctors and other health professionals in the interpretation of colposcopy images, being the diagnosis decision up to the doctor. Other works have proposed CAD systems to help doctors on cervical cancer diagnosis, completing the digital support for this screening method.

Kim [46] developed a pipeline for feature extraction and representation, segmentation of the region of interest, and CIN classification. To represent color and texture features, Kim used spatial pyramids to preserve spatial correlation. To find the region of interest, he used a subset with annotations of the bounding box for the cervix region and compute the similarity between the test image and every image in the subset. The result is given by an optimized bounding box algorithm proposed by Kim et al. [46]. Concerning classification, it was used 2 methods: Support Vector Machine and Majority Vote. The dataset used in this work was the NCI database, where images are labeled as normal, abnormal, CIN2, CIN3, and cancer. Kim considered 2 classes for the

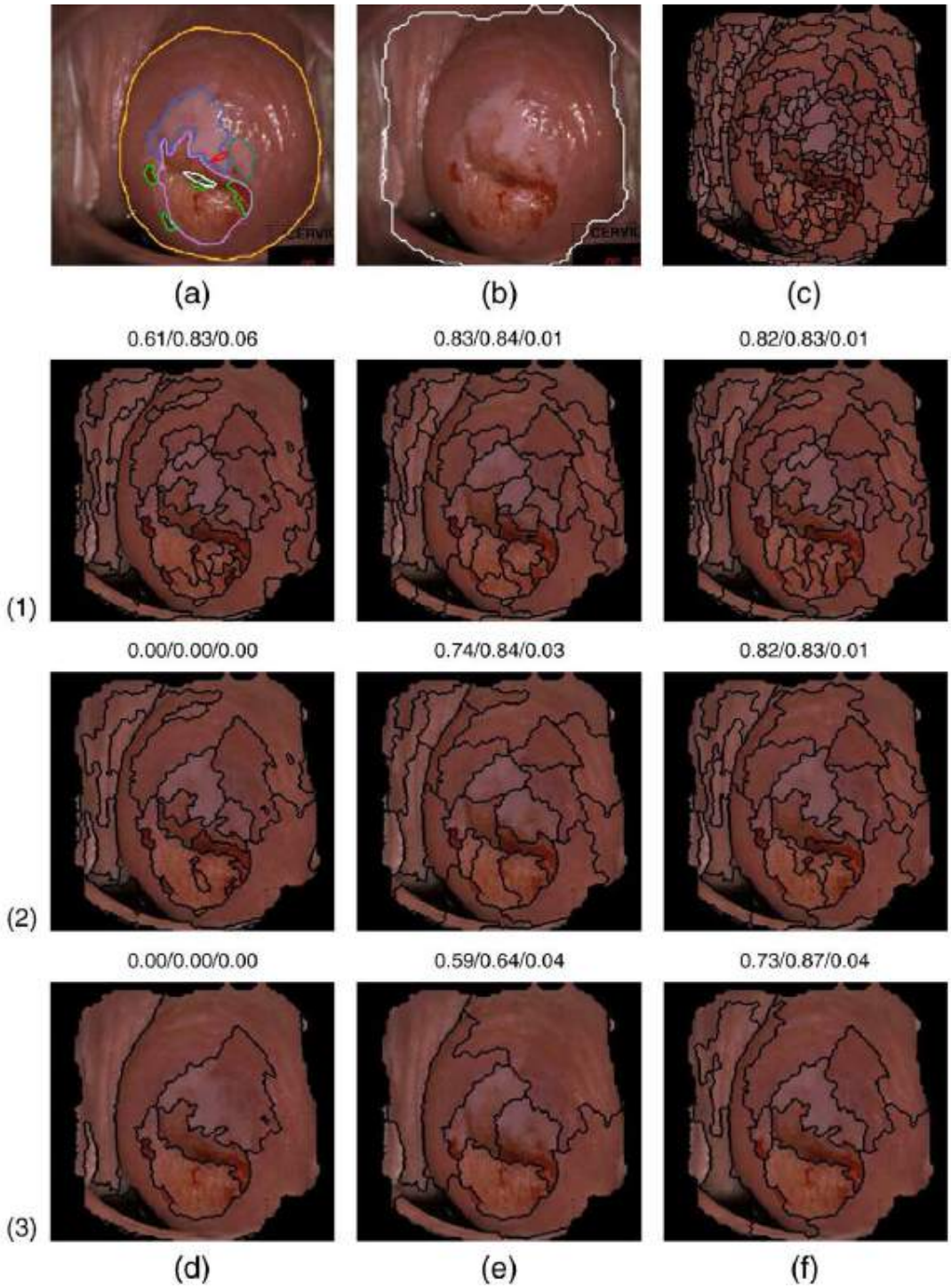


Figure 3.12: (a) Original image with expert markings imposed (yellow for cervix region, blue for AW, and purple for CE); (b) Automatically detected cervix boundary imposed in white; (c) Superpixels boundaries imposed in black on a preprocessed image. Rows (1)–(3) correspond to 50, 30, 10 segments, respectively; Columns (d)–(f) correspond to weighted-mean cut, MinMax cut, normalized-mean cut, respectively. AW Dice/Sensitivity/FP values are listed above each image. Adapted from [39]

classification, one for normal and abnormal and other for CIN2, CIN3, and cancer. Both classifiers used color and texture features previously extracted, concerning the bounding box found in step 2. For Majority Vote, the method used was similar to the one for bounding box: computation of the similarity of each image to every image of the dataset, and minimization of the cost function. The performance of the classifiers was measured using sensitivity and specificity metrics, the results for SVM was 75% and 76% and, for Majority Vote, 73% and 77%, respectively.

Song [47] proposed an approach similar to Kim's, but using multimodal entity coreference. The feature space for this work includes color and texture features extracted from the images and patients data such as age, HPV test, cytology, and pH test. The framework included cervix segmentation using bounding boxes, feature extraction and computation of similarity for patient's data and for cervigrams. The weights for each feature was computed using a gain-based learning approach. Data similarity was computed as the weighted average of similarities of each clinical feature (HPV, pH, cytology, age, and colposcopy). For classification, Song considered 2 classes: negative if normal or abnormal, and positive if CIN2, CIN3, or cancer, and applied a K-means clustering. The best results were achieved when every data type was considered, with an accuracy of 87.43%, a sensitivity of 82.00%, and a specificity of 92.86%.

Liming Hu opted for a Deep Learning approach on the NCI database [48]. The pipeline includes automatic detection of the cervix and prediction of cancer probability. For cervix segmentation, Lu trained an algorithm based on Faster R-CNN. The trained model for segmentation was included in a Transfer Learning process to train other R-CNN for cancer classification. In this work, it was applied data augmentation by performing minor distortions on the original images, such as rotation, mirroring, sheering, and gamma transformation [48]. The model performed a binary classification, clustering labels in the same way than previous authors. To evaluate the performance of the model, Hu split the dataset according to patient's age (younger than 25 years, between 25 and 49, and older than 50 years) and computed the metrics for those 3 subsets. The sensitivity and specificity for each subset was 82.1% and 77.2%, 97.7% and 84.0%, and 92.9% and 83.2%, respectively.

Xu et al has been proposed different approaches to build an accurate model to diagnose cervical cancer [49][50][51][52][53]. The dataset used by these authors was the one from NCI. All the approaches considered two classes: positive for CIN2+ and negative for normal and abnormal results. One of the first work consisted on extracting features from the images, balancing the dataset, and testing different classifiers. In the feature extraction step, three types of features from the images were considered: Pyramid histogram in L*a*b color space (PLAB), Pyramid Histogram of Oriented Gradients (PHOG), and Pyramid histogram of Local Binary Patterns (PLBP). To balance the database, Xu applied an under-sampling technique consisting of keeping the class with fewer samples and randomly remove samples from the larger class until the sizes match [49]. Finally, they compared different classifiers (Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Adaboost, Multilayer Perceptron (MLP), Logistic Regression, SVM, and k-Nearest Neighbors (kNN)) using area under the ROC curve (AUC), accuracy, sensitivity, and specificity as performance metrics. The best classifier for this task was Random Forest, exhibiting the best

results for both imbalanced and balanced datasets. Following the order of the metrics previously mentioned, the performance of RF for the balanced dataset was 84.63%, 80.00%, 84.06%, and 75.94% and for original dataset was 84.83%, 78.24%, 67.54%, and 83,05%. The under-sampling method improved the accuracy and specificity, however, decreased the AUC and the sensitivity.

In a different publication, Xu also considered patients' clinical data [50]. From the previous work to this one, the author kept the feature extraction procedure and the under sampling method and, in addition, he applied isolation of the region of interest. Both text features and image features fed an SVM classifier (one for each kind of features) and the final classification model is a weighted average of text-SVM and image-SVM. The best performance was achieved when the weight for text features was 0.9 and 0.1 for image features with an accuracy of 79.39%, a specificity of 83.03%, and a sensitivity of 76.36% [50].

In further works, Xu et al. used Deep Learning models to extract visual features and predict the risk of CIN [52]. The proposed framework consisted of bounding the region of interest, using ROI to feed a CNN with a structure of an AlexNet, and extracting the features from the last fully connected layer. Image features were further combined with text features to feed a final CNN to learn non-linear correlations across modalities (visual and patients' data). The number of hidden units in the last layer (visual features) is 4096. To avoid overwhelming of the text features, Xu added another fully connected layer with 13 units [52] to compress the vector to 13 dimensions. Finally, visual features (13D) and text features (13D) were concatenated and normalized, using batch normalization [52]. This model achieved better results than the previous works getting an AUC of 94.00%, an accuracy of 88.91%, and a sensitivity of 87.83%.

3.4 Summary

Several works have been developed to improve CAD models concerning colposcopy examination. This chapter provided a background on Machine Learning and Deep Learning techniques, as well as an overview of the published work focused on different tasks, such as image quality assessment and enhancement, semantic image segmentation, medical image registration, detection of abnormalities, and cervical cancer diagnosis.

Despite the differences between tasks, previous works share the same limitations, once they are related with the available datasets. The number of images is relatively small, the annotations are from different nature (cervical cancer risk, TZ type, and segmentation of specific regions), and some images have no annotations at all. It could be a challenge to work with such datasets but also an opportunity to develop and test new methodologies, such as Multitaks Learning.

For this dissertation, the main task is focused on the classification of cervical cancer. In this field, Xu et al. [49][50][51][52][53] are the authors who presented the best results, although there's still space for improvement. Notwithstanding the foregoing, there are ideas and methods proposed by these authors that can be used as a starting point for this dissertation, such as the segmentation of the region of interest as a first step, extraction of canonical features, and implementation of multimodal models.

Chapter 4

Regularization Methodologies For Cervical Cancer Risk Assessment

This dissertation aims to research algorithms to correctly classify the risk of cervical cancer based on colposcopy images. As opposed to previous works, that focused on the application of classical Machine Learning classifiers, this dissertation explores Deep Learning methodologies making use of CNN architectures described in the literature and designing new ones. The database used for these experiments is the NCI/NIH database, previously described in chapter 3, that combines a total of 913 labeled images. The small size of the dataset may cause overfitting, which can be avoided using regularization methodologies. In this chapter, is presented the set of methodologies used to regularize Deep Learning algorithms as well as the results of each methodology.

As shown in figure 4.1, the first step of the proposed framework is the preprocessing of the data, including data augmentation methodologies that are common for every tested model. The subsequent step assesses the benefit of including a segmentation step in the framework. Afterward, Transfer Learning and Multitask Learning methodologies are proposed as well as a regularization methodology making use of features extracted from the images. At the end of the chapter it is presented a description of the training and testing processes followed by the results of each methodology and respective discussion. Methodologies regarding clinical data and imbalanced learning were described only in chapter 5.

4.1 Preprocessing

4.1.1 Database Preprocessing

As mentioned in chapter 3, the subset available from NCI/NIH database include 2,120 cervigrams among other information as patient's ID, patient's age, HPV test, histology results, and time information relating the colposcopy examinations and the histology results. Only 913 cervigrams were labeled, this is, only these images had information about the result of the histology examination. According to the guideline of cervical cancer screening, when the HPV test is negative, there is no risk of cervical cancer and to further examination is performed. To increase the size of the labeled

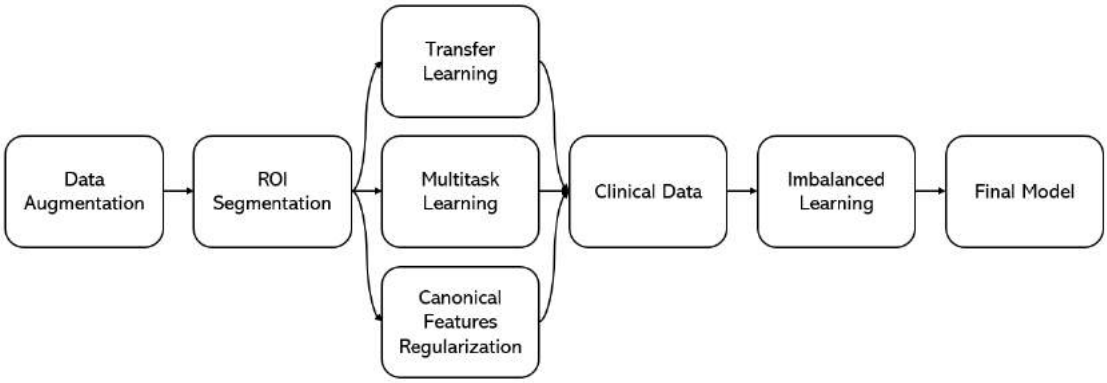


Figure 4.1: Pipeline of the implemented methodologies.

dataset, the cervigrams with a negative result for the HPV test were labeled as low-risk cases, extending the size to 1,886 labeled cervigrams. However, this step increased the imbalance between classes, once the low-risk class has 1,487 samples, while the high-risk class has 399 samples.

4.1.2 Image Preprocessing

Image processing methodologies are commonly included in classification pipelines when the classifiers are classical machine learning models and the features of the image have to be extracted to fed those models. Notwithstanding, deep learning algorithms are able to work with raw data, turning image processing methodologies expendable. Therefore, the preprocessing applied in this framework was image normalization, resize, and data augmentation transformations.

Normalization is a transformation that scales data to fit the range between 0 and 1, regardless the initial range, hence, every input image was normalized. The CNN algorithms implemented in this framework receive the RGB channels of the cervigrams, having an input size of $[224 \times 224 \times 3]$. The cervigrams of the NCI/NIH database have different sizes, however, to fit the neural networks, all of them were resized to 224×224.

Deep learning algorithms can achieve much higher performance compared to traditional machine learning models, being that the deeper the network, the higher the performance range. Yet, the performance of neural networks is strictly dependent on the amount of data available to train those algorithms, and the demand for larger datasets is even bigger for deeper models. This relation is graphically represented in figure 4.2. As previously mentioned, the NCI/NIH database only has 1,886 labeled cervigrams, which is a small amount of data to train deep learning algorithms. Data augmentation is a common approach to attenuate the effect of small databases, like the one used in this dissertation.

To augment the dataset, every image was randomly transformed in each training epoch. The transformations applied included image rotation with a range of 90°, width and height shift, horizontal flip, zoom (in or out), and color saturation. In figure 4.3 there is an example of three random transformations applied to the same image.

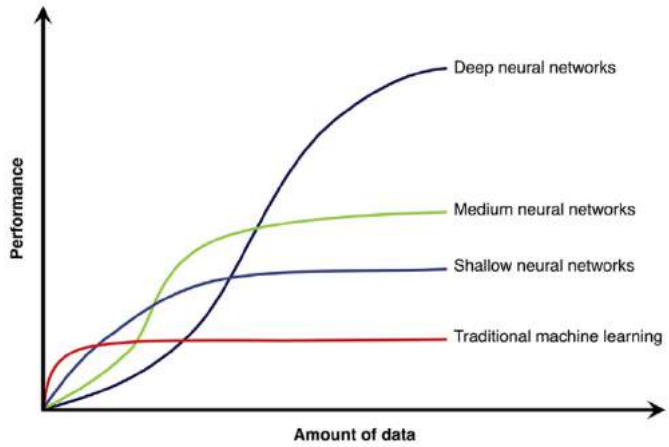
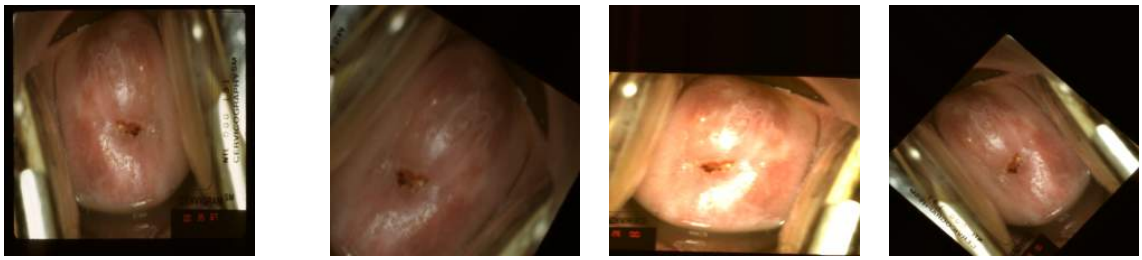


Figure 4.2: Impact of the amount of available data on performance of traditional machine learning and deep learning algorithms. Adapted from [54].



(a) (b) (c) (d)

Figure 4.3: (a) Original cervigram from NCI/NIH database. (b)(c)(d) Three examples of random transformations applied on the cervigram on the left.

4.2 Assessment of Segmentation Relevance

In literature, before feature extraction and image classification, authors commonly extract the region of interest considering the cervix bounding box. The implementation of an automatic tool to extract such a bounding box requires a lot of image processing that increase the computational cost of the algorithm and sometimes leads to errors which compromise the classification task.

To decide if the region of interest extraction step would be included in our pipeline, the images were manually cropped to feed a CNN model. Similar models were trained with three different versions of the dataset - original images, cervix bounding box, and cervix bounding ellipse. The performance of the models were analyzed and compared to assess the importance of the segmentation step.

To extract the cervix bounding box, each image was manually cropped resulting in the rectangle that circumscribes the cervix area (figure 4.4b). The cervix has an elliptical shape, therefore, it was also extracted the cervix bounding ellipse in order to ignore the pixels that do not belong to the cervix area but still appear in the bounding box area. For this purpose, an elliptical mask was

applied to the images resulted from bounding box extraction.

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} \leq 1 \tag{4.1}$$

The equation 4.1 represents the generalized ellipse equation that was applied to each pixel of the mask, where x and y are the coordinates of the pixel, h and k are the coordinates of ellipse's center, and a and b are the radius along each axis. An example of the application of the elliptical mask is represented in figure 4.4c.

(a) (b) (c)

Figure 4.4: (a) Original cervigram from NCI/NIH database. (b) Result of bounding box segmentation. (c) Result of bounding ellipse segmentation.

The architecture of the CNN models trained with each dataset was inspired in VGG architecture, i.e. it is composed by blocks of convolutional layers with kernels of size 3×3 followed by Max Pooling layers. The difference between the two architectures is that each convolutional layer is followed by a pooling layer to rapidly decrease the size of the channels from 224×224 to 7×7. After flattening, there is a fully connected layer of size 30 and ReLU activation followed by another fully connected layer of size 2 with softmax activation.

4.3 Transfer Learning

Transfer Learning is a technique used in Machine Learning field that uses knowledge from one model to apply it in a different task. While traditional Machine Learning models are trained with a dataset to accomplish a specific task, Transfer Learning methodologies take advantage of models previously trained for a task, extracting their knowledge, and using it to train a new model for a different task.

One Transfer Learning approach is the employment of pre-trained neural networks, training them again with new data. A practical example is the application of a CNN trained to classify dogs concerning the size of their ears, transferring this knowledge to a model that classifies dogs concerning the length of their hair. In this case, both models are trained with images with similar objects - dogs. In other applications, the pre-trained model and the ignorant model may not have been trained with images of the same nature. However, the learned knowledge about filtering and

extracting information of an image can be relevant for the new task. Hence, Transfer Learning techniques allows the ignorant learner to achieving satisfactory results even when is trained with a small amount of data. Therefore, in this dissertation, Transfer Learning techniques were used as an attempt to overcome the limitation of the database's size. Besides, augmenting the knowledge from the ignorant learner using a source model usually improves the baseline performance and take less time than training a model from scratch.

The process of retraining a pre-trained model is called fine tuning, and it can be applied to the whole model or just to the last layers. When the datasets are similar, like in dog's example, only the last layers need to be retrained. However, when the dataset for the source model and for the ignorant learner do not share images of the same nature, the fine tuning should be applied to a large part of the whole model.

Some models available in Keras and PyTorch libraries were pre-trained with the ImageNet database, which contains more than 14 million images assigned to over 20.000 categories. The Transfer Learning approach included in this dissertation consisted on fine tuning such pre-trained models available in Keras, in particular, the ResNet model with 50 layers, VGG with 16 layers, and the Inception V3, which the architectures were briefly explained in chapter 3. Despite the large extension of the ImageNet database, it does not include any cervigram or similar images. Then, the pre-trained models were completely fine tuned, when trained with the NCI/NIH database.

4.4 Multitask Learning

Multitask Learning (MTL) is another subfield of Machine Learning where the same model is trained to solve different tasks at the same time. In a nutshell, the structure of a Multitask Learning model can be described as a block of shared layers followed by multiple blocks of task-specific layers, as shown in figure 4.5. This learning process generates more robust models that combine shared features related to the several tasks via shared regularization, being this mechanism more similar to the human brain learning process. MTL can also be considered as a Transfer Learning approach, once the knowledge obtained from one task is transferred to the remaining tasks, however, in Transfer Learning, the knowledge from the source model can be lost to better adapt to the new task, while in Multitask Learning the knowledge from each task is always being updated. Using the dogs' example from the previous subsection, a Multitask Learning model would receive as input images of dogs, returning the type of the ears and the length of the hair at the same time. To solve a multitask, the training dataset should include images with multiple labels (one for each task) and/or multiple subsets of images labeled for one of the tasks.

The cervigrams from the NCI/NIH database are only labeled for the task of predicting the risk of cervical cancer which is insufficient to train a Multitask Learning model. On the other hand, there is the Intel & MobileODT database that contains more than 8,000 cervigrams labeled considering the type of transformation zone of each cervix but does not have any annotation about the risk of cervical cancer. Once the amount of available data for the goal task is considered small for a deep learning approach, we took advantage of the Intel & MobileODT database and joined

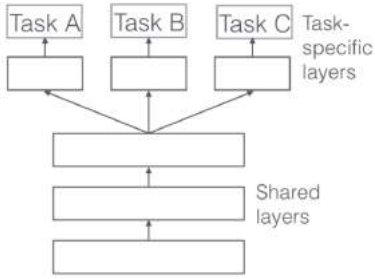


Figure 4.5: Diagram of a generalized multi-task learning model. From [55]

both databases to train Multitask Learning models that simultaneously predict the risk of cervical cancer and the type of the transformation zone. However, since none of the cervigrams are labeled for both tasks, each task needs to be trained, validated, and tested separately but in an alternately way, to update both of them in each epoch.

Three different Multitask learning models were developed in this dissertation. The first and simplest model is similar to the baseline model previously presented with the difference of performing the two tasks aforementioned. The second model integrates the first one and adds a segmentation block with an architecture similar to a U-Net. Finally, there is a third model that performs three tasks: classification of cervical cancer risk, classification of the TZ type, and cervix segmentation.

4.4.1 Model With Two Tasks

The first attempt to include a Multitask Learning approach in this dissertation culminated in the model represented in figure 4.6. This model receives RGB images of size 224×224 followed by a Convolution Block that consists of a set of convolutional layers with Kernels of size 3×3 alternated with Max Pooling Layers. The output of this block is an array of convolved images of size 7×7. Afterward, the array is followed by flattening, which completes the shared part of the model. After flattening, the model is split into two branches, one for each task, that are composed by two fully connected layers. The first one has 30 units and a ReLU activation while the second has the same units as the number of classes for each task (two for Cancer's risk classification and three for TZ type classification) and a softmax activation.

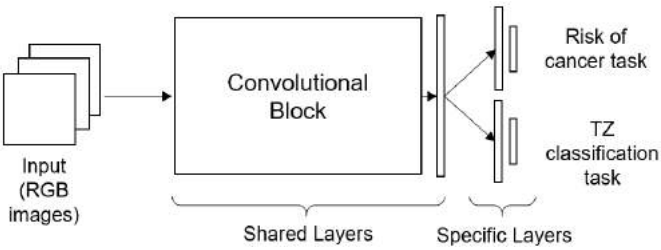


Figure 4.6: Diagram of the model trained for two tasks: classification of the risk of cancer and classification of the transformation zone type.

The loss function used to train this model has two components: one for the task of diagnosing the risk of cervical cancer (L_{CC}) and other for the task of classifying the type of transformation zone (L_{TZ}). Once the data used to train each task is different, the tasks can not be trained at the same time, hence, the loss function, represented in equation 4.2, has a variable $\alpha \in \{0,1\}$ that defines which component is activated during each epoch. Besides the activation factor, each task is weighted by the variable $\omega \in [0,1]$. The cervical cancer related task was trained with a binary cross-entropy loss function (equation 4.3), once it only classifies as low risk or high risk. For the TZ related task, the loss function used is the categorical cross-entropy function which is a generalization of binary cross-entropy. This function is represented in equation 4.4, where M represents the number of classes, y is a binary indicator (0 or 1) that takes the value 1 if class label c is the correct classification for observation o, and p is the predicted probability of observation o to belong to class c.

$$L_{model} = (\alpha)(\omega) L_{CC} + (1-\alpha)(1-\omega) L_{TZ} \tag{4.2}$$

$$L_{CC}(y,p) = -(y\log(p) + (1-y)\log(1-p)) \tag{4.3}$$

$$L_{TZ}(y,p) = -\sum_{c=1}^{M} y_{o,c} log(p_{o,c}) \tag{4.4}$$

4.4.2 Model With Two Tasks and Segmentation

The second MTL model built is represented in figure 4.7. The difference between the previous model and the presented one is the addition of a segmentation block. The main goal of this upgrade is to take advantage of the masks previous extracted to highlight the region of interest.

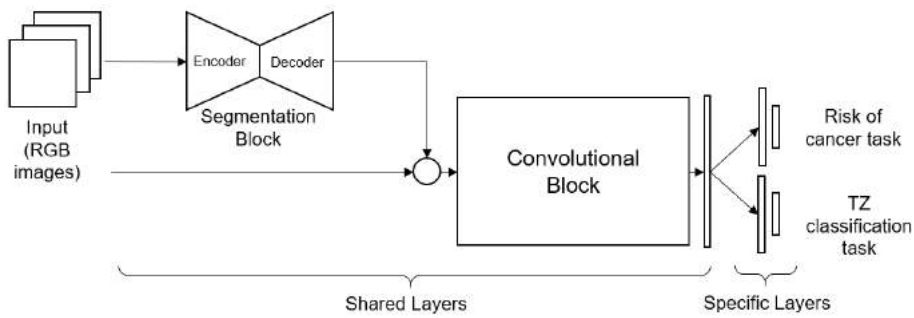


Figure 4.7: Diagram of the model trained for image segmentation, followed by an operation between the original image and the predicted mask, finishing with the performance of two tasks: classification of the risk of cancer and classification of the transformation zone type.

The segmentation block was inspired in the U-net architecture, which is commonly used to segment biomedical images [56]. As shown in figure 4.8, this network has two main parts, an encoder (or contracting path) and a decoder (or expensive path). The encoder component consists

on four blocks of two convolution layers with kernels of size 3×3 followed by Max Pooling with kernels of size 2×2 and the last block, also called as the bottleneck part, is composed by two 3×3 convolutional layers. The decoder component expands the size of the channels to bring them closer to the input size. Therefore, it has four sets composed by one 2×2 upsampling layer, which is an operation that expands each pixel in a square of size 2×2, followed by two 3×3 convolutional layers. Finally, there is a 1×1 convolution that returns the segmentation mask. After each upsampling, the output channels are concatenated with the channels from the correspondent depth of the encoder component. The segmentation block designed for this model differs from the U-net architecture in two aspects: the number of blocks of each component is three instead of four, once the input size of the model is smaller than the U-net's input, and each block only has one convolutional layer instead of two, thus, the encoder structure keep the architecture of the convolutional block from the previous model.

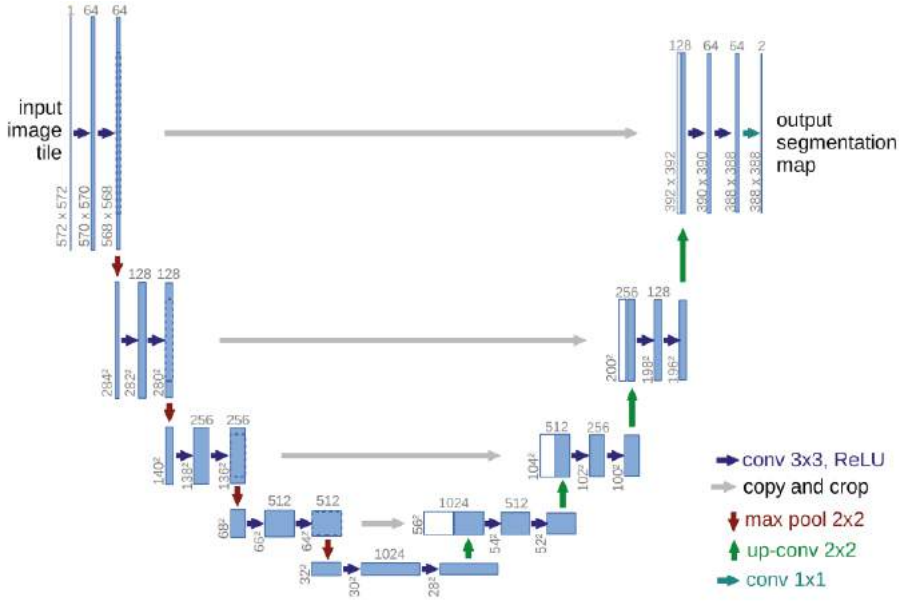


Figure 4.8: Diagram of U-net architecture. Figures from [56]

The segmentation block returns a mask that should mimic the best segmentation option (bounding box or bounding ellipse). After this step, the mask and the original image are subjected to one of the three mathematical operations: multiplication, addition, and concatenation. The result from this operation follows to the second part of the model which has the same architecture of the model with two tasks and no segmentation.

The training process for this model is not quite the same as the first one. At the beginning, the segmentation block was trained individually, using the bounding box or ellipse box masks. Considering the output mask as a 2D array of binary elements, the loss function used to train this part was also the binary cross-entropy (equation 4.3). Posteriorly, the whole model was trained, including the segmentation part, using the alternate mode explained in the previous model (equation 4.2). This enables the algorithm to adapt the segmentation part in order to improve the classification tasks.

4.4.3 Model With Three Tasks

After introducing the segmentation in one of the MTL models, a third approach was designed. The third and last MTL model also includes the segmentation part but, instead of using it to enhance the region of interest, uses it as a third task to be predicted and optimized by the model. Once the encoder structure of the segmentation block is similar to the convolutional block, in the third model, the encoder is the shared component of this MTL model. However, the encoder consists of three sets of one convolutional layer and one Max Polling, while the convolutional block has five sets. Therefore, after splitting the branch for the segmentation task, the other two tasks share a smaller block containing the remaining sets, as represented in figure 4.9.

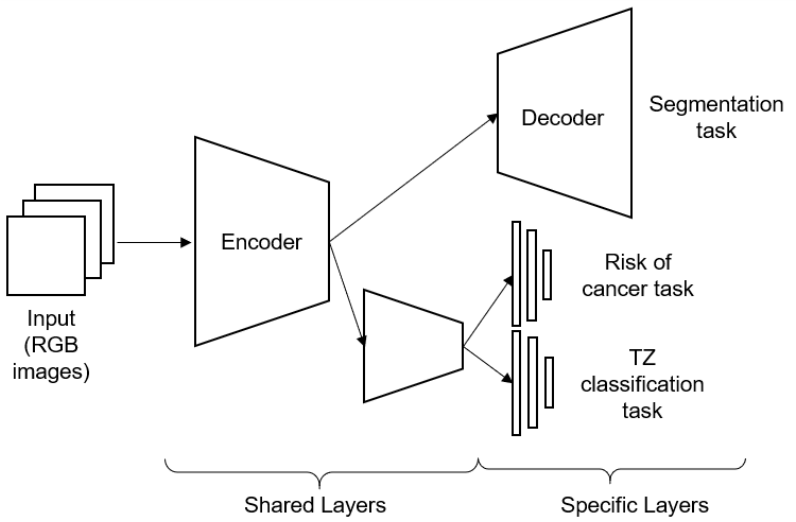


Figure 4.9: Diagram of the model trained for three tasks: cervix segmentation, classification of the risk of cancer, and classification of the transformation zone type.

The manual segmentation mentioned in section 4.2 was only applied to the NCI/NIH, discarding the Intel & MobileODT database in the training of the segmentation task. In consequence, when the loss for the cervical cancer task is activated, so it is the loss for the segmentation task (L_S). On the other hand, when the loss for the TZ task is on, the other two tasks are off, as represented in equation 4.5.

$$L_{model} = (\alpha)(L_{CC} + L_S) + (1 - \alpha) L_{TZ} \tag{4.5}$$

4.5 Regularization With Canonical Features

The approach used by other authors to solve classification problems related to the cervical cancer risk included the extraction of features to feed classification models. Even though convolutional neural networks extract features from the images by applying the convolution of optimized kernels, in this dissertation, it was included a regularization method that forces the models to learn the features that are usually extracted. The implementation of this method contains several steps

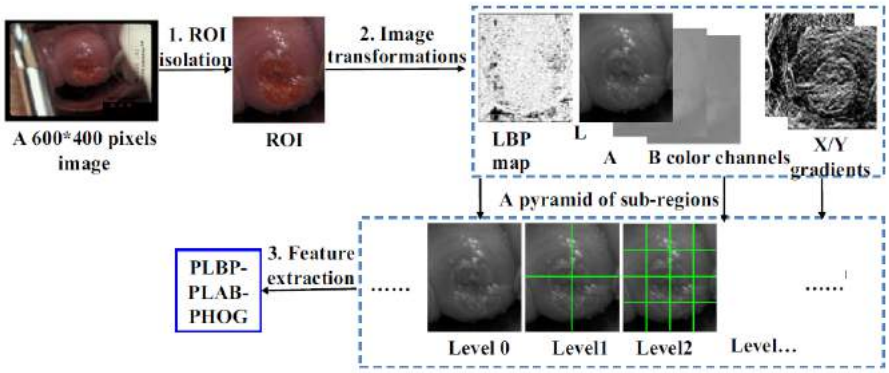


Figure 4.10: Extraction of pyramid features. Figure from [53].

as feature extraction, performance assessment of classical Machine Learning models fed by the extracted features, dimensionality reduction, and regularization of Deep Learning models, which are described below.

4.5.1 Feature Extraction

As explained in the chapter 3, the abnormalities in the cervix that suggest the presence of cervical intraepithelial neoplasia can have a different appearance, therefore, the whole image should be analyzed, instead of focusing on specific characteristics. In [53], Xu T. et al. compares the performance of several Machine Learning models using hand-crafted pyramid features and features extracted from the fully connected layers of a CNN, achieving results that outperform the performance of Pap tests and HPV tests. For that reason, the feature extraction method applied in this dissertation was the pyramidal method described by Xu T. et al. in the aforementioned work and represented in figure 4.10.

The first step performed by Xu et al. was the isolation of the region of interest using the cervix bounding box. After segmentation, the image was transformed into three types of feature maps: L*a*b color space channels, Local Binary Patterns (LBP) map, and Histogram of Oriented Gradients (HOG). Finally, the features were extracted but instead of collecting pixel-wise information, it was extracted multi-scale pyramid histogram features.

The spatial pyramids were built by splitting the image of each level into rectangular sub-regions of the same size. The number of sub-regions for a level l is given by 4^l, i.e. level 0 has 1 region, level 1 has 4 sub-regions, level 2 has 16, and so on. In the end, the levels are concatenated, resulting in a pyramid.

For the color features, the RGB channels were converted into the L*a*b color space, combining that information on 3 3-level pyramids (one for each color channel). The histogram for each sub-region had 16 bins, and each 3-level pyramid has 21 sub-regions, getting a total of 1008 dimensions for the pyramid LAB (PLAB). To extract the LPB maps, the authors considered a rotation invariant version of the original LBP operator, setting it to analyze an 8 equally spaced pixels in a circle of radius 1. To build the pyramid LBP (PLBP), it was considered a 4-level pyramid

(85 sub-regions), represented by histograms of 10 bins, resulting in a pyramid histogram feature of 850 dimensions. The cervigrams were also transformed in HOG maps to extract information about shape and edges. As the PLBP, the pyramid Histogram of Oriented Gradients (PHOG) had 4 levels but the histograms were represented by 8 bins, instead, resulting in a 680 dimension feature. Finally, the features were concatenated culminating in a multi-feature descriptor represented by a vector of size 2538.

4.5.2 Performance of Machine Learning Models

Even though the main goal of this method is to regularize convolutional neural networks with hand-crafted features, it is important to analyze how Machine Learning models behave when fed with these features. To do so, five different classifiers, including an SVM, a kNN with k=5, an MLP with two fully connected layers, a Random Forest with 15 trees, and a Logistic Regression, were trained and tested using the feature vector described above. To overcome the problem of the imbalance, the class weights were set to be inversely proportional to the size of each class. Each classifier was trained and validated with 10-fold cross-validation and the results are represented in table 4.1. Along with these results, in the last row, there are the results from the best model from Xu et al. publication [53].

Table 4.1: Overall performance of Machine Learning classifiers using pyramid features. The table lists the mean ± standard deviation for each metric.

Model	AUC(%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM	74.27 ± 4.49	78.02 ± 5.24	40.76 ± 9.66	85.78 ± 4.85
kNN	73.85 ± 5.49	82.51 ± 3.55	36.74 ± 11.60	92.17 ± 2.49
MLP	69.04 ± 9.69	**83.78 ± 1.94**	26.92 ± 20.59	**95.90 ± 4.50**
RF	**80.59 ± 5.16**	78.79 ± 6.04	**62.39 ± 12.96**	82.23 ± 5.41
LR	71.49 ± 5.75	76.28 ± 2.50	50.59 ± 10.76	81.74 ± 3.05
SVM[53]	80.71 ± 6.15	77.17 ± 6.62	78.55 ± 6.17	75.80 ± 8.39

Analyzing the table 4.1, it is possible to affirm that the replication of the methods proposed by Xu et al. was well performed once the AUC and the accuracy values are similar. Considering the other metrics, our replication has better specificity but worse sensitivity, which means that, despite balancing the class weights, our models are biased towards the larger class. Between our models, the Multilayer Perceptron and the Random Forest present the best results, however, analyzing sensitivity and specificity values, MLP presents more biased results, therefore, the Random Forest was chosen as the best model.

4.5.3 Dimensionality Reduction

A way to regularize a CNN with the canonical features is forcing one of the fully connected layers to learn those features, i.e. forcing the layer to return output values similar to the feature vector.

A limitation with this method is related with the size of the fully connected layers and the size of the feature vector, once this vector has 2538 dimensions while the fully connected layer presented in the previously built models only has 30 units. To overcome this problem, a dimensionality reduction method was implemented, taking advantage of the best model found in the previous step, i.e. the Random Forest classifier.

Random Forests are commonly used as feature selection and dimensionality reduction methods. Building a large forest, it is possible to extract statistical information about each attribute and select the most informative features - feature selection. Another method consists of encoding the information from the feature vector of a sample into the decision path defined by each tree - dimensionality reduction. In a decision tree, the nodes always split one branch in two or more ramifications, i.e there are no converging nodes which implies that each leaf is affiliated to a unique decision path. Consequently, to encode the feature vector, we extracted the index of the leaf where the decision path ends for each tree in the forest. Thus, it was possible to transform a 2538 dimension vector into a 15 dimension array, once the Random Forest model tested before had 15 decision trees.

To ensure that the transformed feature space had relevant information for the classification task, models similar to the Machine Learning classifiers used in the previous subsection were trained and tested with the encoded feature vector. The results of this test were compiled in table 4.2.

Table 4.2: Overall performance of Machine Learning classifiers using pyramid features after dimensionality reduction. The table lists the mean ± standard deviation for each metric.

Model	AUC(%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM	70.64 ± 6.36	**83.56 ± 4.17**	17.60 ± 21.84	**97.29 ± 4.91**
kNN	67.49 ± 4.25	80.40 ± 2.46	28.40 ± 6.09	91.30 ± 2.27
MLP	69.94 ± 5.27	78.36 ± 2.51	41.70 ± 10.65	86.25 ± 2.30
RF	**81.82 ± 6.14**	82.29 ± 3.48	52.85 ± 16.97	88.55 ± 4.08
LR	78.61 ± 4.98	74.68 ± 4.94	**65.47 ± 9.70**	76.51 ± 6.14

It is not possible to conclude if the dimensionality reduction method improved or not the classification task, based on table 4.2, once the RF and LR models show better results but the SVM, the kNN and the MLP models reduced their performance. However, the overall results after dimensionality reduction are similar to the original results which validates the implemented method.

4.5.4 CNN Regularization

The regularization with the canonical features was applied to the best Transfer Learning and Multitask Learning models previous presented, therefore, to implement this method, no extra CNN architectures were built. The goal of this regularization is to guide the learning task, so CNN is

able to extract information that is similar to the hand-crafted features. To implement it, the models were retrained with a new loss function represented in equation 4.6.

$$L_{model} = L_{Classification} + \lambda L_{Regularization} \tag{4.6}$$

The loss function of the model has two parts, one for the classification part, that corresponds to the loss function of each previous model, and a second component related to the regularization. During the training, the models return two outputs (or more in case of MTL). The first is the prediction of the classification task, and the second is the output of the last fully connected layer of each model, which is a 30 dimension vector. To implement the $L_{Regularization}$, the last 15 values of the output vector are disregard and the remaining values are compared to the normalized feature vector, obtained after dimensionality reduction, applying the Mean Squared Error (MSE) function, represented in equation 4.7. Only half of the output from the fully connected layer is considered for regularization to allow the model to extract other information besides the canonical features. The overall loss function also includes a regularization factor (λ) to adjust the influence of this step. For low values, the regularization has no impact, however, when λ is too high, the models neglect the classification task. To better adjust this value, the models were tested for $\lambda \in [0, 0.01, 0.1, 1]$.

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 \tag{4.7}$$

4.6 Training and Testing

Before training and testing, the dataset had to be split into three subsets: train, validation, and test. The splitting step considered the patient identification number instead of the cervigrams, once there are patients with multiple cervigrams that are very similar to each other. To avoid biased results, all the cervigrams from a specific patient were kept in the same subset. The split ratio was 70-15-15, i.e. 70% for training, 15% for validation and 15% for test. However, this is not the real proportion of the size of each subset, because the number of cervigrams per patient varies between 1 and 19.

4.6.1 Training

The goal of the work presented in this chapter was to find the best regularization method to overcome the limitation related to the small size of the database. To perform a fair comparison, the models were set with the same or similar parameters. Every model was built, trained and tested using the TensorFlow's implementation of Keras and the models used for Transfer Learning kept the same structure and parameters set as default in this library. For the original models, previously described, each convolutional layer was set with 32 kernels of size 3×3, the fully connected layers that precede the classification layer all have 30 units, and, excluding the regularization with canonical features, each model or task-related with the classification of the risk of cancer was trained

with a binary cross-entropy loss function (equation 4.3). To avoid overfitting, drop out layers were also added to each model.

Considering the training process *per se*, each model was compiled using the Adam algorithm proposed in [57] for stochastic optimization. The learning rate, previously adjusted to each model, was set to 0.0001 for transfer learning models, and to 0.001 for the remaining algorithms. The batch size was set to 16 for every model, once it was the biggest size that did not raise memory errors. The number of epochs was also previously adjusted to ensure that the model was stabilized, avoiding overfitting, thus, each model or task was trained 150 times (150 epochs). The exception was the segmentation block from the model with two tasks and segmentation, this block was trained with 50 epochs for segmentation but was fine tuned during the training of the two classification tasks, which took 150 epochs, each. Some authors rather keep the best model, i.e. the model from the epoch with the lowest validation loss, and then use it for testing. However, considering that the dataset was split by patient's ID, it was not possible to keep the ratio of classes in every subset. In some cases, almost 90% of validation subset consisted in negative samples. Hence, in the first epochs, the model was biased and every image was classified as negative but those models were saved as the best once the validation loss was very low. To avoid this situation, the models were saved after a given number of epochs.

4.6.2 Testing

Despite the small size of the dataset, there is a big variance between each class. Consequently, two different partitions of the same dataset can lead to very different results. Therefore, to compare the models and the regularization methods, the results are represented as the average and standard deviation of the results from 10-fold cross-validation.

The metrics used to evaluate the models included the metrics on the literature, such as accuracy, the area under the curve (AUC), sensitivity, and specificity, and two more as precision and negative predictive value (NPV). All metrics are described below.

Area Under the Curve: The Receiver Operating Characteristic (ROC) is a curve displayed on a plot where the x-axis represents the false positive rate (FPR), with $FPR = 1 - Specificity$, and the y-axis represents the true positive rate (TPR), also known as sensitivity or recall. Each point of the curve represents a threshold and the values FRP and TRP would have if the model was tested for that threshold. The area under the ROC curve can be used to measure the capacity of the model to separate class. The perfect model would have an AUC near to 1, and a random model or a model that classifies all the samples with the same class would have an AUC of 0.5.

Accuracy: The accuracy is the ratio between the correct predictions and the total of samples tested:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.8}$$

Sensitivity and Specificity: Sensitivity and specificity are represented in the equations below and can be understood as the accuracy for the positive class and the negative class, respectively. In this case, sensitivity represents the fraction of high-risk cases that were correctly predicted, while specificity represents the same fraction for low-risk cases.

$$Sensitivity = \frac{TP}{TP+FN} \tag{4.9}$$

$$Specificity = \frac{TN}{TN+FP} \tag{4.10}$$

$$BalancedAccuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{4.11}$$

Precision and NPV: This two metrics are similar to sensitivity and specificity but have a different meaning for clinical environment. Precision is the fraction between the number of high-risk cases correctly predicted and the total number of positive predicted cases. A low precision leads to unnecessary treatments, which has a high monetary cost. On the other hand, NPV gives the fraction between the true negative cases and all cases predicted as negative. A low NPV leads to misdiagnosed cases, where high-risk patients are treated as low risk, endangering their health.

$$Precision = \frac{TP}{TP+FP} \tag{4.12}$$

$$NPV = \frac{TN}{TN+FN} \tag{4.13}$$

4.7 Results and Discussion

The work developed in this chapter aims to find the best model and regularization method that better performs an automatic cervical cancer screening, overcoming the limitation of the amount of data available, and that will follow for optimization. Thus, the discussion of this part is focused on the proposed methods, ignoring the results achieved by other authors in the literature.

4.7.1 Results After Segmentation

The segmentation of the region of interest using the cervix bounding box was a common step used in literature. However, before applying efforts in segmentation techniques, it was tested the relevance of the cervix segmentation step for deep learning models. For that purpose, the cervix area was extracted using bounding box and bounding ellipse segmentation, as previous described, comparing the classification results with the original dataset. The table 4.3 compiles the results of a baseline model for the three versions of the dataset: original, after bounding box segmentation, and after bounding ellipse segmentation.

Table 4.3: Mean and standard deviation of the baseline model's performance when fed with the original cervigrams, with bounding box segmented images, and with the bounding ellipse segmented images.

	AUC(%)	**Accu. (%)**	**Sens. (%)**	**Spec. (%)**	**Prec. (%)**	**NPV (%)**
Original	**68.25** ± **7.23**	78.86 ± 4.20	**38.35** ± **19.00**	87.37 ± 6.83	39.23 ± 9.65	**87.25** ± **4.49**
B. Box	64.55 ± 4.56	**78.91** ± **6.69**	31.93 ± 12.88	**88.73** ± **8.56**	**39.90** ± **7.85**	86.07 ± 4.06
B. Ellipse	60.36 ± 4.84	76.25 ± 6.52	34.10 ± 14.15	84.77 ± 8.98	32.69 ± 7.89	86.10 ± 3.61

Between the three versions, the bounding ellipse segmented images have the worse performance, being disregarded in the next models. Although the cervix has an elliptical shape, it is not a perfect ellipse, so the mask might ignore cervix zones that would be relevant for the classification task, decreasing its performance. Comparing the original dataset with the bounding box segmented images, the results are very similar and the best metrics' values are divided between the two cases. For the AUC, sensitivity, and NPV metrics, the best performance is achieved using the original images, with a difference of 3.70%, 6.42%, and 1,18% from the bounding box metrics. However, for the metrics which the best performance is achieved with the segmented images, the difference values are almost insignificant, especially considering their high standard deviation.

There are some aspects to infer with these results. The first one is that segmentation methodologies can be disregarded from the framework, once the original images achieve the best results. Also, the similarity between the results from the original images and the bounding box reveals that, even without segmentation, the baseline model is activated to focus on the region of interest, which means the images' background does not affect the classification task.

4.7.2 Transfer Learning Results

Transfer Learning is an approach that helps to overcome dataset limitations, once it uses knowledge learned from other models trained with different datasets. To make use of the advantages of this methodology, three pre-trained models tested in this work: ResNet-50, Inception V3, and VGG-16. During the training, it was applied fine tuning for the entire model, this way, the models can better adjust to the cervix images, improving the feature extraction layers, as well. The results from the pre-trained models are represented in table 4.4 along with the results from the baseline model.

Table 4.4: Mean and standard deviation of the Transfer Learning model's performance.

Model	**AUC(%)**	**Accu. (%)**	**Sens. (%)**	**Spec. (%)**	**Prec. (%)**	**NPV (%)**
Baseline	68.25 ± 7.23	78.86 ± 4.20	**38.35** ± **19.00**	87.37 ± 6.83	39.23 ± 9.65	87.25 ± 4.49
Resnet	74.63 ± 6.54	81.50 ± 4.63	30.64 ± 12.07	92.02 ± 6.02	48.77 ± 10.46	86.53 ± 3.45
Inception	**75.37** ± **6.95**	80.92 ± 4.29	38.12 ± 12.64	89.92 ± 4.14	43.58 ± 10.87	87.47 ± 4.02
VGG	73.67 ± 4.84	**84.41** ± **4.82**	37.66 ± 13.43	**94.09** ± **4.18**	**59.57** ± **13.63**	**87.53** ± **3.65**

All three pre-trained models achieved better performance than the baseline model, as expected.

Between the three best algorithms, the VGG-16 was the model that achieve the best results, therefore, this model was chosen to integrate posterior optimization techniques.

4.7.3 Multitask Learning Results

The implementation of Multitask Learning techniques was a novel approach proposed in this dissertation. As previous described, three models were implemented making use of the NCI/NIH database as well as the Intel & MobileODT database. The first model performs two tasks: the classification of cervical cancer's risk and classification of the cervical transformation zone's type. The second model, besides predicting the two tasks aforementioned, includes a segmentation block to enhance the region of interest. Four versions of this last model were implemented concerning the four operations used to combine the predicted mask and the original image: addition, multiplication, concatenation, and average. Finally, it was implemented a third model that performs 3 tasks: the two aforementioned, and bounding box segmentation. The results of the Multitask Learning models were compiled with the results of the baseline CNN and the VGG-16 algorithm (the best Transfer Learning model tested) in table 4.5.

Table 4.5: Mean and standard deviation of the Multitask Learning models' performance, along with the results from the baseline model and the best Transfer Learning model.

Model	**AUC(%)**	**Accu. (%)**	**Sens. (%)**	**Spec. (%)**	**Prec. (%)**	**NPV (%)**
Baseline	68.25 ± 7.23	78.86 ± 4.20	38.35 ± 19.00	87.37 ± 6.83	39.23 ± 9.65	87.25 ± 4.49
VGG-16	73.67 ± 4.84	84.41 ± 4.82	37.66 ± 13.43	94.09 ± 4.18	59.57 ± 13.63	87.53 ± 3.65
2 Tasks	**75.11 ± 5.24**	84.34 ± 3.80	31.11 ± 13.98	95.67 ± 2.88	61.12 ± 14.40	86.75 ± 4.03
Add	73.86 ± 8.88	**84.86 ± 3.09**	**31.71 ± 15.44**	96.29 ± 2.27	63.80 ± 8.94	**86.92 ± 4.36**
Multiply	62.65 ± 13.01	84.42 ± 3.02	18.50 ± 21.05	**98.20 ± 2.72**	45.99 ± 41.70	85.32 ± 4.22
Concatenate	71.12 ± 11.63	84.74 ± 4.36	29.38 ± 16.90	96.61 ± 3.26	62.74 ± 28.47	86.74 ± 4.58
Average	70.14 ± 11.25	83.93 ± 3.66	30.04 ± 17.04	95.85 ± 3.80	**65.80 ± 20.49**	86.24 ± 3.72
3 Tasks	64.58 ± 13.74	79.19 ± 9.85	25.35 ± 14.45	90.07 ± 10.89	45.43 ± 23.25	85.14 ± 4.07

A careful analysis of table 4.5 allows to make some inferences regarding the quality of the MTL models:

- Concerning the MTL models only, the multiplication version of the two tasks model with segmentation shows one of the worst performance. The multiplication operation can be risky when the model does not perform a perfect segmentation, because it might discard pixels with relevant information, damaging the performance of the classification tasks.

- The 3 tasks model are also not satisfactory. However, it is reasonable that the filters required for a segmentation task may be different from the filters applied on a classification task. Therefore, it is tricky to train the shared part of the CNN, when the 3 tasks are performed.

- Except for the aforementioned models, MTL algorithms exceeded the performance of the baseline model, being considered in further steps of the framework.

- Among the MTL models, the addition version of the two tasks model with segmentation (Add) was the algorithm with the best performance.
- Comparing the Add model with the 2 Tasks model, the metrics' values seem to be quite similar. Besides, both models achieve similar results to the VGG-16, therefore, these three methods are used in further optimization methods.

4.7.4 Feature Regularization Results

Feature regularization applied to cervical images was the second novel approach proposed in this dissertation. The aim of such a methodology is to lead the learning process to learn the features usually extracted in similar problems. To implement this regularization in the previous CNN tested, the models were retrained using a loss function with two components, one for the classification task and other for the regularization part. The regularization influence was adjusted using the factor λ. For $\lambda = 0$ the regularization is discarded from the model training, for $\lambda = 1$ the regularization loss has the same importance as the classification task, and for higher values, the regularization would overwhelm the classification. For this reason, the regularization implementation was tested for $\lambda \in [0, 0.01, 0.1, 1]$.

Table 4.6: Mean and standard deviation of the best models' performance after feature regularization for $\lambda \in [0, 0.01, 0.1, 1]$.

Model	λ	**AUC(%)**	**Accu. (%)**	**Sens. (%)**	**Spec. (%)**	**Prec. (%)**	**NPV (%)**
2 Tasks		75.11 ± 5.24	84.34 ± 3.80	31.11± 13.98	95.67 ± 2.88	61.12 ± 14.40	86.75 ± 4.03
Add	0	73.86 ± 8.88	84.86 ± 3.09	31.71 ± 15.44	96.29 ± 2.27	63.80 ± 8.94	86.92 ± 4.36
VGG		73.67 ± 4.84	84.41 ± 4.37	**37.66 ± 13.43**	94.09 ± 4.18	59.57 ± 13.63	87.53 ± 3.65
2 Tasks		73.50 ± 6.32	85.15± 4.04	26.94± 12.94	97.74 ± 1.91	**73.86 ± 15.22**	86.18 ± 4.33
Add	0.01	70.14 ± 11.25	83.93 ± 3.66	30.04 ± 17.04	95.85 ± 3.80	65.80 ± 20.49	86.24 ± 3.72
VGG		**76.61 ± 6.68**	84.51 ± 4.54	36.40 ± 10.57	94.78 ± 3.11	60.33 ± 13.89	87.46 ± 3.75
2 Tasks		74.64 ± 7.17	84.90 ± 3.51	31.13± 13.23	96.27± 2.09	62.62 ± 14.31	86.89 ± 3.81
Add	0.1	68.83 ± 10.95	84.62 ± 4.13	22.50 ± 14.36	**98.08 ± 1.50**	72.88 ± 15.72	85.50 ± 4.55
VGG		75.18 ± 7.85	84.20 ± 4.29	36.56 ± 14.61	94.29 ± 3.59	59.91 ± 16.56	**87.55 ± 4.00**
2 Tasks		75.17 ± 7.26	**85.67± 3.67**	31.48 ± 12.77	97.12 ± 1.87	70.54 ± 14.99	87.04 ± 3.75
Add	1	71.80 ± 8.06	84.99 ± 3.60	23.51 ± 15.28	97.61 ± 2.05	71.93 ± 20.85	86.08 ± 3.51
VGG		74.19 ± 7.90	82.98 ± 4.37	35.30 ± 17.21	92.88 ± 5.70	54.24 ± 14.09	87.36 ± 3.84

The table 4.6 gather the results for the three best models, regarding the previous analysis, for each λ value. It was expected to observe a direct relation between the regularization factor and the model's performance, however, analyzing the table 4.6, it is not possible to define which λ better suits the problem or even conclude if the feature regularization has a beneficial impact. A fine analysis of the results leads to the following remarks:

- Considering the AUC, the best model was the VGG with a regularization factor of 0.01. However, for 2 Tasks and Add models the AUC decreased when the regularization with $\lambda = 0.01$ was applied. For the 2 Tasks model, the best AUC was achieved for $\lambda = 1$, while the Add better performed when no feature regularization was applied.

- The accuracy of the models hardly changed. For $\lambda = 0.01$, the 2 Tasks and the VGG had a slight gain, but the Add's accuracy decreased. For $\lambda = 0.1$, the accuracy's values were quite similar to the control case ($\lambda = 0$). Finally, for $\lambda = 1$, the first two models improved their performance while the VGG decreased it.

A similar analysis can be done for the Sensitivity, Specificity, Precision, and NPV, once there are gains and losses for some of the models for each λ value. In conclusion, canonical feature regularization does not have a clear impact in the final results, neither positive nor negative, therefore, it is excluded from the further optimization methodologies once its implementation increases the computational cost of the algorithms.

4.8 Summary

The work proposed in this chapter focused on finding a regularized model to perform an automatic cervical cancer screening. The methodologies included data augmentation, assessment of the segmentation impact, Transfer Learning, Multitask Learning with and without segmentation, and regularization with canonical features. The segmentation and the feature regularization did not add value to the CNN models, thus, these methodologies were excluded from the final algorithm.

Among the tested models, two of the proposed MTL architectures and the pre-trained VGG-16 model were the algorithms with the best results, with AUC values around 75%, and accuracy higher than 84%. For this reason, these models will integrate the optimization methodologies proposed in the next chapter.

Chapter 5

Models' Optimization

The chapter 4 focused on the development of methods to regularize the learning process of convolutional neural networks with the aim of finding the best model to predict the risk of cervical cancer development based on cervigrams. In this chapter, the best models previously selected, i.e. the MTL model with two tasks and segmentation, the MTL model with two tasks, and the VGG, are further upgraded employing two approaches: feeding the models with clinical data and applying methodologies to overcome the imbalance data limitation.

5.1 Clinical Data

Clinical history, patient's age, patient's condition, risk behaviors, and test results are very relevant information that should be considered during the screening process. Unfortunately, the NCI/NIH database does not include all that data, nevertheless it contains the age of the patient in 5 years strata, the time interval between the first image collection and the collection of the presented image, the time interval between the present image collection and the worst subsequent histology examination, and the HPV status concurrent with the image.

In the clinical environment, all this information is taken into account during diagnosing, therefore, a realistic algorithm should also consider these variables. In that sense, the models from the previous chapter were transformed into multimodal models that receive simultaneously the cervigrams and the clinical data as input to predict the risk of cervical cancer.

To transform the models into multimodal algorithms, CNN's were set to receive two inputs: an image and a vector. Once the clinical data vector only has 4 dimensions, it is directly concatenated with the vector obtained after the first fully connected layer that precedes the convolutional block. The figure 5.1 represents the implementation described before using the MTL model with 2 tasks as an example. For the remaining models, the implementation was very similar, once the last part of the model follows the same structure.

A problem that usually arises when working with clinical data is the absence of some values, and the NCI/NIH database is no exception. There is a lot of methods to handle missing data that can be split into two categories: deletion and imputation. Deletion includes deleting incomplete

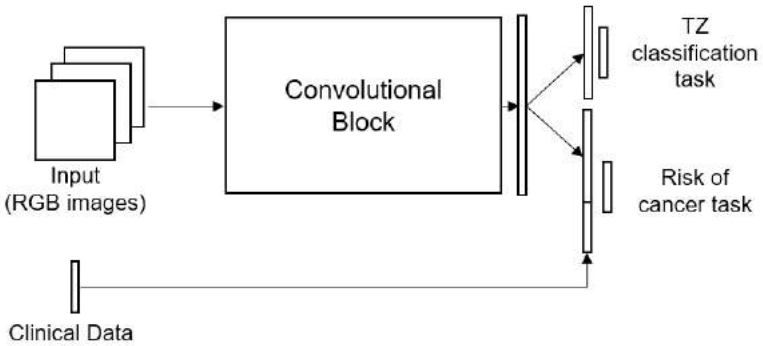


Figure 5.1: Multimodal version of the MTL model with two tasks and no segmentation.

samples or even incomplete rows while imputation replaces missing values using statistical information or more complex models that predict the missing data given the remaining information of the object (in this case, the patient). However, the absence of data might be relevant, for example, when there is no value for the time interval between the present image collection and the worst subsequent histology examination, it is possible that no histology was performed due to the absence of risk factors from the patient. Therefore, missing data was completed using negative numbers to be distinguished from the samples with meaningful values.

Not all features contribute to the classification task, some of them even decrease the performance of the classifiers. To understand the relevance of the clinical data available in the dataset, the baseline CNN was tested for 9 different cases: all features included, only one feature included, and only one feature excluded. In this way, it is possible to evaluate the contribution of each feature as well as the loss due to its absence. Finally, the clinical data was tested alone to check if this data is sufficient to classify the cancer risk. The model used for this test was an MLP with 2 layers, with a structure similar to the last layers of the CNN models, for a fair comparison. Each case was trained and tested applying 40-fold cross-validation to reduce the variance related with the dataset partitions.

5.2 Imbalanced Learning

The imbalance between classes is a limitation present in the NCI/NIH database. As a result, the performance of the models previous tested was biased towards the most represented class. Several methodologies were tested to overcome this problem, including over-sampling and under-sampling methods available in the imbalanced-learn API, a state-of-the-art method envolving ranking strategies, and an over-sampling approach developed in this dissertation. The methods are described below.

5.2.1 Imbalanced-learn API

Imbalanced-learn library has several models to handle imbalanced datasets, however, the employment of these algorithms do not support image objects, requiring feature vectors. In this regard, it was implemented a pipeline to extract the features using the CNN trained models, resampling those features into a synthetic balanced dataset, afterward.

To extract the features, the trained models computed the predictions for each cervigram of the dataset returning the output from the fully connected layer instead of the classification result. The resampling step included 3 different methods: SMOTE, cluster centroids, and SMOTEENN. Synthetic Minority Over-sampling Technique (SMOTE) is an algorithm that increases the number of samples of the minority class, generating artificial samples located in the space between a certain sample and its k closest neighbors [58]. Cluster Centroids is an under-sampling technique that removes samples from both classes, equalizing their sizes. This algorithm includes a k-means estimator that gather the samples from each class into k clusters, returning their centroids. For this test, the k was set to 100, which means that the synthetic dataset only contains 200 samples, in total. The third technique is a combination of an over-sampling algorithm (SMOTE) and an under-sampling technique (Edited Nearest Neighbours). SMOTEENN is a method proposed by Batista, G. et al. in [59] that yield synthetic samples for the minority class while removes samples from both classes, every time they do not match their neighborhood.

After applying to resample techniques to the training data, a Logistic Regression classifier was trained, once this function is similar to the soft-max activation used in the CNN models. The dataset used for testing was not transformed, to achieve more realistic results.

These methods, as well as the following approaches, were tested for the best CNN models found in chapter 4 and for their multimodal versions.

5.2.2 Ranking Model

Cruz, R. et al. proposed in [60] several ranking methodologies in tackling the class imbalance problem. Pairwise ranking algorithms compare each observation against all others predicting each one is "preferred". In this dissertation, it implemented a replication of the RankSVM algorithm proposed in [60], exchanging the SVM classifier for a Logistic Regression.

The implementation requires three steps: pre-processing, training, and post processing. These algorithms rank samples, by comparing them, therefore, the first step is the transformation of dataset into a space of differences, converting the original dataset X into X', where $x'_{ij} = x_i - x_j$, with $y'_{ij} = y_i$, for all pairs that compare observations from different classes. With this transformation, the classes become balanced, once observations from the same class are not compared. The training part uses a Logistic Regression classifier fitted for the space of differences, X'. Considering the decision rule as $w \cdot (x_i - x_j) > 0$, it can be transformed into a scoring function, once $w \cdot (x_i - x_j) > 0 \equiv w \cdot x_i > w \cdot x_j \equiv s(x_i) > s(x_j)$ [60].

A pairwise scoring ranker computes a score for each observation and defines a threshold to predict a class. To find the optimal threshold, the training data is transformed into a score vector,

using the decision rule of the trained classifier. After ordering s_i, the midpoints are computed and used as possible candidates for threshold. The chosen candidate is the one that maximizes the F_1, defined as:

$$F_1 = \frac{2TP}{2TP + FN + FP} \tag{5.1}$$

This method is also not prepared to handle image data, so the features were extracted using the CNN models as described in the previous method.

5.2.3 Over-sampling In Data Augmentation

Unlike the previous methods that make use of the features extracted from the CNN models, this method is applied during the neural networks training, allowing them to fine tune for imbalanced data. This over-sampling technique consists on developing a generator that applies random transformation to the images (data augmentation) and yields the same number of samples for each class, e.g. for a batch size of 16, the generator selects 8 ordered images from the majority class and 8 random images from the minority class, applies random transformations to the images, and finally sends them to model training. The number of steps per epoch is obtained by dividing the length of the majority class by half of the batch size, which means that, during an epoch, each image of the majority class is used only one time, while the images from the minority class are transformed and send to train multiple times.

5.3 Results and Discussion

5.3.1 Clinical Data Analysis

Besides the cervigrams and the results from the worst histology test, the NCI/NIH database includes data about patient's age, result of the HPV test, number of days after image collection (DAIC) date that worst subsequent histology was identified, and the timepoint, i.e. the number of days between the first image collection and the collection of the respective cervigram. This data might be useful for the screening task, thus, the implementation of multimodal models that receive cervigrams and clinical data as input was considered in this dissertation, but previously that data was analyzed to understand which information is more relevant and if any of that could be excluded.

For that analysis, the baseline model was transformed into a multimodal algorithm and was tested for 11 input combinations: only image (baseline results), image and all clinical data, image and clinical data excluding patient's age, image and clinical data excluding patient's HPV test, image and clinical data excluding timepoint, image and clinical data excluding DAIC, image and patient's age only, image and patient's HPV test only, image and timepoint only, image and DAIC only, and only clinical data. The results were computed after a 40-fold cross-validation, being presented in table 5.1.

Table 5.1: Mean and standard deviation of model's performance regarding the input data.

Data	AUC(%)	Accu. (%)	Sens. (%)	Spec. (%)	Prec. (%)	NPV (%)
Baseline	68.25 ± 7.23	78.86 ± 4.20	38.35 ± 19.00	87.37 ± 6.83	39.23 ± 9.65	87.25 ± 4.49
All Data	**90.69 ± 3.25**	**85.73 ± 3.69**	**62.97 ± 13.13**	91.86 ± 4.22	68.33± 11.22	**90.29± 3.90**
w/out age	90.65 ± 3.50	85.71 ± 3.96	62.09 ± 13.69	92.16 ± 4.05	68.63 ± 10.53	90.08 ± 4.26
w/out HPV	79.72 ± 5.91	81.92 ± 4.90	37.22 ± 13.78	93.89 ± 4.35	63.98 ± 14.91	84.87 ± 4.51
w/out timepoint	90.19 ± 3.63	85.31 ± 3.73	61.39 ± 14.63	91.82 ± 3.92	67.61 ± 10.87	89.93 ± 4.41
w/out DAIC	89.23 ± 4.07	84.77 ± 4.24	60.23 ± 13.28	91.37 ± 5.26	66.83 ± 13.00	89.66 ± 4.05
only age	76.63 ± 7.83	81.19 ± 5.49	34.36 ± 14.70	93.19 ± 5.53	60.51 ± 18.74	83.21 ± 8.34
only HPV	89.35 ± 4.52	84.95 ± 5.35	60.97 ± 13.32	90.94 ± 9.87	**68.78 ± 12.15**	88.77 ± 7.21
only timepoint	75.64 ± 13.49	82.31 ± 4.96	33.37 ± 13.49	**95.46 ± 3.08**	66.75 ± 16.63	84.27 ± 4.71
only DAIC	78.14 ± 7.10	82.21 ± 5.37	37.15 ± 12.55	94.15 ± 12.55	64.98 ± 16.92	84.85 ± 4.59
only Clinical	82.28 ± 4.91	80.03 ± 4.11	54.65 ± 12.37	86.81 ± 4.12	52.44 ± 10.27	87.79 ± 4.24

The analysis of the table above allows us to affirm that the inclusion of the clinical data boosts the model's performance once the results of the combination image + clinical data, exceeds the baseline results for all metrics. Using only clinical data seems to be more accurate (AUC of 82.28% and accuracy of 80.03%) than predicting the risk of cancer based on cervigrams only (AUC of 68.25% and accuracy of 78.86%). However, these two sources are not redundant, once the model that combines both image and clinical data (AUC of 90.69% and accuracy of 85.73%) exceeds the performance of the models that analyze them separately.

Focusing on the cases where only one of the clinical features were excluded, it is possible to draw some conclusions:

- The results are very similar and almost achieve the performance of the model that includes all data, except for the combination that excludes the HPV test, which presents the worst performance. Hence, this feature is considered as the most relevant for the screening task, which is very reasonable since this test represents the first step of the cervical cancer screening framework.
- The other three combinations display similar results, therefore, it is not possible to pick the second most relevant feature based on that information.
- Even without the HPV test, the performance exceeds the baseline model, which means the remaining clinical features are pertinent for the screening task.

The group of cases that combine the image with only one clinical feature also provide significant information:

- The four combinations exceed the baseline model, therefore, all clinical should integrate the multimodal model.
- The model trained only with cervigrams and HPV test results achieved the best performance, which supports the aforementioned conclusion about the relevance of this feature.
- Second best performance is achieved when the image is combined with the DAIC feature. DAIC is the number of days between image collection date and the worst histology test. For

some images, the DAIC is higher than 365 days, therefore, it is possible that by the time the image was collected, there were no visible abnormalities on the cervix. Including this feature might help the algorithm to adjust the relevance of the abnormalities.

5.3.2 Multimodal Results

Regarding the previous results, the 4 clinical features were included in the best models found in chapter 4. After transforming them into multimodal algorithms, the three models were trained and tested using 10-fold cross-validation. The results from these tests are presented in the table 5.2.

Table 5.2: Mean and standard deviation of the multimodal model's performance.

Model	AUC(%)	Accu. (%)	Sens. (%)	Spec. (%)	Prec. (%)	NPV (%)
2 Tasks	75.11 ± 5.24	84.34 ± 3.80	31.11± 13.98	95.67 ± 2.88	61.12 ± 14.40	86.75 ± 4.03
Add	73.86 ± 8.88	84.86 ± 3.09	31.71 ± 15.44	96.29 ± 2.27	63.80 ± 8.94	86.92 ± 4.36
VGG	73.67 ± 4.84	84.41 ± 4.37	37.66 ± 13.43	94.09 ± 4.18	59.57 ± 13.63	87.53 ± 3.65
MM-2Tasks	**91.57 ± 1.28**	**88.37 ± 1.81**	**62.60 ± 9.76**	93.91 ± 2.40	68.14 ± 11.55	**92.24 ± 2.91**
MM-Add	89.78 ± 1.57	86.26 ± 2.36	57.41 ± 13.18	92.27 ± 3.57	61.87 ± 10.99	91.26 ± 2.75
MM-VGG	89.61 ± 1.99	87.55 ± 3.85	43.57 ± 14.24	**97.57 ± 0.89**	**74.15 ± 10.03**	88.88 ± 3.91

The effect of the clinical data addition was coherent since all models improved their performance when became multimodal. All metrics improved for the three models, except the specificity for the 2 Tasks and the Add model. These unimodal models have a poor sensitivity, which means they are biased to classify images as negative, therefore, a small decrease on the specificity is not relevant when counterposed with a big improvement of the sensitivity metric.

After implementing multimodality, the architecture that achieves the best results was the 2 Tasks model, nevertheless, the three models were included in further methodologies.

5.3.3 Imbalanced Learning Results

Until this part, the methodologies proposed did not address the problem of the imbalanced classes, in consequence, the models usually present a high specificity but a poor sensitivity, once the positive class is the smaller. To overcome this problem, five methods were applied to the three models previously used (2 Tasks, Add, and VGG), for their both unimodal and multimodal versions. The imbalanced learning methods are described in the present chapter and include the following techniques: SMOTE, Cluster Centroids, SMOTEENN, a pairwise ranking algorithm, and over-sampling in data augmentation. The results were split into three tables according to the architecture of the model, being analyzed separately. Thus, the table 5.3 compiles the results for the 2 Tasks model, the table 5.4 displays the performance of the Add model after applying the mentioned methods, and, finally, table 5.5 presents the results for the VGG model.

The analysis of the table 5.3 shows that imbalanced learning methods accomplished that purpose of balancing the classes, once the sensitivity of the model increased after applying the methods. However, the best overall performance is achieved by the models without imbalanced methods applied. For the unimodal model, there is no unanimity about the method that better suits the

Table 5.3: Mean and Standard deviation of the 2 Tasks model's performance before and after applying each imbalanced learning method.

Method	AUC(%)	Accu. (%)	Sens. (%)	Spec. (%)	Prec. (%)	NPV (%)
w/out	**75.11 ± 5.24**	84.34 ± 3.80	31.11± 13.98	**95.67 ± 2.88**	61.12 ± 14.40	86.75 ± 4.03
SMOTE	66.40 ± 6.78	75.69 ± 12.28	45.44 ± 14.79	82.50 ± 16.05	46.21 ± 16.82	82.90 ± 16.56
Centroids	64.07 ± 6.39	74.80 ± 11.60	47.11 ± 8.77	81.03 ± 14.50	42.20 ± 18.34	87.75 ± 3.35
SMOTEEN	66.78 ± 6.76	76.14 ± 9.08	51.46 ± 14.20	82.10 ± 12.05	42.84 ± 17.76	88.61 ± 3.98
Ranking	74.26 ± 6.73	**84.56 ± 3.21**	36.16 ± 15.16	94.93 ± 2.38	**62.91 ± 16.77**	87.50 ± 3.98
DA	65.17 ± 12.20	51.65 ± 28.60	**70.31 ± 27.25**	48.94 ± 38.99	29.62 ± 14.98	**88.95 ± 4.74**
MM-w/out	**91.57 ± 1.28**	**88.37 ± 1.81**	62.60 ± 9.76	**93.91 ± 2.40**	**68.14 ± 11.55**	92.24 ± 2.91
MM-SMOTE	81.59 ± 2.77	82.47 ± 3.19	80.39 ± 7.50	82.79 ± 5.03	50.41 ± 11.61	95.37 ± 1.69
MM-Centroids	81.18 ± 2.47	81.92 ± 2.82	80.20 ± 6.57	82.17 ± 4.30	48.64 ± 10.36	95.34 ± 1.49
MM-SMOTEEN	81.90 ± 3.39	79.79 ± 2.59	**87.03 ± 8.69**	78.09 ± 3.33	46.23 ± 10.57	**96.71 ± 1.68**
MM-Ranking	91.13 ± 2.19	84.13 ± 3.90	68.74 ± 25.29	87.90 ± 7.87	58.58 ± 19.63	92.57 ± 5.01
MM-DA	90.43 ± 2.25	82.09 ± 3.32	79.45 ± 8.91	82.42 ± 5.80	49.15 ± 10.02	95.31 ± 1.80

problem. The results of the ranking are very similar to the control case (model without imbalanced learning method), so there is no gain from this method. SMOTE, Cluster Centroids and SMOTEENN achieved identical performance, with a slight improvement for the SMOTEENN method. Over-sampling in data augmentation (DA) was the model with the best sensitivity, 70.31%, however, there is a huge decrease of specificity and precision, therefore, it is considered as the worst model.

Focusing on the multimodal model, the best overall performance is achieved when no imbalanced learning method is concerned. Between the methods applied, the SMOTEENN had the best performance. For this case, over-sampling in data augmentation achieved results similar to SMOTE and Cluster Centroids, which contradicts the results from the unimodal model. As observed before, the ranking method is closer to the control case than the remaining methods.

Table 5.4: Mean and Standard deviation of the Add model's performance before and after applying each imbalanced learning method.

Method	AUC(%)	Accu. (%)	Sens. (%)	Spec. (%)	Prec. (%)	NPV (%)
w/out	**73.86 ± 8.88**	**84.86 ± 3.09**	31.71 ± 15.44	**96.29 ± 2.27**	**63.80 ± 8.94**	86.92 ± 4.36
SMOTE	69.94 ± 9.59	83.08 ± 4.05	49.04 ± 21.01	90.83 ± 4.43	53.37 ± 12.12	89.04 ± 5.07
Centroids	69.40 ± 9.47	83.23 ± 4.32	47.40 ± 20.79	91.39 ± 4.87	55.17 ± 14.15	88.76 ± 5.05
SMOTEEN	70.20 ± 9.32	83.01 ± 4.01	49.85 ± 20.84	90.55 ± 4.96	53.49 ± 12.76	**89.20 ± 4.98**
Ranking	70.88 ± 12.01	82.24 ± 3.40	33.76 ± 23.91	93.17 ± 7.28	59.81 ± 23.60	86.82 ± 4.89
DA	72.65 ± 9.07	69.16 ± 19.28	**59.99 ± 71.64**	71.64 ± 24.47	37.64 ± 15.82	89.02 ± 4.60
MM-w/out	89.78 ± 1.57	**86.26 ± 2.36**	57.41 ± 13.18	**92.27 ± 3.57**	**61.87 ± 10.99**	91.26 ± 2.75
MM-SMOTE	80.11 ± 2.49	76.06 ± 13.83	79.18 ± 9.47	75.62 ± 15.57	42.75 ± 15.79	95.0 ± 2.34
MM-Centroids	80.01 ± 2.83	80.90 ± 2.85	78.89 ± 8.60	81.14 ± 5.07	47.06 ± 10.59	95.12 ± 1.88
MM-SMOTEEN	81.70 ± 2.78	79.55 ± 3.19	**85.18 ± 8.96**	78.22 ± 5.46	45.80 ± 9.88	**96.34 ± 2.10**
MM-Ranking	89.73 ± 1.96	79.68 ± 6.68	75.71 ± 20.30	81.37 ± 11.63	49.96 ± 18.48	94.33 ± 4.59
MM-DA	**90.31 ± 1.70**	81.09 ± 4.00	82.59 ± 9.28	80.60 ± 5.79	47.67 ± 8.71	95.83 ± 2.13

Concerning the Add model (table 5.4), the best overall performance is also achieved when no imbalanced learning technique is applied. For the unimodal model, the results follow the same trend observed in the 2 Tasks model: SMOTE, Cluster Centroids, and SMOTEENN present similar results, the over-sampling in data augmentation has the worst accuracy, specificity and precision, and the ranking method present results similar to the control case. For the multimodal model,

SMOTEENN and DA present the best results among the five methods, boosting the sensitivity metric. Concerning the ranking method, its behavior was more related to the remaining methods than with the control case, contrasting with the previous results.

Table 5.5: Mean and Standard deviation of the VGG-16 model's performance before and after applying each imbalanced learning method.

Method	AUC(%)	Accu. (%)	Sens. (%)	Spec. (%)	Prec. (%)	NPV (%)
w/out	73.67 ± 4.84	**84.41 ± 4.37**	37.66 ± 13.43	**94.09 ± 4.18**	59.57 ± 13.63	87.53 ± 3.65
SMOTE	69.61 ± 7.45	82.67 ± 4.20	49.26 ± 14.87	89.95 ± 3.66	51.11 ± 14.56	89.12 ± 4.00
Centroids	68.43 ± 6.87	82.79 ± 5.03	44.09 ± 15.03	91.72 ± 3.78	53.99 ± 13.77	87.89 ± 5.42
SMOTEEN	69.86 ± 7.47	83.00 ± 4.56	49.44 ± 14.71	90.28 ± 3.80	52.12 ± 13.52	89.17 ± 4.00
Ranking	74.61 ± 9.86	82.70 ± 5.35	33.37 ± 19.87	93.01 ± 7.75	**60.16 ± 21.94**	87.02 ± 4.06
DA	**75.87 ± 5.21**	72.30 ± 5.30	**61.57 ± 11.86**	74.77 ± 7.66	35.14 ± 10.52	**90.12 ± 3.66**
MM-w/out	89.61 ± 1.99	**87.55 ± 3.85**	43.57 ± 14.24	**97.57 ± 0.89**	**74.15 ± 10.03**	88.88 ± 3.91
MM-SMOTE	82.57 ± 2.83	81.45 ± 3.73	84.57 ± 7.16	80.57 ± 5.59	48.60 ± 11.30	96.35 ± 1.41
MM-Centroids	82.22 ± 2.35	81.86 ± 3.59	83.02 ± 6.74	81.42 ± 5.51	48.74 ± 10.72	96.05 ± 1.36
MM-SMOTEEN	82.64 ± 3.15	79.42 ± 3.03	87.80 ± 8.89	77.48 ± 5.08	45.58 ± 9.51	96.97 ± 1.87
MM-Ranking	**91.64 ± 2.06**	70.29 ± 12.09	**95.25 ± 4.26**	65.08 ± 14.82	38.15 ± 10.85	**98.65 ± 1.09**
MM-DA	89.67 ± 2.39	83.34 ± 3.30	75.50 ± 11.18	84.60 ± 5.91	51.43 ± 8.06	94.66 ± 2.34

The last table to be analyzed is table 5.5 that shows the results for the VGG-16 model. Again, the control cases reveal the best overall performance, both in unimodal and multimodal models. The ranking method and the control case show similar performance for the unimodal model, however, for multimodality, the ranking algorithm is considered as the best imbalanced method once it boosted the AUC, the sensitivity and the NPV. SMOTE, Cluster Centroids and SMOTEENN achieved similar results, both in unimodal and multimodal models. Concerning unimodality, the over-sampling in data augmentation presented the best results, but the method was overwhelmed by the ranking method for the multimodal experience.

Choosing the best classifying model is not a simple task and several aspects should be concerned. First of all, it is clear that including clinical data boost the model's performance, however, the dissertation aims to find the best image-based model for cervical cancer screening, so the final remarks should consider the best image-based model and the best multimodal model.

The second aspect is the applicability of the algorithm. The model should support medical decision during a colposcopy examination, which is a screening test, while the real diagnosis is only performed by biopsy. For a screening model, it is more important to minimize the number of false negatives than minimize the total of false positives. A false positive has a cost of an unnecessary biopsy, while a false negative leads to a misdiagnosed patient that can develop cervical intraepithelial neoplasia and, consequently, cervical cancer, which might cost a human life. Usually, the metric chosen to assess the capability of a model to better predict the negative cases is the specificity, however, a perfect specificity implies that all negative cases were classified as negative but it does not give information about the false negatives, which are related with human life cost. Instead, in this dissertation, that capability is measured using the NPV metric that indicates the percentage of true negatives among the predicted negative cases.

Taking these considerations, there are four categories for the best models: the best overall image-based, the best overall multimodal, the best image-based screening model, and the best

multimodal screening model. Hence, the models chosen for the four categories are the image-based Add model without any imbalanced learning methodology, the multimodal 2 Tasks model without imbalanced learning, the image-based Add model after SMOTEENN imbalanced learning, once it combines one of the best NPV (89.20%) with a high specificity (90.55%), and the multimodal VGG after Ranking implementation (NPV of 98.65%), respectively.

5.3.4 Final Remarks

Finding the best hyperparameters for a CNN model is a time consuming process. For that reason, the hyperparameters were adjusted for the first models but the grid search process was only applied to the best models, at the end of the experiments. The hyperparameters tuned were the learning rate ($LR \in [0.01, 0.001, 0.0001]$, the moments for the Adam optimizer ($\beta 1 \in [0.9, 0.999]$ and $\beta 2 \in [0.9, 0.999]$), the number of epochs ($NE \in [100, 125, 150, 175, 200]$), and the task weight, $\omega \in [0.5, 0.75, 0.9]$ (only for the MTL models). For Add and 2Tasks models, the best parameters were LR=0.001, $\beta 1$=0.9, $\beta 2$=0.999, NE=150, and ω=0.5. For the VGG model, the best parameters were LR=0.0001, $\beta 1$=0.9, $\beta 2$=0.999, and NE=175. The results of the best models after grid search are represented in table 5.6.

Table 5.6: Results of the best models after Grid Search.

Method	AUC(%)	Accu. (%)	Sens. (%)	Spec. (%)	Prec. (%)	NPV (%)
Img Add	73.86 ± 8.8	84.86 ± 3.09	31.71 ± 15.44	96.29 ± 2.27	63.80 ± 8.94	86.92 ± 4.36
Img Add+SMOTEEN	70.20 ± 9.32	83.01 ± 4.01	49.85 ± 20.84	90.55 ± 4.96	53.49 ± 12.76	89.20 ± 4.98
MM 2Task	91.57 ± 1.28	88.37 ± 1.81	62.60 ± 9.76	93.91 ± 2.40	68.14 ± 11.55	92.24 ± 2.91
MM VGG+Ranking	91.26 ± 2.28	72.06 ± 5.49	95.42 ± 4.12	67.11 ± 6.96	38.01 ± 7.90	98.62 ± 1.30

As presented in chapter 3, several authors proposed automatic systems for colposcopy support, including classification algorithms to predict the risk of cervical cancer. To assess the relevance of the contribution of the presented dissertation, the best models were compared with the proposed systems found in the literature. The table 5.7 gathers the literature results for image-based models while the 5.8 compiles the results for multimodal models.

Sato, M. et al. was the only author who used only deep learning models, achieving the worst results. Song, Kim, and Xu focused on classical Machine Learning methodologies or combination between those models and deep learning techniques. The table 5.7 shows two results for the author Xu, T et al. once they presented results for the best model in normal conditions (A) and results for a model constrained to achieving a specificity of 90% (B). Considering the accuracy and specificity metrics, the best performance was achieved by our Add model, however, Xu's model presented the best AUC and sensibility. It is also clear that, despite efforts with imbalanced learning methodologies, our models were not able to achieve the sensibility that other authors present. Comparing the constrained Xu's model with the Add+SMOTEENN results, the sensibility and specificity values seem to be very similar, but the accuracy of our model exceeds the first by more than 12%, achieving a better overall performance.

Regarding the addition of clinical data to the model, only Song and Xu implemented multimodal algorithms which performance is represented in table 5.8. The approach that achieved the best results was Xu's implementation, nonetheless, the clinical data available for Xu's team included HPV test, patient's age, cervical pH, and cytology result, which is relevant information for the cervical cancer screening that was not available for this dissertation. Considering the approach proposed by Song, individual analysis of the metrics reveals that each metric has been overcome by one of the 2 Tasks and VGG + Ranking models. However, Song's model presents a better

Table 5.7: Comparison between the proposed models' performance and the best results found in literature for automatic cervical cancer screening using image-based models.

Model	AUC (%)	Accu. (%)	Sens. (%)	Spec. (%)
Sato, Masaku et al. [61]	-	50.00	-	-
Song, Dezhao et al. [47]	-	81.93	74.14	89.71
Kim, Edward et al. [46]	-	-	75.00	76.00
(A) Xu, Tao et al. [53]	**82.31**	78.41	**80.87**	75.94
(B) Xu, Tao et al. [53]	-	70.50	51.00	90.00
Add	73.86	**84.86**	31.71	**96.29**
Add+SMOTEENN	70.20	83.01	49.85	90.55

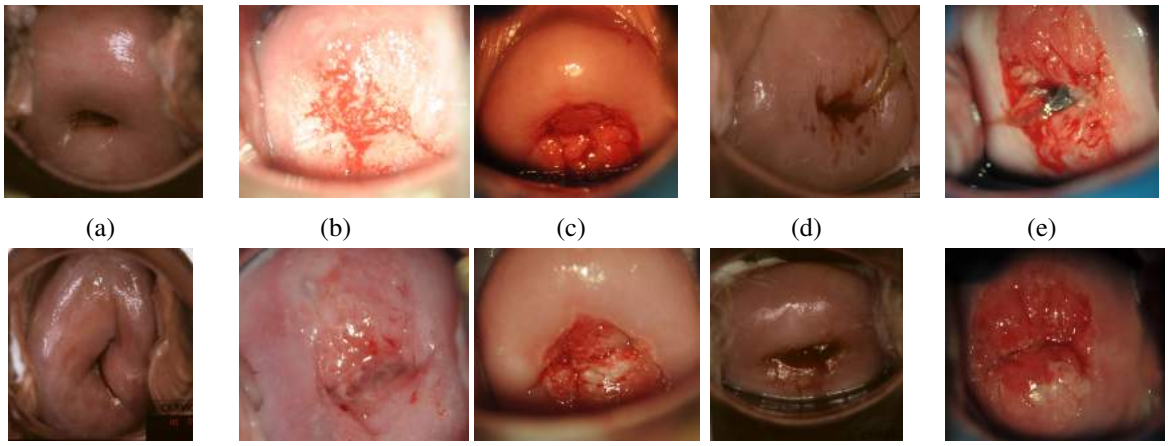
balance between sensitivity specificity that was not achieved by our models. Even so, Song had access to the whole NCI/NIH database that includes data from 10,000 patients [47], therefore, it is not fair to compare our results with this model, once the amount of data available was one of the biggest limitations of this work.

Table 5.8: Comparison between the proposed models' performance and the best results found in literature for automatic cervical cancer screening using multimodal models.

Model	AUC (%)	Accu. (%)	Sens. (%)	Spec. (%)
Song, Dezhao et al. [47]	-	87.79	82.79	92.82
Xu, Tao et al. [52]	**94.00**	**88.91**	87.83	90.00
2 Tasks	91.57	88.37	62.60	**93.91**
VGG+Ranking	91.26	72.06	**95.42**	67.11

A closer analysis of the results presented in this dissertation and in literature makes us wonder if there is a ceiling for the model's performance considering this database. Machine Learning and Deep Learning techniques have been refined during the previous decades which makes relatively easy to achieve good classification results using simple techniques, e.g. fine tuning a pre-trained network with new images. Notwithstanding, several authors concentrated efforts on the cervical cancer screening task using the NCI/NIH database and none of them achieved an accuracy of 90% or higher, being even lower for image-based models.

A look over the database's cervigrams may answer the previous question. For that reason, some images of the database were compiled in figure 5.2. The first row gathers five images from the negative class while the second row combines five positive cervigrams. Observing each row separately it is possible to identify a large intraclass variability, the problem is the similarity between images from different classes. Looking at figure 5.2, it is possible to pair the cervigrams of the first row with the second row by their aspect, yet, they represent different patients and different risk degree, regarding cervical cancer. Thus, it is very hard for a model to learn how to distinguish these cases.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 5.2: (a),(b),(c),(d), and (e) are cervigrams from the NCI/NIH database classified as normal. (f),(g),(g),(i), and (j) are cervigrams labeled for the positive class, i.e. diagnosed with cervical neoplasia or cervical cancer.

5.4 Summary

This chapter focused on two purposes: including clinical data in the model and overcoming the imbalanced classes problem. The clinical data have a clear influence on the screening decision, which was confirmed by the performance boost that the models demonstrate after multimodal implementation. Regarding the imbalanced learning techniques, every method tested accomplished the task of increasing the sensibility, but, in some cases, that implicated a decrease on the model's specificity.

To select the best models, four categories were considered: the best overall image-based model, the best overall multimodal model, the best image-based screening model, and the best multimodal screening model. Among the multimodal models, the 2 Tasks architecture achieved the best overall performance, with an AUC of 91.57% and an accuracy of 88.37%. The VGG-16 + Ranking model achieved the best sensibility (95.42%) and NPV (98.62%), which means that if a cervigram of a patient returns a negative result, there is 98.65% probability of being right.

When compared with other methodologies proposed in the literature, our models were able to achieve similar overall performance but did not exceed their results, which raises the question of how much it is possible to improve using the NCI/NIH database.

Chapter 6

Conclusion

Cervical cancer, as well as other cancers, is more invasive and deadly in advanced stages, having a more efficient treatment when detected in the early stages. Cervical cancer screening programs promote early diagnosis and assessment of cancer risk, providing early treatment and follow up of patients with suspicious test results. Both cytology and colposcopy examinations are complex tasks that depend on specialists opinion, which is a limitation for low-income countries where there are few medical resources. To avoid this dependency and to simplify this task, some authors have proposed automatic systems to support cervical cancer screening, especially in colposcopy examination where the acquisition step is more simple (only requires the collection of cervical images and medical data from patients) unlike cytology, where specialists have to collect cells from the patient.

The main goal of this dissertation was to develop an accurate image-based algorithm to support a medical decision for cervical cancer screening. A few authors proposed classification models with the same aim, making use of conventional Machine Learning techniques. The number of Machine Learning applications is ever-growing, encompassing the medical area. More recently, conventional Machine Learning models have been replaced by Deep Learning techniques that are very promising and tend to outperform older approaches. With that in mind, this dissertation focused on Deep Learning applications to reach the proposed aim.

Deep Learning algorithms require a large amount of data to boost their performance, however, the database available for this dissertation was relatively small, which raised a limitation. To overcome this problem, several methods were developed and tested and the best results were achieved when applied Transfer Learning and Multitask Learning techniques. The performance growth was expected, once these techniques make use of data beyond the NCI/NIH database. The models used for Transfer Learning were pre-trained with the ImageNet database and the Multitask Learning algorithms combined the NCI/NIH and the Intel & MobileODT databases. Other techniques as image segmentation and regularization using canonical features added no gain to model's performance, therefore, they were not considered in the second part of the dissertation.

Besides the cervigrams, NCI/NIH database gathers some clinical data such as patient's age, HPV test result, and temporal information. To take advantage of that information, the clinical

data were analyzed to better understand the relevance of each feature regarding the screening task. The HPV test was the feature that caused more impact on the model, although every feature contributed to the performance gain. For that reason, all clinical data were used in the multimodal models, which were designed to receive both cervigrams and clinical data as input. Despite the small number of clinical features, the multimodal models achieved much better results than image-based models, which makes us wonder how the models would perform if more data, such as DST history, cytology results, and sexual history, was collected.

Besides the small amount of data, the dataset had a second limitation: imbalanced classes. In chapter 5 several approaches were tested to overcome that problem including SMOTE, Cluster Centroids, SMOTEENN, a pairwise ranking algorithm, and over-sampling in data augmentation. Among those models, the SMOTEENN and the ranking method achieved the best results. However, every method decreased the overall performance of the models, being only considered for the screening categories, where sensibility and NPV overwhelm the remaining metrics. Nonetheless, the proposed methodologies were not able to outperform literature results. In the other hand, a quick look over the database is enough to realize how similar images from different classes can be, which hampers model's training, so, it is possible that the proposed methods already achieved the performance ceiling that is feasible for this database, however, it is not possible to be sure of that.

6.1 Future Work

There is a lot to improve in this challenging task of finding the best automatic system for cervical cancer screening, but the biggest limitation seems to be related to the database. As mentioned before, the imbalance between classes and the small amount of data are limitations intrinsic to the database, that can be fixed by implementing the adjustable methods or by collecting more data.

It is also important to rethink the image collection procees. Colposcopy is a dynamic examination that combines four steps. For that reason, gynecologists affirm that taking a decision based on a single image is a tricky task that might lead to misdiagnosed cases. In consequence, some questions were raised during this dissertation. Is it possible to correctly screen cervical cancer based on a single image? Which clinical data should be added to the screening model? As a dynamic examination with 4 steps, would it achieve better results getting a cervigram from each phase? Which one is more promising, an image-based model or a video-based model?

Unfortunately, it is not possible to answer these questions based on the presented work, yet they should be considered in further projects, especially in data collection protocols.

Bibliography

[1] WHO | Cervical cancer. *WHO*, 2018. URL: https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/.

[2] World Health Organization. Cervical Cancer Statistics. Technical report, 2018. URL: http://gco.iarc.fr/today.

[3] Human papillomavirus (HPV) and cervical cancer. URL: https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer.

[4] CUF Instituto de Oncologia. Cancro do colo do útero. URL: https://www.saudecuf.pt/oncologia/o-cancro/cancro-do-colo-do-utero.

[5] Instituto Nacional de Estatística. Anual - INE, Óbitos por causas de morte. URL: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0008281&contexto=bd&selTab=tab2&xlang=pt.

[6] Ewert Bengtsson and Patrik Malm. Screening for cervical cancer using automated analysis of pap-smears. *Computational and mathematical methods in medicine*, 2014, 2014.

[7] Ashrafun Nessa, Joya Shree Roy, Most Afroza Chowdhury, Quayuma Khanam, Romena Afroz, Charlotte Wistrand, Marcus Thuresson, Malin Thorsell, Isaac Shemer, and Elisabeth Andrea Wikström Shemer. Evaluation of the accuracy in detecting cervical lesions by nurses versus doctors using a stationary colposcope and gynocular in a low-resource setting. *BMJ open*, 4(11):e005313, 2014.

[8] Female Reproductive Organ Anatomy: Overview, Gross Anatomy, Microscopic Anatomy. URL: https://emedicine.medscape.com/article/1898919-overview.

[9] Anatomy and Physiology of the Female Reproductive System – Anatomy and Physiology. URL: https://opentextbc.ca/anatomyandphysiology/chapter/27-2-anatomy-and-physiology-of-the-female-reproductive-system/.

[10] Walter Prendiville and Rengaswamy Sankaranarayanan. *Colposcopy and treatment of cervical precancer*. International Agency for Research on Cancer, World Health Organization, 2017.

[11] Cecelia H Boardman. Cervical cancer, 2018. URL: https://emedicine.medscape.com/article/253513-overview.

[12] National Health Service. Cervical cancer - symptoms, 2018. URL: https://www.nhs.uk/conditions/cervical-cancer/symptoms/.

[13] National Health Service. Cervical cancer - prevention, 2018. URL: https://www.nhs.uk/conditions/cervical-cancer/prevention/.

[14] John W Sellors and Rengaswamy Sankaranarayanan. *Colposcopy and treatment of cervical intraepithelial neoplasia: a beginner's manual*. Diamond Pocket Books (P) Ltd., 2003.

[15] World Health Organization. Atlas of colposcopy – principles and practice. URL: http://screening.iarc.fr/atlascolpodetail.php?Index=49&e=.

[16] Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos. k-nearest neighbor classification. In *Data Mining in Agriculture*, pages 83–106. Springer New York, 2009. URL: https://doi.org/10.1007/978-0-387-88615-2_4, doi:10.1007/978-0-387-88615-2_4.

[17] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[18] Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[19] Vikramaditya Jakkula. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37, 2006.

[20] Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[25] Héctor-Gabriel Acosta-Mesa, Nicandro Cruz-Ramírez, and Rodolfo Hernández-Jiménez. Aceto-white temporal pattern classification using k-nn to identify precancerous cervical lesion in colposcopic images. *Computers in biology and medicine*, 39(9):778–784, 2009.

[26] Kelwin Fernandes, Jaime S Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian conference on pattern recognition and image analysis*, pages 243–250. Springer, 2017.

[27] Kelwin Fernandes, Jaime S Cardoso, and Jessica Fernandes. Automated methods for the decision support of cervical cancer screening using digital colposcopies. *IEEE Access*, 6:33910–33927, 2018.

[28] Jia Gu and Wenjing Li. Automatic image quality assessment for uterine cervical imagery. In *Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment*, volume 6146, page 61461B. International Society for Optics and Photonics, 2006.

[29] Holger Lange. Automatic glare removal in reflectance imagery of the uterine cervix. In *Medical Imaging 2005: Image Processing*, volume 5747, pages 2183–2193. International Society for Optics and Photonics, 2005.

[30] Shiri Gordon, Gali Zimmerman, Rodney Long, Sameer Antani, Jose Jeronimo, and Hayit Greenspan. Content analysis of uterine cervix images: initial steps toward content based indexing and retrieval of cervigrams. In *Medical Imaging 2006: Image Processing*, volume 6144, page 61444U. International Society for Optics and Photonics, 2006.

[31] Farnaz Rouhbakhsh, Fardad Farokhi, and Kaveh Kangarloo. Effective feature selection for pre-cancerous cervix lesions using artificial neural networks. *International Journal of Smart Electrical Engineering*, 1(3), 2012.

[32] Viara Van Raad and Andrew P Bradley. Active contour model based segmentation of colposcopy images of cervix uteri using gaussian pyramids. In *6th International Symposium on Digital Signal Processing for Communication Systems (DSPCS'02)*, 2002.

[33] Kelwin Fernandes and Jaime S Cardoso. Ordinal image segmentation using deep neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[34] Héctor-Gabriel Acosta-Mesa, B Zitova, HV Rios-Figueroa, Nicandro Cruz-Ramirez, A Marin-Hernandez, Rodolfo Hernandez-Jimenez, Bertha E Cocotle-Ronzon, and Efrain Hernandez-Galicia. Cervical cancer detection using colposcopic images: a temporal approach. In *Computer Science, 2005. ENC 2005. Sixth Mexican International Conference on*, pages 158–164. IEEE, 2005.

[35] Juan D García-Arteaga, Jan Kybic, and Wenjing Li. Automatic colposcopy video tissue classification using higher order entropy-based image registration. *Computers in biology and medicine*, 41(10):960–970, 2011.

[36] Héctor-Gabriel Acosta-Mesa, Nicandro Cruz-Ramírez, Karina Gutiérrez-Fragoso, Rocío-Erandi Barrientos-Martínez, and Rodolfo Hernández-Jiménez. Assessing the possibility of identifying precancerous cervical lesions using aceto-white temporal patterns. In *Decision Support Systems, Advances in*. InTech, 2010.

[37] Juan D García-Arteaga and Jan Kybic. Automatic landmark detection for cervical image registration validation. In *Medical Imaging 2007: Computer-Aided Diagnosis*, volume 6514, page 65142S. International Society for Optics and Photonics, 2007.

[38] Abhishek Das, Avijit Kar, and Debasis Bhattacharyya. Detection of abnormal regions of precancerous lesions in digitised uterine cervix images. In *Electrical Engineering Congress (iEECON), 2014 International*, pages 1–4. IEEE, 2014.

[39] Shiri Gordon and Hayit Greenspan. An agglomerative segmentation framework for non-convex regions within uterine cervix images. *Image and Vision Computing*, 28(12):1682–1701, 2010.

[40] Wenjing Li and Allen Poirson. Detection and characterization of abnormal vascular patterns in automated cervical image analysis. In *International Symposium on Visual Computing*, pages 627–636. Springer, 2006.

[41] Qiang Ji, John Engel, and Eric Craine. Texture analysis for classification of cervix lesions. *IEEE Transactions on medical imaging*, 19(11):1144–1149, 2000.

[42] Yeshwanth Srinivasan, Enrique Corona, Brian Nutter, Sunanda Mitra, and Sonal Bhattacharya. A unified model-based image analysis framework for automated detection of precancerous lesions in digitized uterine cervix images. *IEEE Journal of Selected Topics in Signal Processing*, 3(1):101–111, 2009.

[43] Yeshwanth Srinivasan, Brian Nutter, Sunanda Mitra, Benny Phillips, and Eric Sinzinger. Classification of cervix lesions using filter bank-based texture mode. In *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on*, pages 832–840. IEEE, 2006.

[44] Abhishek Das, Avijit Kar, and Debasis Bhattacharyya. Early detection of cervical cancer using novel segmentation algorithms. *Invertis Journal of Science & Technology*, 7(2):91–95, 2014.

[45] Sheng Huang, Mingchen Gao, Dan Yang, Xiaolei Huang, Ahmed Elgammal, and Xiaohong Zhang. Unbalanced graph-based transduction on superpixels for automatic cervigram image

segmentation. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 1556–1559. IEEE, 2015.

[46] Edward Kim and Xiaolei Huang. A data driven approach to cervigram image analysis and classification. In *Color Medical Image analysis*, pages 1–13. Springer, 2013.

[47] Dezhao Song, Edward Kim, Xiaolei Huang, Joseph Patruno, Héctor Muñoz-Avila, Jeff Heflin, L Rodney Long, and Sameer K Antani. Multimodal entity coreference for cervical dysplasia diagnosis. *IEEE Trans. Med. Imaging*, 34(1):229–245, 2015.

[48] Liming Hu, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S Jaiswal, Brian Befano, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *JNCI: Journal of the National Cancer Institute*, 2019.

[49] Tao Xu, Cheng Xin, L Rodney Long, Sameer Antani, Zhiyun Xue, Edward Kim, and Xiaolei Huang. A new image data set and benchmark for cervical dysplasia classification evaluation. In *International Workshop on Machine Learning in Medical Imaging*, pages 26–35. Springer, 2015.

[50] Tao Xu, Xiaolei Huang, Edward Kim, L Rodney Long, and Sameer Antani. Multi-test cervical cancer diagnosis with missing data estimation. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94140X. International Society for Optics and Photonics, 2015.

[51] Tao Xu, Edward Kim, and Xiaolei Huang. Adjustable adaboost classifier and pyramid features for image-based cervical cancer diagnosis. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 281–285. IEEE, 2015.

[52] Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. Multimodal deep learning for cervical dysplasia diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 115–123. Springer, 2016.

[53] Tao Xu, Han Zhang, Cheng Xin, Edward Kim, L Rodney Long, Zhiyun Xue, Sameer Antani, and Xiaolei Huang. Multi-feature based benchmark for cervical dysplasia classification evaluation. *Pattern recognition*, 63:468–475, 2017.

[54] An Tang, Roger Tam, Alexandre Cadrin-Chênevert, Will Guest, Jaron Chong, Joseph Barfett, Leonid Chepelev, Robyn Cairns, J Ross Mitchell, Mark D Cicero, et al. Canadian association of radiologists white paper on artificial intelligence in radiology. *Canadian Association of Radiologists Journal*, 69(2):120–135, 2018.

[55] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[57] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[58] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[59] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[60] Ricardo Cruz, Kelwin Fernandes, Jaime S Cardoso, and Joaquim F Pinto Costa. Tackling class imbalance with ranking. In *2016 International joint conference on neural networks (IJCNN)*, pages 2182–2187. IEEE, 2016.

[61] Masakazu Sato, Koji Horie, Aki Hara, Yuichiro Miyamoto, Kazuko Kurihara, Kensuke Tomio, and Harushige Yokota. Application of deep learning to the classification of images from colposcopy. *Oncology letters*, 15(3):3518–3523, 2018.