A Proof of Concept Imaging System for Automated Cervical Cancer Screening in Peru

Mabel Karel Raza Garcia

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Linda G. Shapiro, Chair

Sherrilynne Fuller

Ira J Kalet

Program Authorized to Offer Degree:
Biomedical and Health Informatics

University of Washington

Abstract

A Proof of Concept Imaging System for Automated Cervical Cancer Screening in Peru

Mabel Karel Raza Garcia

Chair of the Supervisory Committee:
Professor Linda G. Shapiro
Computer Science and Engineering

Cervical cancer is the second most frequent cancer in women around the world and affects half a million women per year. The World Health Organization (WHO) estimates that 275,000 women die every year, and 80% to 85% of these deaths occur in low-resource countries in Africa and South America [18]. In Peru, cervical cancer has the highest incidence and the second highest mortality rates of cancers among women [18]. Currently, the screening techniques such as the Papanicolau (Pap) test, in which some cells from the cervix are examined under a microscope to detect potentially pre-cancerous and cancerous cells [20], and the Visual Inspection with Acetic Acid (VIA), in which the surface layer of the cervix is examined through visual inspection after washing it with 3% to 5% acetic acid (vinegar) for one minute [20], are part of the national health policy in Peru [36]. The Pap test is mainly used in urban areas in Peru. However, there are some challenges related to spreading the Pap test throughout the whole country: lack of quality and standardization of the readings of Pap smears [31], shortage of trained personnel, uneven processing of samples resulting in diagnosis and treatment delays, and lack of even basic laboratory infrastructure, all of which impacts greatly on the sustainability of this procedure in remote and/or poor settings.

Extensive research has shown that computational solutions are a viable and suitable aid for overcoming these barriers [42] [17]. However, the majority of these solutions are commercial products that are not affordable for developing countries, such as Peru. In this context, developing a strategy, algorithm and open source computational implementation that recognizes normal vs. abnormal Pap smears can ultimately provide a cost-effective alternative for developing countries. The dissertation-specific objectives are to: 1) determine the characteristics of normal vs. abnormal Pap smears through expert consultation and relevant literature, 2) collect Pap smear data sets and run preliminary experiments to compare two pattern recognition algorithms in terms of features and classification performance, and 3) assess the accuracy, sensitivity and specificity of the proposed cervical cancer screening approach for classifying normal vs. abnormal Pap smears compared to experts' review.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ACCURACY: statistical measure of how well a binary classification test correctly identifies condition based on the proportion of true results (true positives and true negatives).

CARCINOMA IN SITU (CIS): pre-cancerous cells are confined to the cervix and have not spread to other parts of the body.

CSE SERVER: remote computer server in the Paul G Allen Center for Computer Science and Engineering at the University of Washington.

DATA SET: set of images collected digitally.

HIGH GRADE SQUAMOUS INTRAEPITHELIAL LESION (HSIL): the Besthesda system category to diagnose high cell abnormality in the cervix or with features suspicious for invasive cervical cancer in a Pap test. HSIL includes moderate and severe dysplasia, CIS, CIN 2 and CIN 3.

HUMAN PAPILLOMAVIRUS (HPV): a virus that infects humans in which the high-risk types 16 and 18 can lead to cervical cancer.

HUMAN T-LYMPHOTROPIC VIRUS (HTLV): a human RNA retrovirus in which the type I can cause types of cancer such as t-cell leukemia and lymphoma.

IEEE-XPLORE: digital library for scientific and technical literature by the Institute of Electrical and Electronics Engineers (IEEE).

LOW GRADE SQUAMOUS INTRAEPITHELIAL LESION(LSIL): the Besthesda system category to diagnose low cell abnormality in the cervix in a Pap test. LSIL includes HPV, mild dysplasia, CIN 1.

ODDS RATIO: statistical measure to describe the strength of association between two variables in a logistic regression model.

PAP SMEAR: (Papanicolau test or Pap test) medical procedure in which some cells from the cervix are examined under a microscope to detect potentially pre-cancerous and cancerous cells.

PSEUDO R2: statistical measure of goodness-of-fit in a logistic regression model.

PUBMED: database for biomedical literature from MEDLINE by the US National Library of Medicine - National Institutes of Health.

RGB IMAGE: color image that has three channels red, green and blue.

SCREENING: the systematic application of a medical test in an asymptomatic population.

ACKNOWLEDGMENTS

The author wishes to express profound gratitude to the University of Washington, where she has had the opportunity to be trained, and especially to my advisor Linda Shapiro for her dedication, guidance, knowledge and patience; to the Biomedical and Health Informatics program for their teachings, and especially to Professor Sherrilynne Fuller, Professor Ira Kalet and Professor William Lober; to the Universidad Peruana Cayetano Heredia for their extended long-term collaboration with the University of Washington through the Quipu-Fogarty Program for funding my studies and this project, and especially to Dr. Patricia Garcia, Professor Mirko Zimic, Professor Jesus Castagnetto, Dr. Marcela Lazo and Alicia Alva for their support and contributions to this project; to the Multimedia Group in Computer Science and Engineering, and especially to Dr. Selim Aksoy, Ezgi Mecan, Nicola Dell, Waylon Brunette (Change group) and Dr. Ravensara Travillian; to all collaborators; to Florence Patten for her vast reserve of knowledge in cytopathology; to Dr. Magaly Blas and Dr. Isaac Alva for the first data set from the Ainbo project, and to Dr. Jaime Caceres-Pizarro, Dr. Patricia Arboleda-Ezcurra, Dr. Moises Rojas-Mezarina and Dr. Alejandro Dagnino-Varas for the biggest and second data set, and for its respective authorization to Dr. Aida Palacios-Ramirez and Dr. Jaime Cok-Garcia from the Hospital Nacional Cayetano Heredia. The author also would like to express profound gratitude, respect and love to Master Choa Kok Sui for the constant blessings; to my dear husband Luis for his unconditional love and support; and to my parents and siblings for their love and blessings.

DEDICATION

To Master Choa Kok Sui, for the constant blessings. To my dear husband Luis and my family Raza-Garcia for their unconditional support and love.

Chapter 1

INTRODUCTION

Cervical cancer is the second most frequent cancer in women around the world and affects half a million women per year. The World Health Organization (WHO) estimates that 275,000 women die every year, and 80% to 85% of these deaths occur in low-resource countries in Africa and South America [18]. In Peru, cervical cancer has the highest incidence and the second highest mortality rates of cancers among women [18]. WH0/ICO (last official statistics 2010 [39]) also reports that Peru has 4446 women with diagnosis of cervical cancer, and 2098 cervical cancer deaths per year. In addition, the population of women at risk for developing cervical cancer (ages 15 and over) is 9.51 million in Peru [39]. Screening techniques such as the Papanicolau (Pap) test, Visual Inspection with Acetic Acid (VIA) and the HPV DNA test in adult women are key steps to prevent cervical pre-cancer and cancer [20]. Currently, the Pap test and VIA are part of the national health policy in Peru [36]. The Pap test is a conventional cytology technique in which some of the squamous cells are scraped off the cervix, and then the sample is analyzed in the laboratory [20]. VIA is an alternative screening technique in which the surface layer of the cervix is examined through visual inspection after washing it with 3% to 5% acetic acid (vinegar) for one minute [20].

The conventional Pap test is mainly used in urban areas in Peru. However, there are some challenges related to spreading the Pap test throughout the whole country. One challenge is the lack of quality and standardization of the readings of Pap smears across the country [31]. There is also a shortage of trained personnel such as

pathologists who can read and accurately interpret these smears. Another challenge is the delay on reading and giving results of the Pap smears. Thus, there is loss of follow-ups and delayed clinical treatments of women. Finally, another big challenge is the lack of even basic laboratory infrastructure and sustainability in remote and/or poor settings.

A literature review has shown that computational solutions are a viable and suitable aid for overcoming these barriers [42] [17]. However, the majority of these solutions are commercial products (e.g BD FocalPoint [16], ThinPrep [13], and CHAMP-Dimac [12]) that are not affordable or sustainable for developing countries such as Peru. In addition, the existing commercial providers do not share their algorithms due to patents and software licensing issues. In this context, developing a strategy, algorithm and open source computational implementation that recognizes normal versus abnormal Pap smears can give a unique opportunity to test new current sophisticated computer vision (imaging informatics) techniques; and provide state of the art classifiers for Pap smear images in Peru.

This work presents the proof of concept for an automated cervical cancer screening system in Peru that can ultimately provide a cost-effective alternative for developing countries in terms of cervical cancer screening. This work has the potential to introduce a regular and standard program of Pap smear screening with an increased probability of a suitable follow-up in Peru. This initiative would reduce the incidence and mortality rates of cervical cancer in Peru, and, in addition, could prove useful and generalizable to other developing countries.

Chapter 2 sets the background and significance of cervical cancer screening in Peru and the extensive work done in this research area. Chapter 3 presents the preliminary studies conducted in collaboration with the Bioinformatics Unit at Universidad Peruana Cayetano Heredia (Lima-Peru) and the Department of Computer Science and Engineering at the University of Washington (Seattle-U.S.). Chapter 4 describes the design and methods to achieve the automated cervical cancer screening system

in Peru to classify normal vs. abnormal Pap smears. Chapter 5 presents the various computational experiments and their classification results. Finally, Chapter 6 discusses the conclusions, contributions, limitations, and projected future work in this area in Peru.

Chapter 2

BACKGROUND AND SIGNIFICANCE

2.1 Impact of Cervical Cancer

Cervical cancer is the second most frequent cancer in women worldwide and affects half a million women per year. WHO estimates that 275,000 women die every year, and 80% to 85% of these deaths occur in developing countries [18]. WH0/ICO also estimates that the population of women at risk for developing cervical cancer (ages 15 and over) is 2,337 million around the world [40]. The cause of cervical cancer is related to the high-risk types of human papillomavirus (HPV 16 and 18). The prevalence of HPV infection in Latin America is among the highest in the world [9].

Screening techniques such as the Pap test, VIA and the HPV DNA test in adult women are important steps to prevent cervical pre-cancer and cancer. The Pap test is a conventional cytology technique that has been conducted for over 50 years in women around the world. In this technique, some of the squamous cells are scraped off the cervix and then the sample is examined under the microscope in the laboratory [20]. Health professionals pay attention to a special area called the transformation zone (where the flat and columnar cells meet), because it is very vulnerable to attack by HPV viruses. VIA is an alternative screening technique in which the surface layer of the cervix is examined through visual inspection after washing it with 3% to 5% acetic acid (vinegar) for one minute [20]. The HPV DNA test is also a screening technique that uses a small swab to collect some samples from the cervix, and then the DNA test is conducted [20].

2.2 Cervical Cancer in Peru

Cervical cancer in Peru has the highest incidence and the second highest mortality rates of cancers among women [18]. The estimated incidence rate for cervical cancer is 34.5 per 100,000 and the estimated mortality rate is 16.3 per 100,000 Peruvian women per year [18]. WH0/ICO (last official statistics 2010 [39]) also reports that Peru has 4446 women with diagnosis of cervical cancer, and 2098 cervical cancer deaths per year. In addition, the population of women at risk for developing cervical cancer (ages 15 and over) is 9.51 million in Peru [39]. Thus, cervical cancer is still an important public health issue declared as a national priority in Peru [36].

In Peru, the Pap test and VIA are included in the "Manual of Standards and Procedures for the Prevention of Cervical Cancer" as two of the main standard screening techniques [36] [31]. According to the Program for Appropriate Technology in Health (PATH), an international non-profit organization for global health, the conventional Pap test is mainly used in urban areas in this country [51]; and the liquid-based cytology (LBC) test is only used in one or two laboratories in the capital [1]. However, there are some challenges related to spreading the Pap test across Peru.

One challenge is the lack of quality and standardization of the readings of Pap smears throughout the whole country [31]. A study in the Peruvian Amazonia reported that the Pap test had a low sensitivity of 42.54% and specificity of 98.68% for detecting carcinoma *in situ* or cervical cancer in a population of 5,435 women [1]. Another challenge is the shortage of trained personnel such as pathologists who can read and accurately interpret these smears. The study in the Peruvian Amazonia also reported that the trained personnel would require more regular training and supervision to conduct the screening techniques in this rural region [1].

Another main challenge is related to the lack of processing samples in a timely manner, and sub-standard quality control procedures [31]. To illustrate this, the public health research group at the Universidad Peruana Cayetano Heredia in Lima-

Peru collected data showing that the time delay for women to receive their Pap results is around 3 to 9 months at one health network of the Ministry of Health in Callao-Peru (described in Appendix A) [45]. Finally, one more key challenge is the lack of even basic laboratory infrastructure and sustainability in remote and/or poor settings. In addition, to my knowledge, there is not currently any automated cervical cancer screening system in Peru.

2.3 Related Work

2.3.1 Current commercial solutions

A literature review has shown that computational solutions are a viable and suitable aid for overcoming these challenges [42] [17]. For example, BD Diagnostics, which acquired the TriPath Imaging company, is a global cancer diagnostics company in the U.S. that offers products for cervical cytology screening with FDA approval. BD Diagnostics provides the BD FocalPoint (formerly known as the AutoPap system) as automated imaging system [16]. The major components of the software are: a) image segmentation, b) global detection, c) local detection, d) cell discrimination and e) slide classification [27]. HOLOGIC, a women's healthcare company in the U.S., offers the ThinPrep Imaging System that performs around 70% of the Pap tests in the U.S. This system also has FDA approval [13]. The main stages of ThinPrep are: a) automated slide screening with identification of 22 fields of view, b) automated labeling of suspicious or abnormal cells (no further review for normal cells), c) cytotechnologist's slide review, and d) automated slide labeling (dots) for pathologist's review [15]. Dimac, a digital image company in Denmark - Europe, provides cervical cancer diagnosis through the Cytology and Histology Analysis Modular Package (CHAMP system) [12]. Its patented algorithm has seven components: a) image acquisition, b) color standardization and segmentation, c) nucleus identification and delimitation, d) classification, e) cell measurement, f) slide grading and g) slide categorization [29].

These commercial systems focus on the primary screening phase in which the definitive normal slides are dropped and the abnormal slides are given to the experts for further analysis and diagnostics confirmation. Unfortunately, these solutions are commercial products that are not affordable nor sustainable for developing countries such as Peru. In addition, the existing commercial providers do not share their algorithms due to patents and software licensing issues.

2.3.2 Current academic solutions

Extensive research has shown that there have been many efforts to address cervical cancer screening and diagnosis using computer-based solutions. Wied *et al.* [53] in 1978 proposed a kind of computer recognition of ectocervical cells using 200 different features of density, texture and shape under different spatial resolutions (microscope magnifications). Bartels *et al.* [4] in 1981 continued the work on computer recognition of ectocervical cells and determined the most significant features for discrimination between different cell types from the ectocervix. Lee *et al.* [26] in 1992 presented a multilayer processing strategy for automated Pap smear screening using 68 features of size, shape, density and texture for classifier training. Then, in 1997 they reproduced the complete cell features provided by the well-known cytopathology expert Stanley Patten that lead to the rule-based algorithmic cell classification approach of the AutoPap system (now BD FocalPoint) [28]. Jantzen *et al.* [24] in 2005 provided Pap smear benchmark data for comparing classification methods; they tested 20 features and classified the cells into seven classes. Schilling *et al.* [50] in 2007 used texture analysis, contour grouping and pattern recognition techniques to detect and classify cervical cells using phase-contrast microscopy; they extracted 80 features and chose the best 20 features for classification. Mat-Isa *et al.* [32] in 2008 proposed an automated diagnostic system using automatic feature extraction and an intelligent diagnostic using 4 features. Kale [25] in 2010 proposed an unsupervised screening system to rank cervical cell images using 14 distinct cells features. Plissiti *et al.* [46]

in 2011 presented a fully automated method for cell nuclei detection in Pap smear images using morphological reconstruction and clustering; they obtained a sensitivity of 90.75% and a specificity of 75.28% with the fuzzy C-mean (FCM) classification technique. Finally, Bergmeir *et al.* [5] in 2012 implemented a system that handles full resolution images and proposed a new segmentation algorithm for nuclei using a voting scheme and prior knowledge.

2.3.3 Current open-source and accesible solutions

To my knowledge, there are few computational solutions for cervical cancer screening and other diseases that use microscopic images, which are readily accessible and open-source [5]. Below, I describe one authorized copyrighted pattern recognition algorithm for Tuberculosis and one open-source pattern recognition algorithm for cervical cancer that I tested in my preliminary studies. I decided to evaluate these two algorithms, because both show promise and are accessible and open-source solutions.

The TB/MODS pattern recognition algorithm

The TB/MODS pattern recognition algorithm was developed in the Bioinformatics Unit at the Universidad Peruana Cayetano Heredia. Alicia Alva working under the supervision of Dr. Mirko Zimic programmed the algorithm using the programming language C with a Fourier library [10]. The goal was to automatically diagnose Mycobacterium tuberculosis in a MODS culture using geometrical and illumination features. This algorithm has shown a sensitivity of 99.1% and a specificity of 99.7% for diagnosing tuberculosis compared to the expert's review in the MODS assay [2]. Figure 2.1 shows the main steps of the algorithm to process MODS images. Each step is described below:

Image Processing: The RGB image is converted to grayscale. Then, a global contrast filter is applied. The next step is to apply a global binarization using the

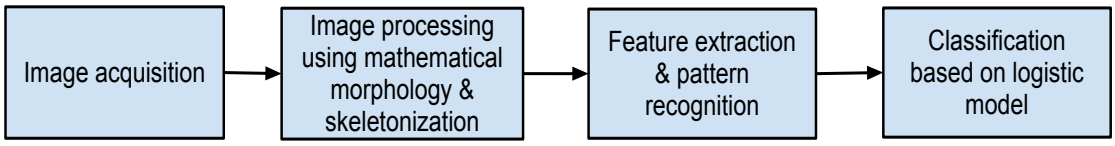


Figure 2.1: Flowchart of the TB/MODS pattern recognition algorithm

Otsu binary threshold-selection algorithm [41] to reduce the grayscale image to a binary image calculating a threshold value from all the pixels of the image. The following steps include application of a median filter using a 3x3 window size and labeling the object numbers. Then, boundary objects were removed to prevent bias. Later, some mathematical morphology operations, such as dilation and erosion, were performed to reduce noise. Hole filtering was also applied to remove the white regions inside the binarized objects, and area filtering was done to remove the extremes of area [2]. Then, the skeletonization process applied an image foresting transform that is a graph-based approach to shrink the image and obtain the skeletons of the objects with their main topological features. Finally, image recoloring is conducted to facilitate further analysis [2].

Feature extraction and pattern recognition: In total, 54 features were used to construct the feature descriptor for the TB/MODS classification. The full 54 features are listed and described in Appendix B. The four single major features obtained were: a) the skeleton - identifies the trunk and its branches, and the border points of the object; b) object thickness - quantifies the average thickness based on the cross-sections and measurement of the thickness, light refraction and circularity; c) overall form of the object - average deviation of the objects shape to a straight line (linearity of the skeleton) and; d) lateral curvature - this feature was calculated using a Fourier transform [2].

Classification based on logistic model: Two models were built using training and testing sets: one called the object model and the other called the photo-model. In the object model, 54 features were used to perform a univariate analysis using simple logistic regression. Then, features with the highest Pseudo R2 and odds ratios were considered and the highly correlated features were dropped using a multiple regression analysis that obtained the best prediction object model. This model identifies Mycobacterium tuberculosis with 96.81% sensitivity and 96.32% specificity. The area under the ROC curve was 0.988 [2]. In the photo-model, each object model was applied to each photo where the best-eight objects were selected. This photo-model was built with the same statistical methodology as the object model using multiple logistic regressions in a step-backward approximation (removing the least significant features). All these models were analyzed using the Stata software [52]. The best photo-model identifies Mycobacterium tuberculosis with 99.1% sensitivity and 99.7% specificity. The area under the ROC curve was 0.999 [2].

The UW/CSE pattern recognition algorithm

The UW/CSE pattern recognition algorithm was developed in the Department of Computer Science and Engineering (CSE) at the University of Washington. Nicola Dell and Waylon Brunette, CSE PhD students, programmed the algorithm using the programming language C++ with the OpenCV library [8]. The goal was to automatically screen Pap images. This algorithm had a classification accuracy of 88.2% in single Pap images using the k-means clustering algorithm with features related to k-means data, area, shape and color of nucleus and cytoplasm, and a classification accuracy of 81% in whole Pap images using all the previous features plus texture and extropy features [14].

In this project, I was a close collaborator providing the Ainbo data set (200 Pap images) described in Chapter 3, and the key features that recognize normal and abnormal cells in a Pap smear according to the experts. Figure 2.2 shows the flowchart

of the UW/CSE pattern recognition algorithm to process Pap images. Each step is described below [14]:

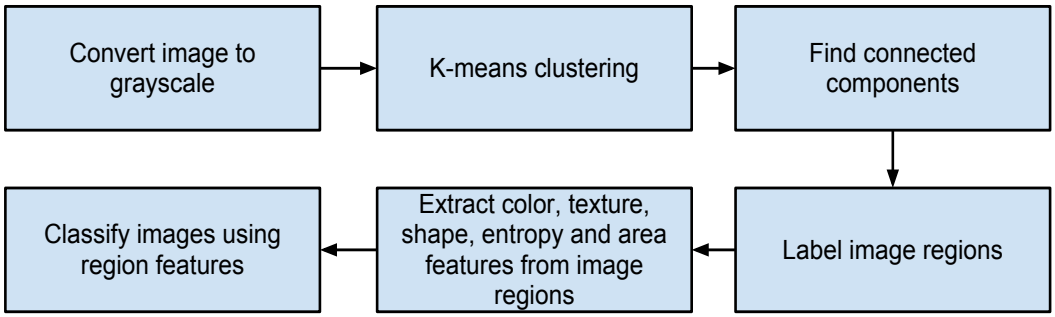


Figure 2.2: Flowchart of the UW/CSE pattern recognition algorithm [14]

K-means clustering: First, the Pap image was converted to normalized gray scale. Then, a k-means clustering with initial random seeds was run on the normalized gray scale image. Values of K less than 10 were used to cluster different cell regions [14].

Connected-components and labeling image regions: Next, a connected components algorithm was used to identify the main regions of the Pap image. The next step was to label each component of the image as nucleus, cytoplasm or background [14].

Feature extraction: Then, 5 different types of features - color, texture, shape, entropy and area - were extracted from the regions. In total, the feature descriptor was built using 12 features: a) k-means cluster information for the whole image, b) LBP histogram information from nucleus components, c) LBP histogram information from cytoplasm components, d) average pixel intensity of nucleus components, e) average pixel intensity of cytoplasm components, f) average roundness of nucleus components,

g) average roundness of cyploplasm components, h) entropy of nucleus components, i) entropy of cyploplasm components, j) total area of nucleus components, k) total area of cytoplasm components, and l) ratio of total area of nucleus components to total area of cytoplasm components [14].

Classification using machine learning algorithms: Finally, several different classifiers from Weka [21], an open source collection of machine learning algorithms, were trained to classify the regions with the default parameters. The best accuracy result to recognize single normal vs. abnormal Pap image was 88.2%, achieved using the Random Forest classifier [14]. The best accuracy result to recognize whole normal vs. abnormal Pap image was 81%, achieved using the Multilayer Perceptron classifier [14].

Chapter 3

PRELIMINARY STUDIES

The objectives in the preliminary studies were to 1) describe the list of key characteristics of cervical cells according to experts consultation and relevant literature; and 2) collect Pap smear data sets and run preliminary experiments to compare two pattern recognition algorithms in terms of features and classification performance as stated in Objective 1 and Objective 2 of this dissertation respectively. The following is a detailed description on how these preliminary studies were conducted.

3.1 Characteristics of Cervical Cells

The knowledge of the characteristics of cervical cells is the first key step to establish the pattern recognition of cervical cancer screening according to the experts. To accomplish this goal, I had weekly meetings dedicated exclusively to analysis the cytopathology of Pap smears images with a senior expert cytotechnologist at the University of Washington, and then contacted expert pathologists from the Department of Clinical Pathology and Pathological Anatomy at the Hospital National Cayetano Heredia in Lima-Peru [22].

In a study period of two quarters, Florence Patten, the senior expert cytotechnologist, emphasized three main aspects: a) Pap smears are like fingerprints due to the unique and wide range of diversity of these samples. This assertion means that the proposed algorithm should be able to handle different types of Pap smear images with a huge training data set to be implemented in a real-setting in Peru. b) Background is not clear in abnormal Pap smears. The only exception to this case is the presence of another type of invasive cancer from some other part of the body that shows a clear

background but with cancer cells in a Pap smear. The proposed algorithm should be able to screen the whole Pap smear image. c) Then nucleus is the most important part for screening a Pap smear. The nucleus shows high chromosome activity in abnormal Pap smears. This study of the pattern recognition of normal and abnormal Pap smears images was conducted using the Bethesda System website atlas [34]. The Bethesda System is the U.S. standard diagnosis system for cervical cancer to report Pap smear results [34].

In addition to study the key medical criteria employed to diagnose the Pap smear, I contacted pathologists Dr. Aida Palacios-Ramirez, Dr. Jaime Cok-Garcia and Dr. Jaime Caceres-Pizarro from the Department of Clinical Pathology and Pathological Anatomy at the Hospital Nacional Cayetano Heredia in Lima-Peru [22] to obtain a real-setting Pap smear data set from Peru. Hospital Nacional Cayetano Heredia is a governmental hospital, which like all of them, follows the U. S. Bethesda standard system for cervical cancer diagnosis. This research collaboration was possible through the long-standing collaboration between the University of Washington and Universidad Peruana Cayetano Heredia with its partnerships in Lima-Peru through the QUIPU-Fogarty program [23].

As a result of expert's consultation, I identified normal (Table 3.1) and abnormal (Table 3.2) cervical cell characteristics to establish the cervical cancer pattern recognition according to the experts. The Pap smear images are from the Ainbo and Cayetano Heredia data sets - described in the next Section 3.2.

To complement the knowledge of pattern recognition of cervical cells, I performed a thorough literature review employed PubMed [37] for the biomedical aspects of the problem, and IEEE Xplore [35] for the engineering and computer science aspects of this research. The overarching goal was to know the key features employed by cervical cancer pattern recognition algorithms, and the state of the art in identifying important characteristics of cervical cells in Pap smears. A summary of the results of the key features used in pattern recognition algorithms from the literature review is shown

Table 3.1: List of characteristics of normal cervical cells

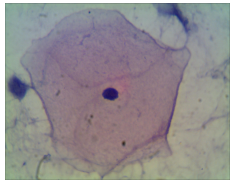
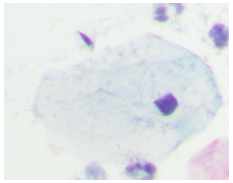
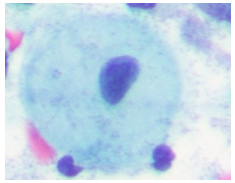
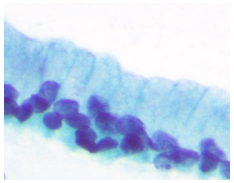
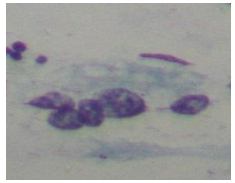
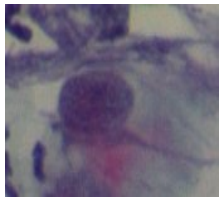
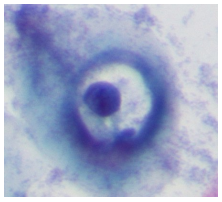
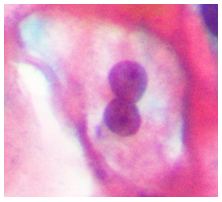
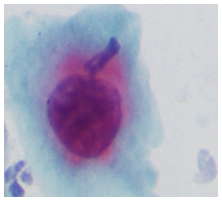
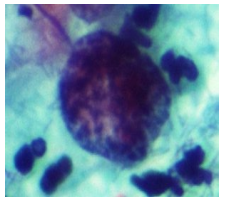
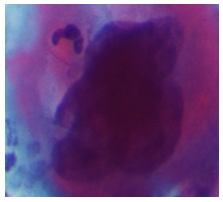
	Squamous cells		**Metaplastic cells**	**Glandular cells**	
Characteristics	**Superficial**	**Intermediate**	**Metaplastic**	**Columnar**	**Endometrial**
Pap image (section 3.2)					
Shape	polygonal/oval/flat	polygonal	round/oval	cylindrical/columnar	cuboidal/round
Nucleus shape	very small-solid	vesiclar	round	round-oval	round
Nucleus area	25 sq um	35 sq um	50 sq um	40 sq um	35 sq um
N/C ratio	very large	stable	small	highly variable	n/a
Cytoplasm shape	polygonal	polygonal	round/oval	columnar	cuboidal
Cytoplasm area	1500 sq um	1500 sq um	n/a	n/a	n/a
Texture	thin-transparent	thin-transparent	thick-dense	no homogeneous	very thin
Cell border	well-defined	well-defined	well-defined	well-defined	well-defined
Cell arragement	isolated/sheets	isolated/sheets	isolated/sheets	isolated/sheets	isolated/cluster
Object analogy	“dinner plate”	“dinner plate”	“fried egg”	“honeycomb”	“grape cluster”
Additional note	presence of estrogen	presence of progesterone	inmature squamous proliferation	showing specimen adequacy	seen prior to / during period
Location	extocervix	extocervix	transformation zone	endocervix	endocervix

Table 3.2: List of characteristics of abnormal cervical cells

		LSIL		HSIL		
Characteristics	**ASCUS**	**LSIL**	**LSIL + HPV**	**HSIL**	**CIS**	**Carcinoma**
Pap image (Section 3.2)						
Nucleus shape	twice normal size	large & light	single nucleus or multinuclei	large & dark	large & dark & deformed	very dark & variable
N/C ratio	altered	altered	altered	very altered	very altered	very altered
Cytoplasm shape	polygonal	polygonal	polygonal	round/oval	round	variable
Cytoplasm area	mimics normal	mimics normal	mimics normal	reduced size	very limited	variable
Texture	mimics normal	koilocytotic change	koilocytotic change	dense	dense & sparse	highly variable
Chromatin particles	fine/oval granular	fine granular	fine granular	grossly granular	fine/coarse granular	very granular & big nucleoli
Chromatin distribution	even	even	even	uneven	uneven	irregular

in Table 3.3 and Table 3.4. We see that researchers tried to reduce the number of significant features across time, and to move from cell segmentation to cytoplasm and nucleus segmentation. Plissiti *et al.* [46] in 2011 were able to automatically identify cell nuclei in conventional Pap smears. The majority of features in this literature review were related to shape and illumination.

3.2 Data Sets

The experiments presented in this work are illustrated using two different data sets from Peru:

3.2.1 The Ainbo data set

The Ainbo data set consists of 200 images of Pap smears (100 normal, 100 abnormal (50 LSIL, 50 HSIL)). The Pap smear slides were collected by the AINBO project, which evaluates the association between HPV and HTLV in women in the jungle of Peru. This study aimed to define if this association can increase the risk of developing cervical cancer in this population [7]. The project leaders, Dr. Magaly Blas and Dr. Isaac Alva, provided 341 anonymous Pap smear slides with the Bethesda ground truth classification done by a pathologist. I digitalized a subset of these slides (200 images) using a Zeiss microscope with 100x objective and 10x ocular (1000x magnification), maximum microscope illumination, and a 3MPx digital camera. All the facilities were provided by Dr. Mirko Zimic in the Bioinformatics Unit at Universidad Peruana Cayetano Heredia in Peru. Figure 3.1 shows different normal and abnormal Pap smears images from this data set.

3.2.2 The Cayetano Heredia data set

The Cayetano Heredia data set was collected by the Department of Clinical Pathology and Pathological Anatomy at the Hospital Nacional Cayetano Heredia [22], a governmental hospital, using the normal cytology preparation technique. The pathologist

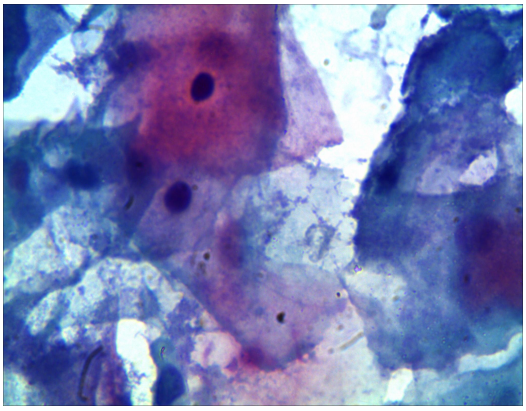
Table 3.3: Summary list of features in the literature review

Year	Author	No. Features	Features
1978	Wied *et al.* [53]	200	Global extinction, cell and nucleus histograms, cell contour size and regularity, cell concavities, 3rd highest extinct, N/C ratio, Fourier ellipse, cell and nucleus area, transitions (the complete list was not published).
1981	Bartels *et al.* [4]	6	CYBEST system: nuclear staining, chromatin, cytoplasmic area, nuclear boundary, nuclear shape and number of nuclei.
		10	TICAS system: nuclear area/cell area ratio, the average difference of optical density (OD) values between adjacent points in the nucleus, the values of the cell OD frequency histogram, the standard deviation of OD values for all points in the nucleus, the values of the nuclear OD, the area of the nucleus in sq *um*, the length of the shorter semiaxis of the nucleus contour fitting oval and normalized ratio.
		10	TUDAB Project: N/C ratio, total optical density in the nucleus, cell area, cell perimeter, area of chromatin in nucleus, nuclear perimeter, variance of OD values in the nucleus, variance of OD values in cytoplasm, nuclear area and nuclear perimeter squared/nuclear area.
1997	Lee *et al.* (AutoPap) [27]	68	Nuclear area, cell area, cell shape, relative nuclear area, nuclear hypechromasia, chromatin particles, chromatin distribution, micronucleoli, macronucleoli, cytoplasmatic texture, cell borders and cell arrangement (the details of all the 68 features was not published).
2005	Jantzen *et al.* [24]	20	Nucleus and cytoplasm area, N/C ratio, nucleus and cytoplasm brightness, nucleus and cytoplasm shortest diameter, nucleus and cytoplasm longest diameter (continued below).

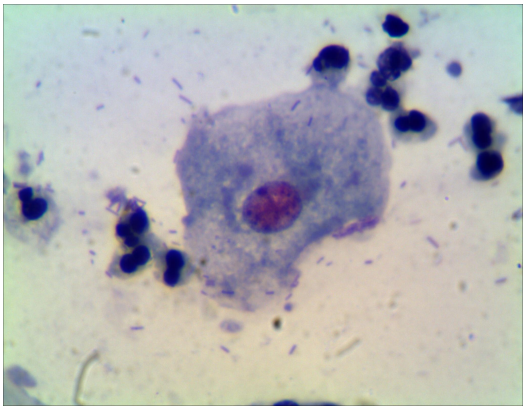
Table 3.4: Summary list of features in the literature review - continued

Year	Author	No. Features	Features
2005	Jantzen *et al.* [24]	20	Nucleus and cytoplasm enlongation, nucleus and cytoplasm roundness, nucleus and cytoplasm perimeter, nucleus position, maxima in nucleus, maxima and minima in nucleus and cytoplasm.
2007	Schilling *et al.* [50]	80	Length of the semi-major axis of E, avg. amplitude on the boundary of E, avg. angle between edge orientation and corresponding E orientation, error of fit between E and edge segments, aspect ratio of E, avg. level and variance of intensity inside E (E - ellipse) (the details of all the 80 features was not published).
2008	Mat-Isa *et al.* [32]	4	Nucleus size, nucleus grey level, cytoplasm size and cytoplasm grey level.
2010	Kale's thesis [25]	14	Nucleus and cytoplasm area, N/C ratio, nucleus and cytoplasm brigthness, nucleus longest diameter, nucleus shortest diameter, nucleus enlongation, nucleus roundness, nucleus perimeter, nucleus maxima/minima and cytoplasm maxima/minima.
2011	Plissiti *et al.* [47]	16	Entropy B in green, perimeter, foreground-background contrast in red, std histogram LBP hyperbola in green, circularity, foreground-background contrast in green, mean histogram LBP circle in blue, entropy B in red, mean histogram LBP hyperbola in blue, smoothness B in green, std histogram LBP circle in red, entropy A in green, foreground-background contrast in blue, std histogram LBP hyperbola in red, smoothness A in red and 3rd moment of neighborhood in blue (A - boundary area and B - bounding box of A).

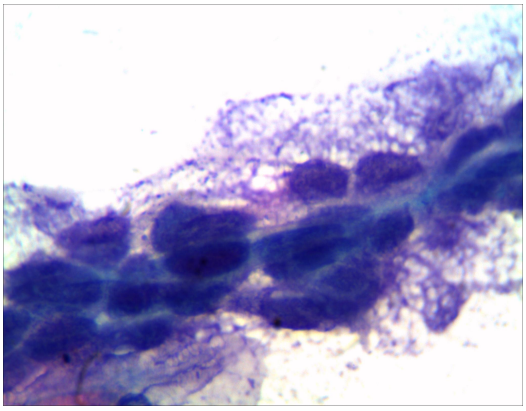
(a) Normal Image (single cells)



(b) Normal Image (overlapping cells)



(c) Abnormal image (LSIL)



(d) Abnormal image (HSIL)

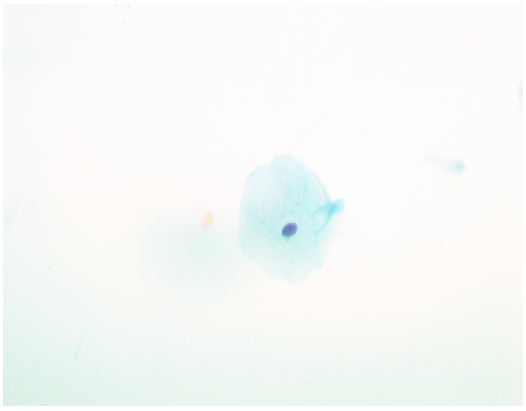
Figure 3.1: The Ainbo data set (1000x magnification)

Dr. Jaime Caceres-Pizarro and his senior medical residents Dr. Patricia Arboleda-Ezcurra, Dr. Moises Rojas-Mezarina and Dr. Alejandro Dagnino-Varas provided 966 anonymous images (502 normal, 464 abnormal (250 LSIL, 214 HSIL)) with their respective Bethesda ground truth classification. They also directly marked the most representative cells on the Pap images. These digital images were taken at the hospital with a Nikon Microscope E3 with 40x objective and 10x ocular (400x magnification), maximum microscope illumination, and a DS-Fi1 digital camera. This data set is more realistic due to the challenges associated with poor contrast and overlapping cells, which are common conditions on conventional Pap smears from health clinics throughout the country. Figure 3.2 shows Pap smears images from the Cayetano Heredia data set.

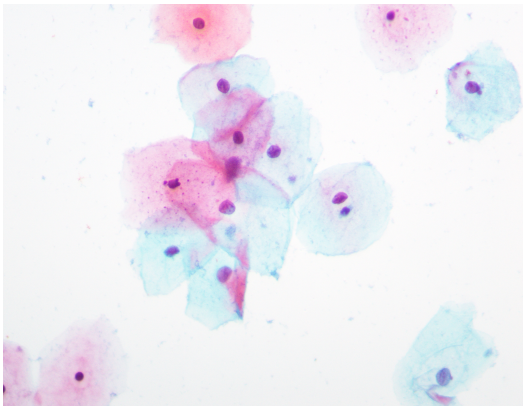
In summary, we can see that both data sets have different magnifications and laboratory staining. The image magnification can make a huge difference in the image pre-processing and processing. Thus, the expert pathologists recommended images of 400x magnification, which as it happens is one widely used in Peru for diagnostic purposes. The label accompanying these data sets was provided in different formats, which required consolidation and term normalization. The Ainbo data set was given with the ground truth of the anonymous slides in a spreadsheet file, whereas the Cayetano Heredia data set was provided with the ground truth of the images and the most representative cells labeled on the same JPG file.

3.3 Experiment A: Adapting the TB/MODS Pattern Recognition Algorithm

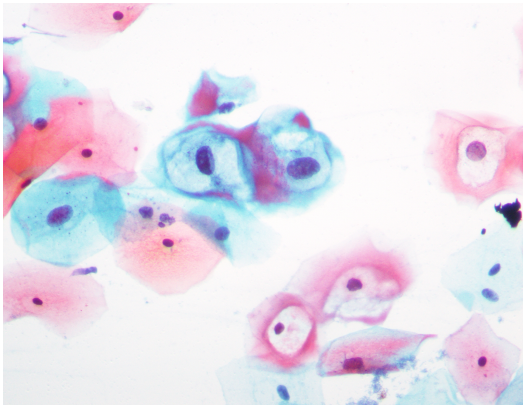
Experiment A aimed at adapting the TB/MODS pattern recognition algorithm (sen. 99.1% and spe. 99.7% for tuberculosis diagnosis compared with the expert's review in the MODS assay [2]), to recognize normal vs. abnormal cell nuclei in the Pap images compared to the expert pathologists' ground truth. The latest studies mentioned that the segmentation of cell nuclei is the most important, because nuclei show significant



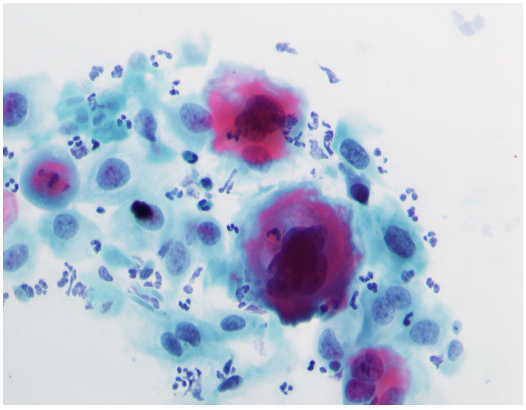
(a) Normal Image (single cell)



(b) Normal Image (overlapping cells)



(c) Abnormal image (LSIL)



(d) Abnormal image (HSIL)

Figure 3.2: The Cayetano Heredia data set (400x magnification)

changes when disease is present [46]. Thus, I start adapting the TB/MODS pattern recognition algorithm to identify nuclei, for which a small subset of the Cayetano Heredia data set (276 images - 227 normal and 49 LSIL) for image processing was used comprised by 70 cell nuclei (35 normal and 35 LSIL) for training classification. The abnormal images were just from low-abnormality (LSIL) according to the ground truth of the experts. The main steps of the adapted algorithm are described in detail below.

3.3.1 Image Pre-processing

This process was the key part to provide suitable filters, parameters and methods for the image processing. I tested different parameters to be able to adjust the image processing of the TB/MODS pattern recognition algorithm to recognize nuclei of the Pap smears images. I used a small subset of the Cayetano Heredia data set. I selected the seven most representative Pap smear images from the normal category. The flowchart on Figure 3.3 shows the main steps of the process:

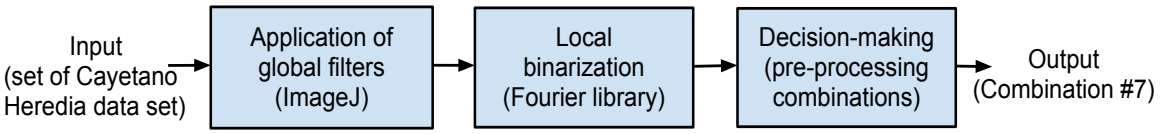


Figure 3.3: Flowchart of the Pap image pre-processing

Application of global filters using ImageJ

First, the RGB Pap image was uploaded by ImageJ, an open source image processing software [48], that allowed testing different filters and combinations. I tried distinct filters such as Gaussian blur for smoothing the image, enhanced contrast for strengthening the contrast, and median filter with different window sizes for reducing

the image noise. The exploration with ImageJ was very helpful, because it provided great insights about the key filters and parameters that could handle the Pap smear images.

Local binarization using the Fourier Library

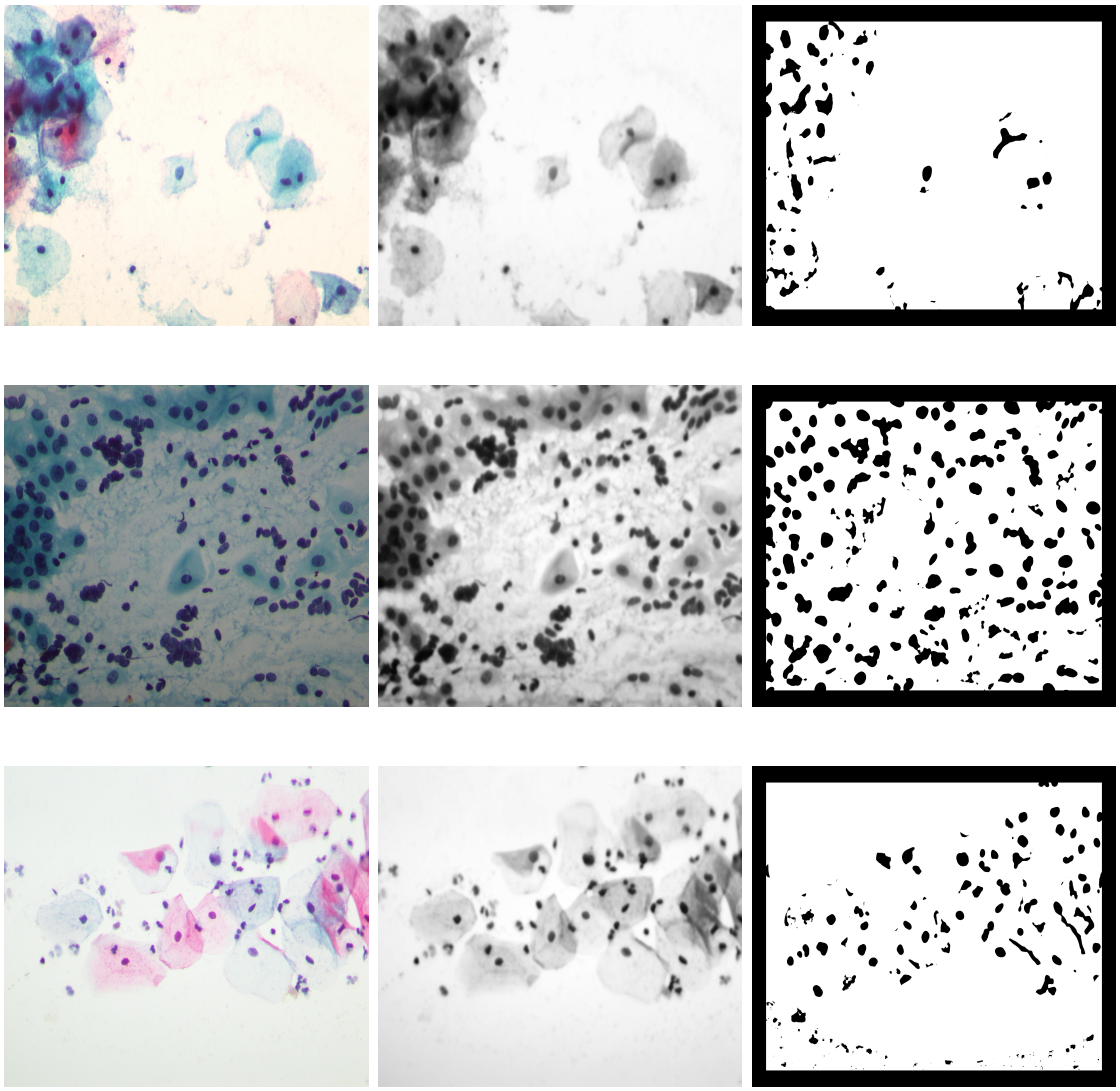
Then, I tested three local binarization methods using an open source library in C called Fourier lib [10]. In this type of images, local binarization is better than global binarization due to an uneven distribution of the illumination of the Pap smear images. The local threshold methods used were: a) Bernsen, which is based on comparing if the local threshold is below the contrast threshold [6], b) Niblack, which is based on the calculation of the local mean and standard deviation [33], and Sauvola, which is based on a variation of the Niblack method [49].

Decision-making

Finally, I had different combinations of filters, parameters and local binarization methods. After analizing the pre-processed images, the combination that better recognized nuclei on the Pap smears images was selected. Table 3.5 shows the most important combinations. The best combination at recognizing nuclei on Pap smears was # 7. This combination consisted of first applying the Gaussian blur filter. Next, an enhanced contrast filter was run to produce a strong contrast. The next step was to convert the RGB image to an 8-bit gray scale image. Then, a median filter with a 3x3 window size was applied. At the end, a local binarization using the local threshold method of Niblack [33] was run. Figure 3.4 shows the best pre-processing result from combination #7.

Table 3.5: Best combinations of Pap image pre-processing

		Filters [parameter]				
Combo	**Normal Image**	**1st Filter**	**2nd Filter**	**3rd Filter**	**4th Filter**	**Local Binarization**
1	1003, 1006, 1009	gray scale [16 bits]	Gaussian blur [6]	median [4]	n/a	Bernsen, Niblack, Sauvola
2	1003, 1006, 1009	Gaussian blur [4]	gray scale [8 bits]	median [2]	n/a	Bernsen, Niblack, Sauvola
3	1003, 1006, 1009	Gaussian blur [8]	gray scale [16 bits]	median [4]	n/a	Bernsen, Niblack, Sauvola
4	1003, 1006, 1009	Gaussian blur [10]	Enhanced contrast [0.2]	gray scale [8 bits]	median [10]	Bernsen, Niblack, Sauvola
5	1003, 1006, 1009	Gaussian blur [4]	Enhanced contrast [0.4]	gray scale [8 bits]	median [8]	Bernsen, Niblack, Sauvola
6	1003, 1006, 1009	Gaussian blur [2]	Enhanced contrast [0.4]	gray scale [8 bits]	median [10]	Bernsen, Niblack, Sauvola
7	1003, 1006, 1009	Gaussian blur [6]	Enhanced contrast [0.2]	gray scale [8 bits]	median [3]	Bernsen, Niblack, Sauvola
8	1003, 1006, 1009	Gaussian blur [8]	Enhanced contrast [0.1]	gray scale [8 bits]	median [6]	Bernsen, Niblack, Sauvola



(a) Normal Pap images (b) Gray-scale Pap images (c) Binary Pap images

Figure 3.4: Results of the best Pap image pre-processing (combination # 7)

3.3.2 Image Processing

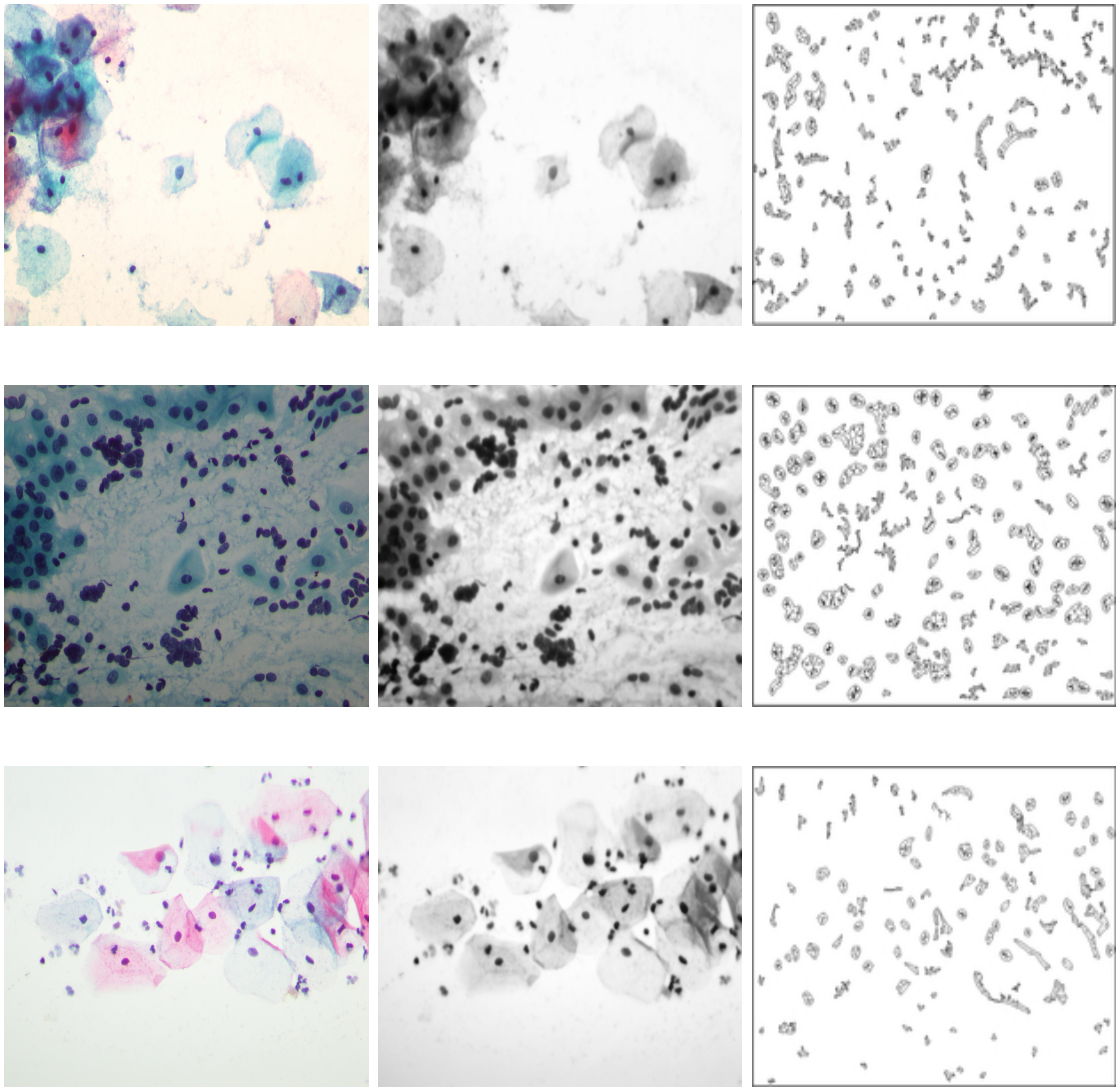
The best combination of Pap image pre-processing was included in the TB/MODS pattern recognition algorithm. Alicia Alva, first author of the TB/MODS pattern recognition algorithm and research assistant at the Bioinformatics Unit, adapted and ran the TB/MODS pattern recognition algorithm with the parameters of the best combination. Figure 3.5 presents the results of the image processing of some normal Pap images. The output of the image processing was: a) the features of each object (spreadsheet files), b) the object images (jpg files), and c) the original, gray-scale and edge-skeleton of the Pap images (jpg files) as shown on Figure 3.6.

3.3.3 Manual Labeling and Mapping

The expert pathologists provided the Bethesda ground truth image classification and directly marked the most representative cells on the same Pap images. I did a manual labeling of cell nuclei in the previously processed Pap images according to the experts. I labeled 227 normal Pap images (highlighted green) and 49 abnormal LSIL Pap images (highlighted blue). Figure 3.7a shows an example of labeling normal cell nuclei and Figure 3.7b of labeling abnormal cell nucleus in a processed Pap image. Later, I did a manual mapping of each nucleus with its object number by looking for the object in the output folder "object images", and reading its file name, which was named by the image name plus the object number automatically assigned by the algorithm (e.g."1006a-object-0057.jpg"). See Figure 3.7c and Figure 3.7d.

3.3.4 Feature Extraction

The feature extraction step aimed to provide the 54 features for each cell nuclei. A CSV (comma-separated values) file was created using a data set of 70 cell nuclei (35 normal and 35 abnormal (LSIL)). Then, different scripts automatically joined the normal and abnormal object files, generated after the image processing, to one single



(a) Normal Pap image (b) Gray-scale Pap image (c) Edge-skeleton Pap image

Figure 3.5: Results of the image processing of Pap smears

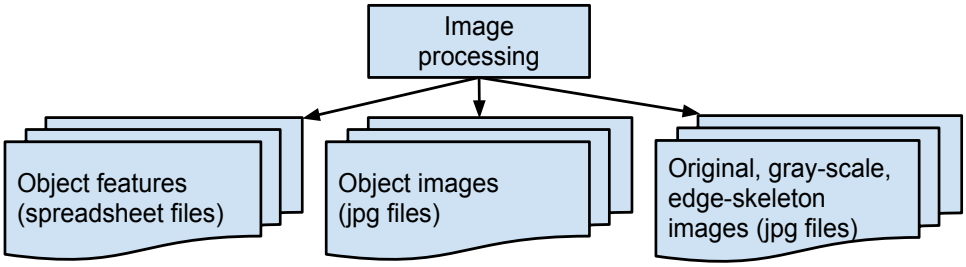
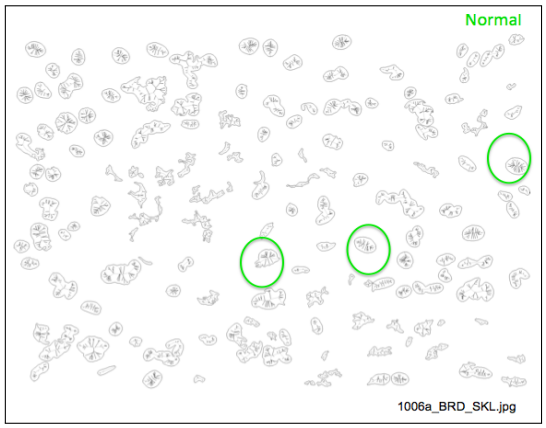
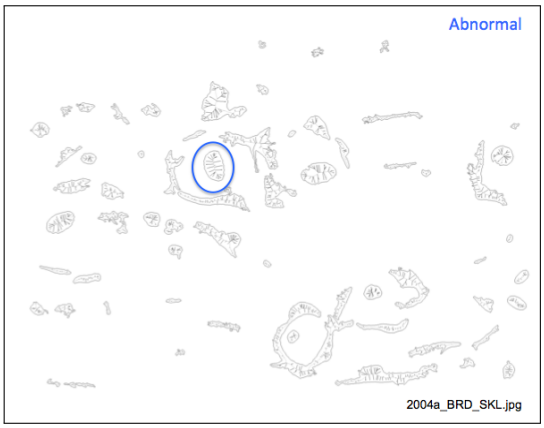


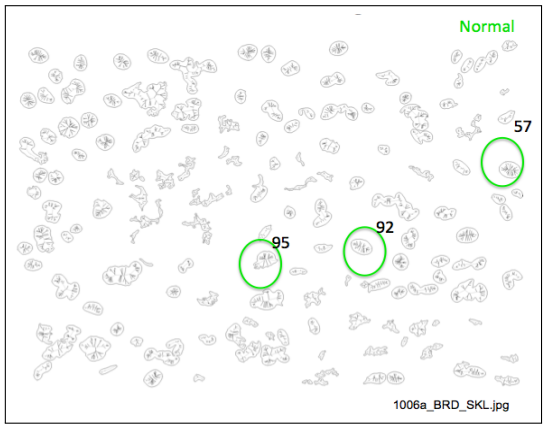
Figure 3.6: Output files of the image processing of Pap smears



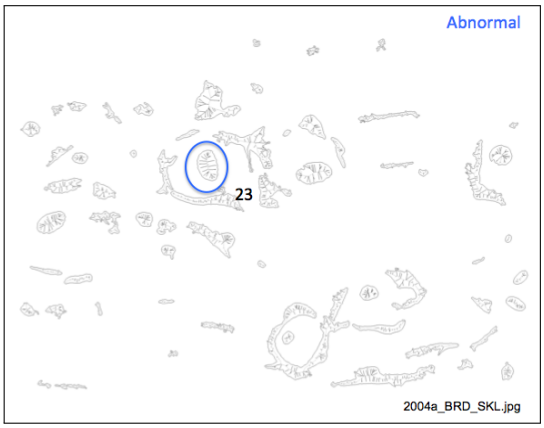
(a) Manual labeling of normal nuclei



(b) Manual labeling of abnormal nuclei



(c) Manual mapping of normal nucleus



(d) Manual mapping of abnormal nucleus

Figure 3.7: Manual labeling and mapping in processed Pap images

spreadsheet file with all 54 features of each nucleus. The 54 features were used as a feature descriptor for classification.

3.3.5 Classification Results

I built four object models in this preliminary study using simple and multiple logistic regression. First, the 54 features were used to perform a univariate analysis. Then, features with the highest pseudo R2 and odds ratio were included, and the highly correlated features were removed, keeping the most significant predictors. Filtered feature predictors were tested using multiple logistic regressions in a step-backward approximation (removing the least biologically meaningful features). All these models were analyzed using the Stata software [52]. See Figure 3.8 for the flowchart of this classification process and Table 3.6 for the classification results. We can see that the fourth object model identified Pap normal nuclei vs. abnormal nuclei with the highest sensitivity of 97.14% and the highest specificity of 97.14% compared to the expert pathologists ground truth. The area under the ROC curve, which illustrated the performance of a binary classifier, of the best model was 0.9935 as shown in Figure 3.9.

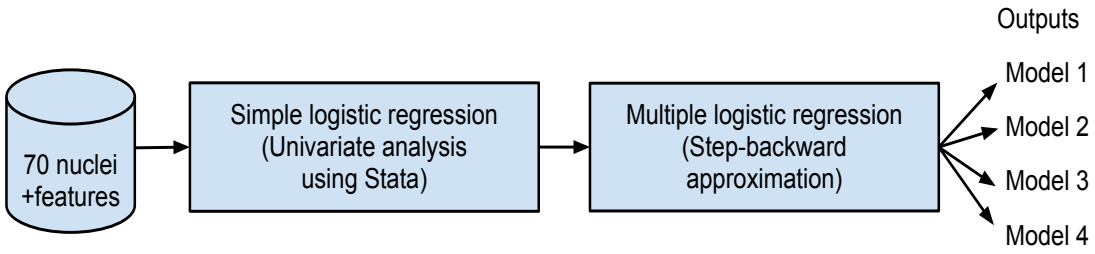


Figure 3.8: Experiment A: Flowchart of the classification process

According to the statistical analysis, the six features that best classified normal vs. abnormal cell nuclei are shown in Table 3.7.

Table 3.6: Experiment A: Classification models

Model	No. Features	Sensitivity (%)	Specificity (%)
1	4	74.28	88.57
2	3	82.85	80.00
3	2	74.28	82.85
4	6	97.14	97.14

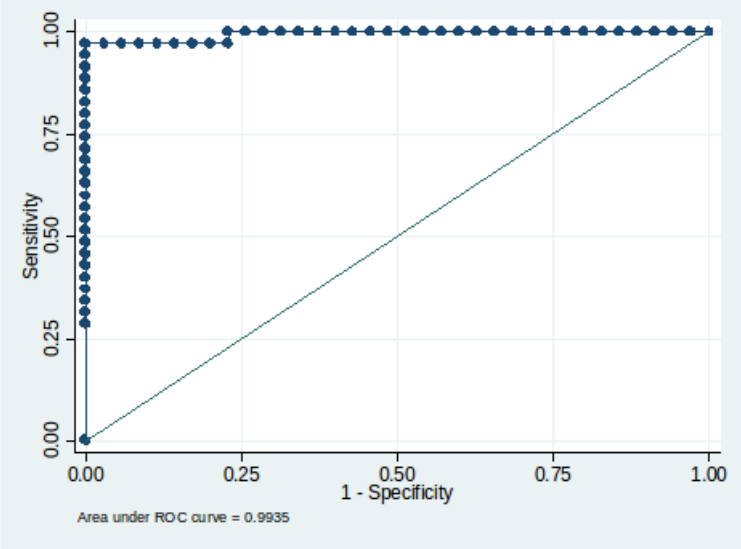


Figure 3.9: Experiment A: The ROC curve of the best classfication model

3.4 Experiment B: Testing the UW/CSE Pattern Recognition Algorithm

Experiment B was to test the UW/CSE pattern recognition algorithm [14] with the Cayetano Heredia data set. In this case, I followed the same methodology to identify nucleus, cytoplasm and background. I start testing with a subset of the Ainbo data set (35 single images 18 normal and 17 abnormal) to get familiar with algorithm; and

Table 3.7: Experiment A: Features of the best classification model

Features			
English Name	**Spanish Name**	**Type**	**Description**
Avg of shape deviation	prom for	geometric	average of the shape's deviation of the object to a straight line
Perimeter	perimetro	geometric	perimeter of the object
Light refraction 22	bire22	illumination	brightness compared to the mean brightness of the image at 22%
Light refraction 8 in blue	bire8_b	illumination	brightness compared to the mean brightness of the image at 8% in the blue mask
Mean image-illumination	mediafoto	global	mean of the whole image-illumination
Std dev image-illumination	dsfoto	global	standard deviation of the whole image-illumination

then I used a subset of the Cayetano Heredia data set (35 whole images - 18 normal and 17 abnormal) for training and testing classification respectively. The main steps of the testing are described in detail below.

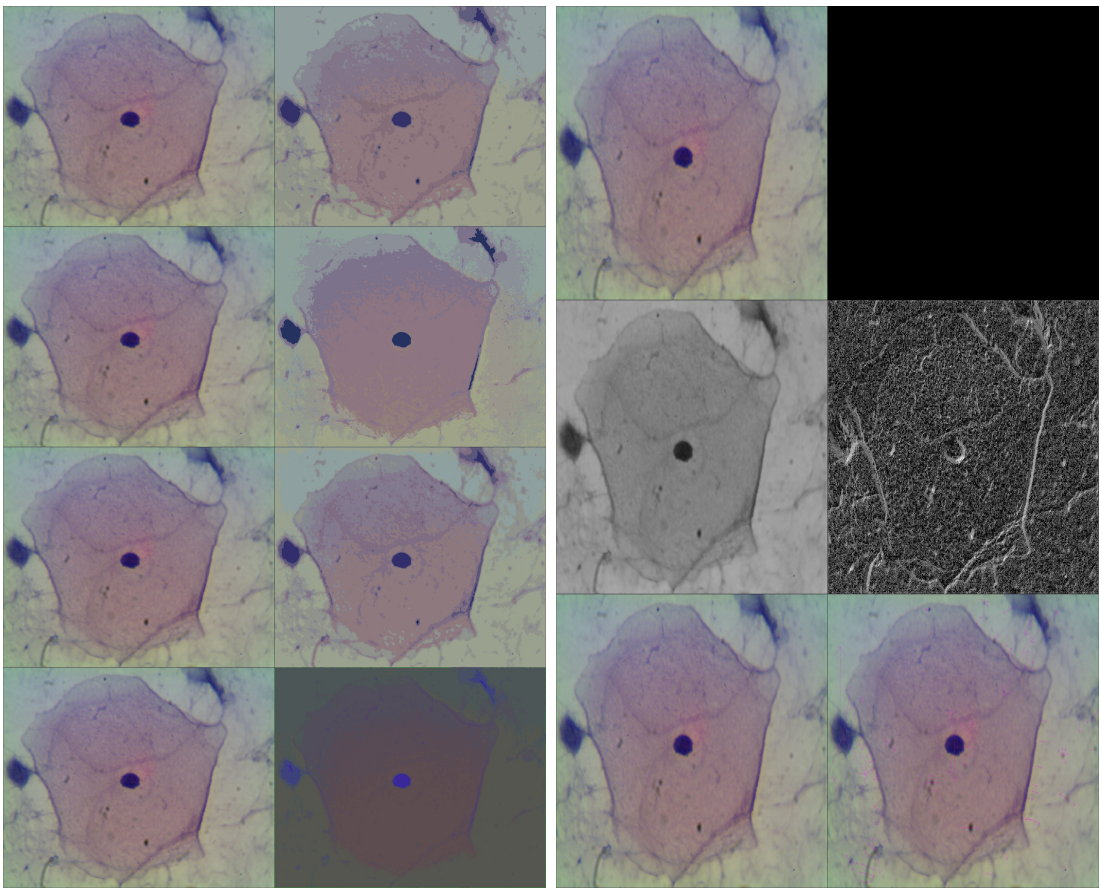
3.4.1 K-means clustering

First, I tested the k-means algorithm with different features on the Ainbo data set in a remote connection with the CSE server. The UW/CSE pattern recognition algorithm has different parameters that use the k-means algorithm. For example, the parameter [A] applies the k-means algorithm with initial random seeds with different color threshold as shown in Figure 3.10a; and the parameter [W] applies the k-means algorithm with the texture feature - local binary pattern (LBP) [38] and the scale-invariant feature transform (SIFT) [30] using the ground truth images as shown in Figure 3.10b.

Second, I ran different scripts with the Cayetano Heredia data set that tested different k sizes (6 to 10). I only included 35 images due to time processing constraints in the remote server at UW/CSE. It took almost two full days to obtain results with this data set.

3.4.2 Connected-components and labeling image regions

The ouputs of segmentation and region identification from the CSE/UW pattern recognition algorithm are shown in Figure 3.11. The original Pap image is in the first column. Cell identification and labeling is shown in the second column. The third one shows the image mask for nucleus feature extraction, and the fourth column shows the image mask for the cytoplasm regions. Figure 3.12 shows light cytoplasm incorrectly classified as background. Nucleus is also incorrectly classified as cytoplasm.



(a) Applying k-means with different color threshold

(b) Applying k-means with LBP and SIFT using ground truth images

Figure 3.10: Experiment B: Examples of testing the k-means algorithm

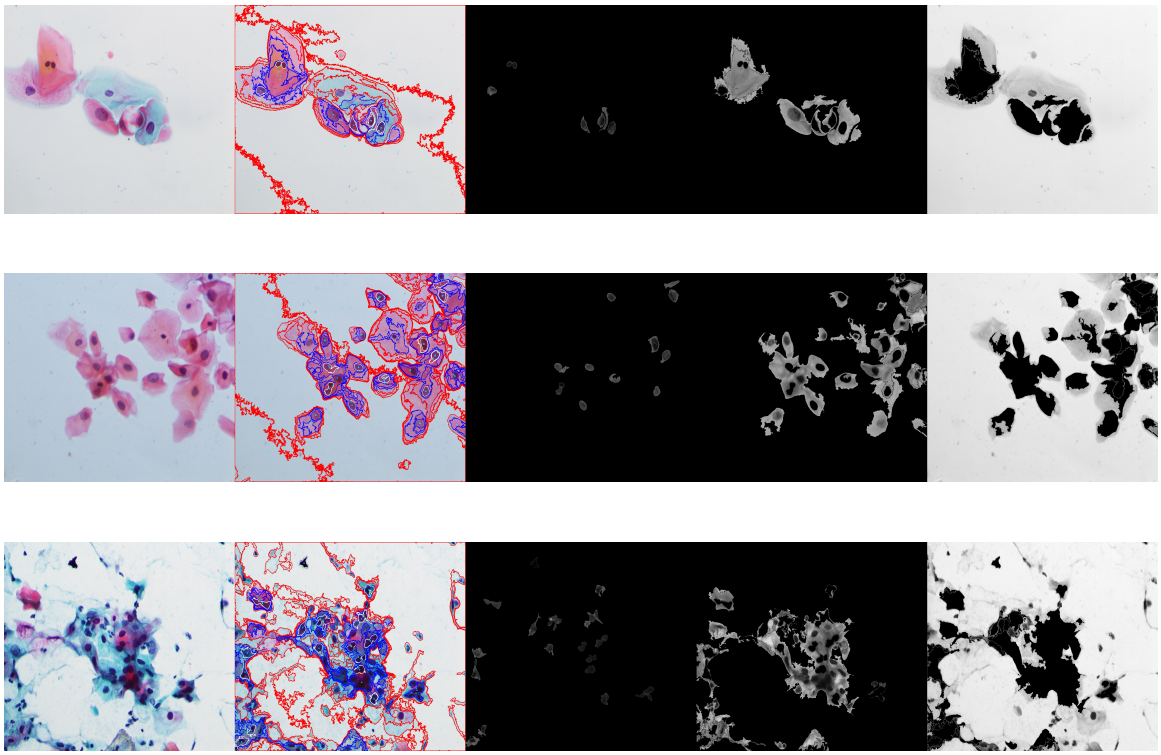


Figure 3.11: Experiment B: Segmentation results showing quite successfull region identification. Original Pap images in the 1st column; cell identification and labeling - nucleus (white), cytoplasm (blue), and background (red) in the 2nd column; image mask showing nuclei in the 3rd column; image mask showing cytoplasm in the 4rd column; and gray-scale Pap images in the 5th column.

(a) Nuclei were incorrectly labeled as cytoplasm as shown in the 2nd column

(b) Any nucleus was segmented as shown in the 3rd column, but some cytoplasms were segmented as shown in the 4th column

(c) Light cytoplasm was incorrectly labeled as background as shown in the 2nd column

Figure 3.12: Experiment B: Segmentation results showing some errors

3.4.3 Classification Results

Finally, I ran the classification of normal vs. abnormal Pap images using WEKA [21]. The classifiers were trained with the Cayetano Heredia data set. The training set consisted of 18 whole Pap images, and the testing set consisted of 17 whole Pap images. Then, I ran ten different machine-learning algorithms with the default parameters with different sets of features. The accuracy classification results with k equal to 7 are shown in Table 3.8. The best result is the set of features D (entropy, color, area, shape and ratio) that recognized normal vs. abnormal Pap images with an accuracy of 82.3% by the Random Tree classifier.

3.5 Comparison of the Two Pattern Recognition Algorithms

Both pattern recognition algorithms have advantages and limitations. The adapted TB/MODS algorithm was semi-automatic, and identified just nucleus. The UW/CSE algorithm was full-automatic, and identified nucleus, cytoplasm and background. Table 3.9 compares the overall characteristics of the adapted TB/MODS pattern recognition algorithm and the UW/CSE pattern recognition algorithm.

In terms of features comparison, the TB/MODS pattern recognition algorithm used six features to classify normal vs. abnormal cell nuclei; and the UW/CSE pattern recognition algorithm used ten features to classify Pap smear images. The best features of the adapted TB/MODS algorithm were features related to illumination and geometric features; and the best features of UW/CSE algorithm were related to shape, illumination, color and texture. A notable difference was that the TB/MODS only identified features for nuclei, and the UW/CSE identified features for nuclei, cytoplasm and background. Furthermore, in the TB/MODS approach, the nuclei were hand-selected for training while in the UW/CSE approach, the algorithm was fully automated. Table 3.10 shows the features comparison of two methods in detail.

In terms of algorithms performance, the adapted TB/MODS algorithm shows a

Table 3.8: Experiment B: Classification results

Set of features A (all features); set of features B (kmeans color and count, entropy, area, shape and n/c ratio); set of features C (color, area, shape and n/c ratio); set of features D (entropy, color, area, shape and n/c ratio); and set of features E (kmeans color and count, area, shape and n/c ratio).

	Accuracy (%)														
	Set of Features A (color threshold)			**Set of Features B (color threshold)**			**Set of Features C (color threshold)**			**Set of Features D (color threshold)**			**Set of Features E (color threshold)**		
Classifiers	**107**	**110**	**113**	**107**	**110**	**113**	**107**	**110**	**113**	**107**	**110**	**113**	**107**	**110**	**113**
Naive Bayes	64.7	64.7	64.7	64.7	52.9	52.9	58.8	52.9	52.9	64.7	52.9	52.9	64.7	52.9	52.9
Bayes Net	64.7	58.8	58.8	70.5	58.8	58.8	76.4	64.7	64.7	70.5	58.8	58.8	76.4	64.7	64.7
Logistic	70.5	70.5	70.5	58.8	58.8	58.8	41.1	35.2	35.2	70.5	70.5	70.5	58.8	47.0	47.0
Multilayer Perceptron	47.0	58.8	58.8	58.8	58.8	58.8	64.7	58.8	58.8	64.7	70.5	70.5	58.8	58.8	58.8
SMO	58.8	58.8	58.8	52.9	52.9	52.9	47.0	58.8	58.8	58.8	70.5	70.5	52.9	47.0	47.0
Classification Via Clustering	41.1	47.0	47.0	41.1	47.0	47.0	52.9	58.8	58.8	64.7	64.7	64.7	64.7	47.0	47.0
Threshold Selector	64.7	58.8	58.8	52.9	58.8	58.8	47.0	47.0	47.0	64.7	70.5	70.5	52.9	52.9	52.9
Random Forest	70.5	64.7	64.7	58.8	52.9	52.9	70.5	76.4	76.4	64.7	76.4	76.4	64.7	76.4	76.4
Random Tree	52.9	52.9	52.9	64.7	64.7	64.7	47.0	47.0	47.0	70.5	82.3	82.3	47.0	52.9	52.9
Decision Table	64.7	67.7	64.7	64.7	64.7	64.7	64.7	64.7	64.7	64.7	64.7	64.7	64.7	64.7	64.7

Table 3.9: Comparison of the two pattern recognition algorithms

Characteristics	**Adapted TB/MODS algorithm**	**UW/CSE algorithm**
Mode	semi-automatic	full-automatic
Data set	70 cell nuclei	35 Pap images
Region identification	nucleus	nucleus, cytoplasm and background
Classification mode	training (70 nuclei)	training (17 images) and testing (18 images)
Classifiers	logistic regression (Stata)	10 machine learning algorithms (Weka)
Performance	97.14% sensitivy, 97.14% specificity, and 97.14% accuracy	87.50% sensitivy, 77.78% specificity, and 82.35% accuracy
Programming language	C	C++
Code documentation	partially-documented	well-documented
License	authorized copyright	open source
Year	release 2012	release 2011

Table 3.10: Feature comparison of the two pattern recognition algorithms

Feature	**Adapted TB/MODS algorithm**	**UW/CSE algorithm**
Type	shape and illumination	shape, illumination, color and texture
Descriptor	54 features	12 features
Significant features	6 features	10 features
Name	avg of shape deviation, perimeter, light refraction at 22%, light refraction at 8% in blue, mean image-illumination and std dev image-illumination	color, entropy, area, shape and n/c ratio

sensitivity of 97.14%, specificity of 97.14% and accuracy of 97.14%, whereas, the UW/CSE shows a sensitivity of 87.50%, specificity of 77.78%, an accuracy of 82.35%. Table 3.11 shows each a lgorithm‘s performance in terms of accuracy classification, sensitivity (sen), specificity (spe), true positive (TP), false positive (FP), true negative (TN) and false negative (FN). From this table, we can see that the adapted TB/MODS algorithm had better performance classifying normal vs. abnormal nuclei. However, the UW/CSE algortihm classified whole normal vs. abnormal images (nucleus, cytoplasm and background).

In summary, both methods have advantages and limitations. It is important to note that these experiments were not run under the same conditions, because the data sets were of different sizes, the region identifications were dissimilar, and the classification methods were distinct in methodology. In general, both methods showed that they can process and classify Pap smear images from the Cayetano Heredia data set.

Table 3.11: Algorithms performance

				Confusion Matrix	
Algorithm	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**TN** **FN**	**FP** **TP**
Adapted TB/MODS	97.14	97.14	97.14	34/70	1/70
				1/70	34/70
UW/CSE	82.35	87.50	77.78	7/17	2/17
				1/17	7/17

Chapter 4
DESIGN AND METHODS

Objective 3 of this dissertation was to assess the accuracy, sensitivity and specificity of the proposed cervical cancer screening approach for classifying normal vs. abnormal Pap smears compared to experts' review. This chapter describes the design of the proposed cervical cancer screeening system based on the two algorithms described in the preliminary studies in Chapter 3.

The proposed strategy for the cervical cancer screening algorithm was to use the best components of both methods, and the overall goal was to fully automate the adapted TB/MODS pattern recognition algorithm for cervical cancer screening adding texture features and classification methods from the UW/CSE pattern recognition algorithm.

4.1 System Design

The proposed new combined algorithm is based on a tree-structured approach as shown in Figure 4.1. The left branch is an adapted version from the TB/MODS algorithm to cervical cancer pattern recognition (described in the image pre-proccessing in Chapter 3 and the classification models in Chapter 5). The right branch analyses the whole image using the Local Binary Pattern (LBP) texture feature from the UW/CSE algorithm.

At the top level of the tree on the left branch, after an image processing step which segments the image into separate objects (as discussed in Chapter 3), most Pap smear images are sent to the left side, and some inadequate images are sent to the right side. The inadequacy of some images was due to illumination changes during the image

acquisition by the expert pathologists. Then, the remaining adequate images were sent to the feature extraction process. The algorithm extracted 54 features from each segmented object. Later, these objects were sent to the object classification model to be classified as nucleus or non-nucleus (artifacts). After that, the algorithm only focused on the nucleus objects to classify which ones were normal and which ones were abnormal nuclei. The final image classifier on the left branch uses the number of abnormal objects, the highest probabilities of the abnormal objects, and the mean of the highest probabilities of the abnormal objects according to each of the four object models (described in Chapter 5. This image classifier also includes some global geometric and illumination features. All these features were used as a feature descriptor to classify the image as normal or abnormal using the logistic regression method in Stata [52].

The right branch of the tree shown in Figure 4.1 was perfomed independently as it is part of a collection of computer vision techniques of the UW/CSE algorithm. This approach applies the LBP texture operator to batch the whole image. There is no segmentation step, and no images are rejected. The LBP operator produces a 256-bin histogram representing the texture of the image [38]. The values of this histogram are the feature descriptor that feed into a classifier on the bottom right. This classifier used only the LBP texture features classified by different machine learning algorithms in Weka [21].

Both the whole image classifier on the left, which uses the number of abnormal objects, the highest probabilities of the abnormal objects, and the mean of the highest probabilities of the abnormal objects according to each of the four object models plus global geometric and illumination features; and the whole image classifier on the right, which uses only the LBP texture features are able to identify the most significant features used in their best classifications with the simple logistic regression classifier in Weka [21] and the step-backward approximation (removing the least biological meaningful features). The final step in our cervical cancer screening algorithm is

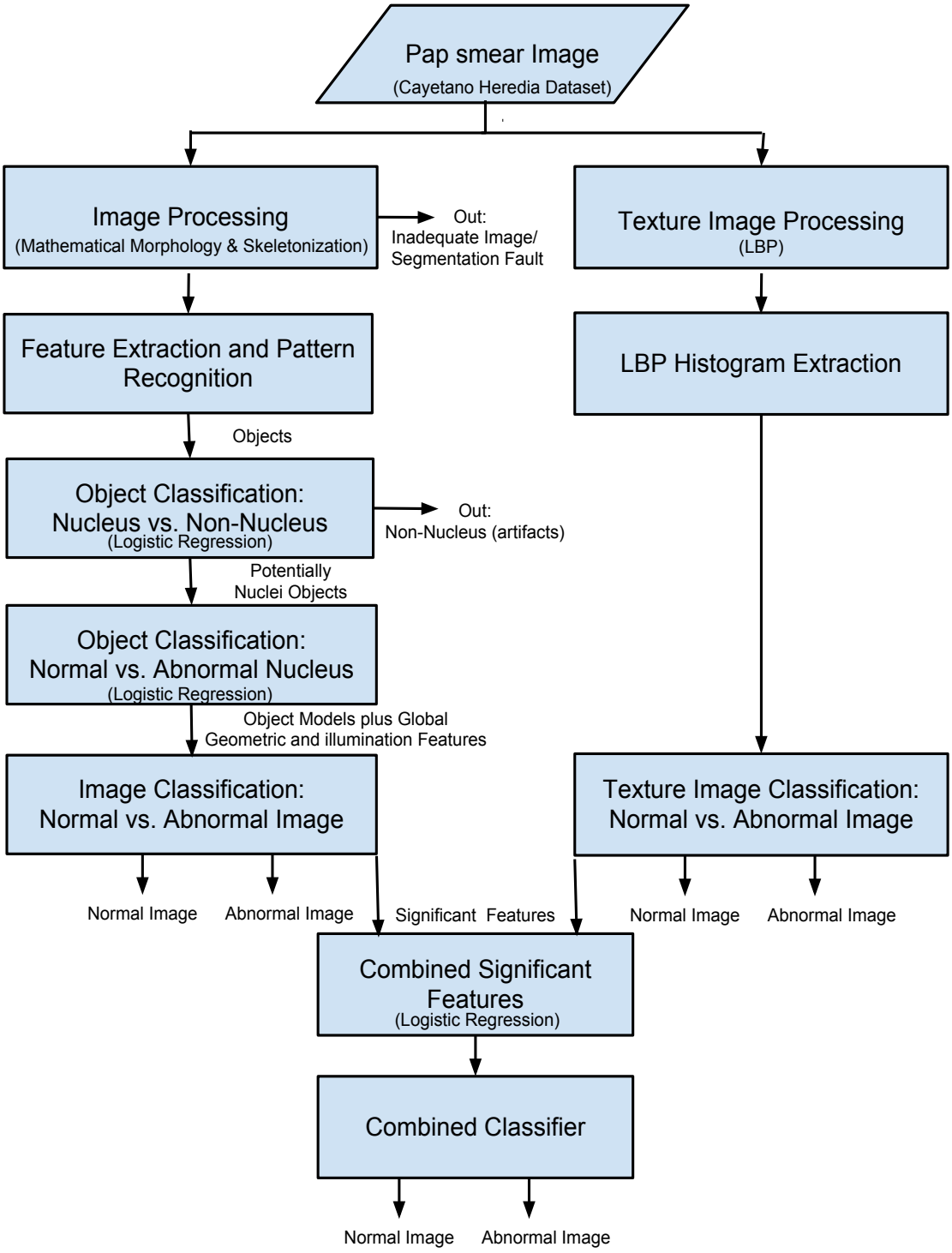


Figure 4.1: The proposed cervical cancer screening algorithm

to combine the most significant features from each classifier into a single feature descriptor, and train a final classifier with all these significant features.

4.2 Data Sets

On the left branch of the tree, I used different sizes of data sets for the object and image classification, due to the availability of the data provided by the expert pathologists from the Hospital Nacional Cayetano Heredia. First, I used 534 objects from 162 images in the object classification of nucleus vs. non-nucleus. Then, I used 267 objects (half of the set) from the same 162 images in the object classification of normal vs. abnormal nucleus. Finally, I enlarged the data set to have 378 images. After removing some images due to inadequacy for illumination changes, the final data set was 213 images for the classification of normal vs. abnormal image (bottom left of the flowchart). On the right branch of the tree, I used the same data set of 213 images for the texture image processing and texture image classification.

4.2.1 Data Set of Objects: Nucleus and Non-Nucleus

The initial data set contained objects from 288 images from a subset of the whole Cayetano Heredia dataset, but 126 images were removed due to the objects selected. I hand-selected 534 objects (267 nucleus and 267 non-nucleus) from the remaining 162 images of the Cayetano Heredia dataset according to the marked cells by the expert pathologists. The 267 nucleus belong to three categories: 151 normal nucleus of 98 normal images, 41 LSIL nucleus of 29 LSIL images, and 75 HSIL nucleus of 35 HSIL images. The 267 non-nucleus were artifacts such as polys (white blood cells) and background artifacts (see Figure 4.2). These 534 objects were used for algorithm training to classify nucleus vs. non-nucleus.

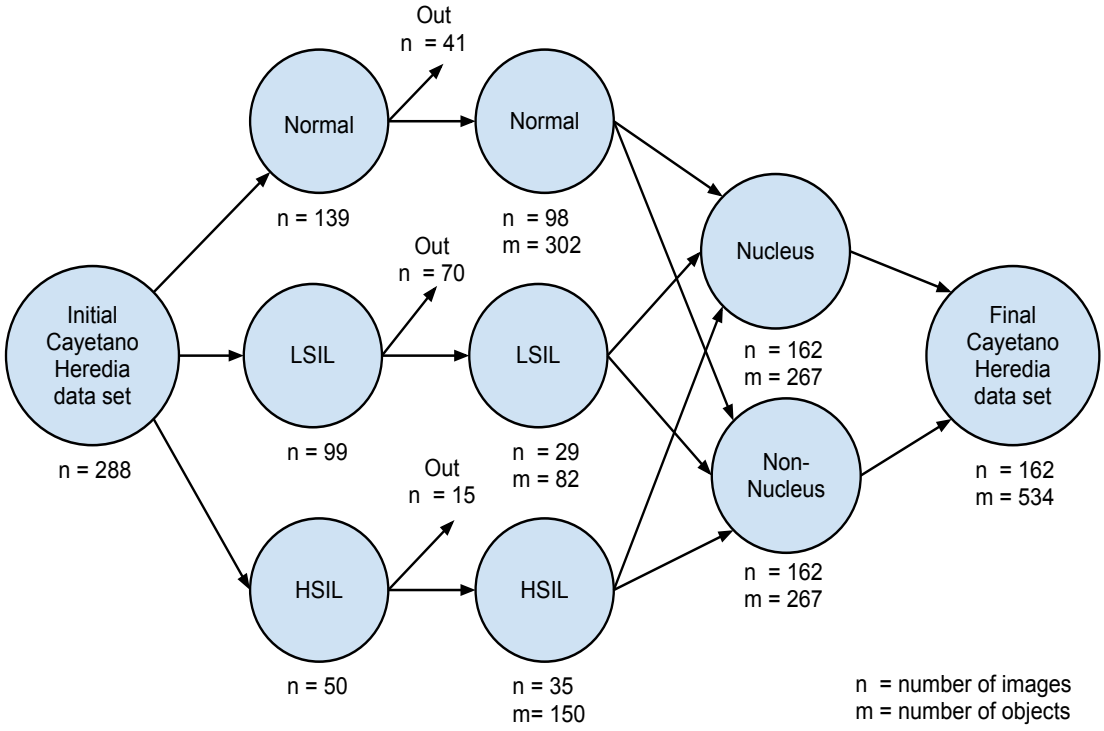


Figure 4.2: Data set of objects: nucleus and non-nucleus

4.2.2 Data set of Objects: Normal and Abnormal Nucleus

I selected half of the data set of objects of nucleus and non-nucleus. This data set contained the 267 nuclei from the same 162 images (151 normal nuclei, 41 LSIL nuclei, and 75 HSIL nuclei) (see Figure 4.3). These 267 nucleus were used for training in the classification of normal vs. abnormal nucleus.

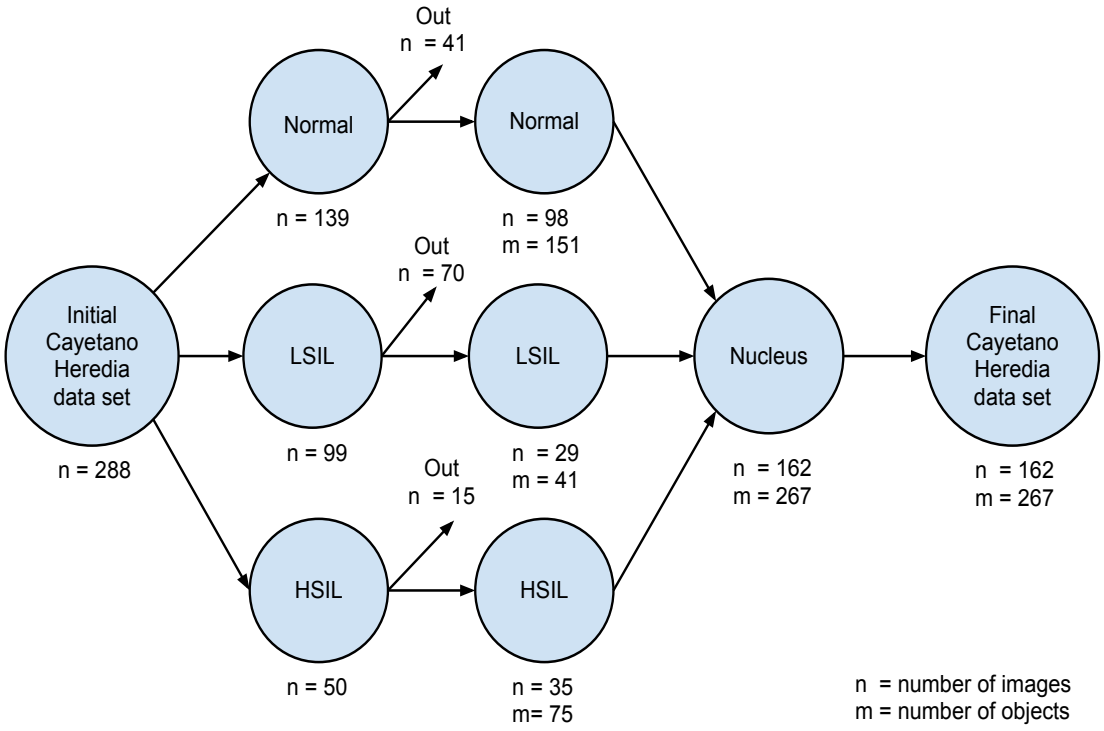


Figure 4.3: Data set of objects: normal and abnormal nucleus

4.2.3 Data set of Images for the Adapted TB/MODS Algorithm

I hand-selected a subset of 378 images (165 normal, 52 LSIL and 161 HSIL) of the whole Cayetano Heredia dataset (966 images). These images were used for algorithm training, except that 165 images were removed due to the inability of the

adapted TB/MODS algorithm to process them due to illumination changes and inadequate/indeterminate images. The object model classification described in Chapter 5 of the adapted TB/MODS algorithm are based on some illumination features as features heritage of the Tuberculosis diagnosis using MODS in which light refraction is part of the key pattern recognition for the experts. In addition, the Cayetano Heredia data set contains some inadequate images with too many blood cells.

Only 213 images (89 normal, 32 LSIL and 92 HSIL) passed to the classification phase (see Figure 4.4).

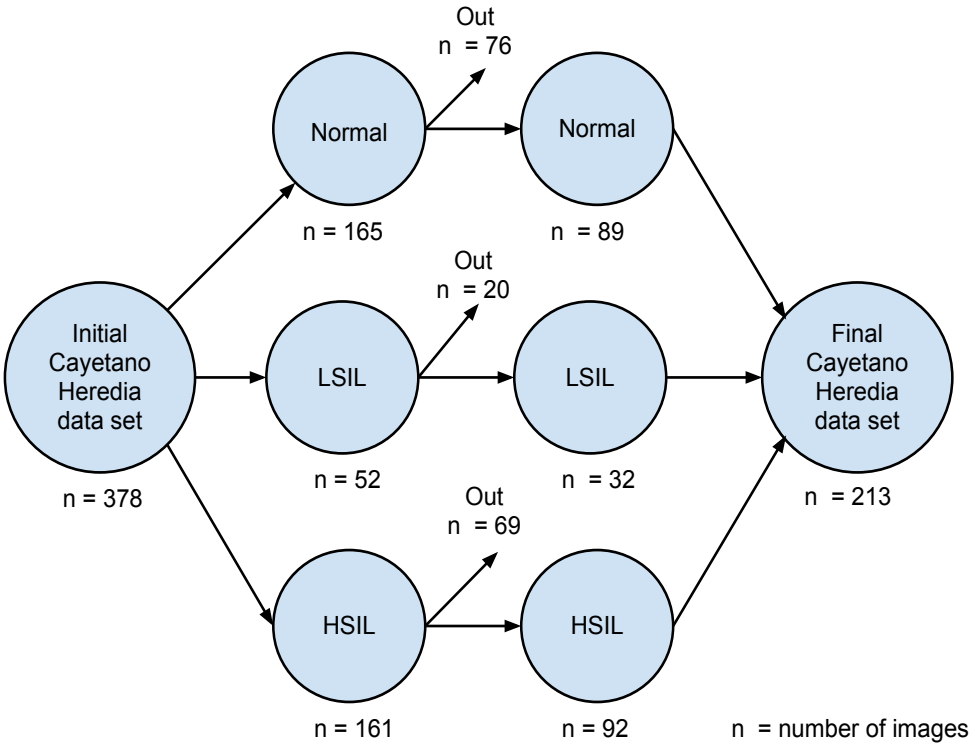


Figure 4.4: Data set of images for the adapted TB/MODS algorithm

4.2.4 Data set of Images for the UW/CSE Algorithm of LBP Texture

I used the same data set of 213 images (89 normal, 32 LSIL and 92 HSIL). No images were removed in this right branch of the tree since the LBP algorithm could handle all of them (see Figure 4.5).

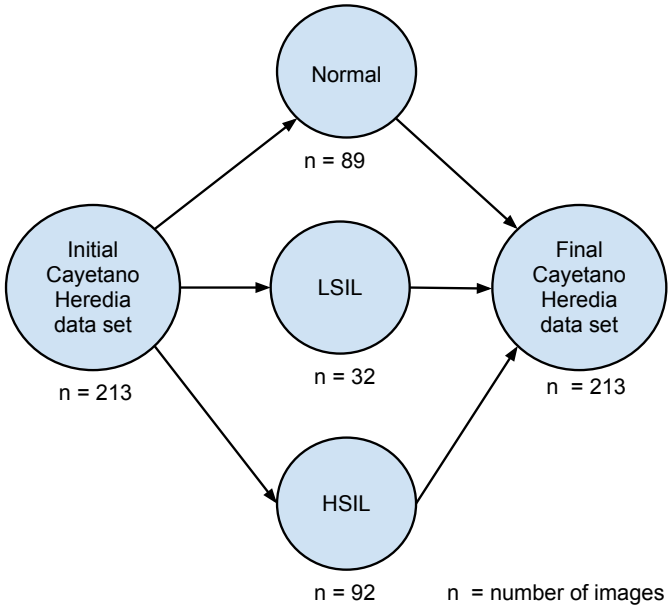


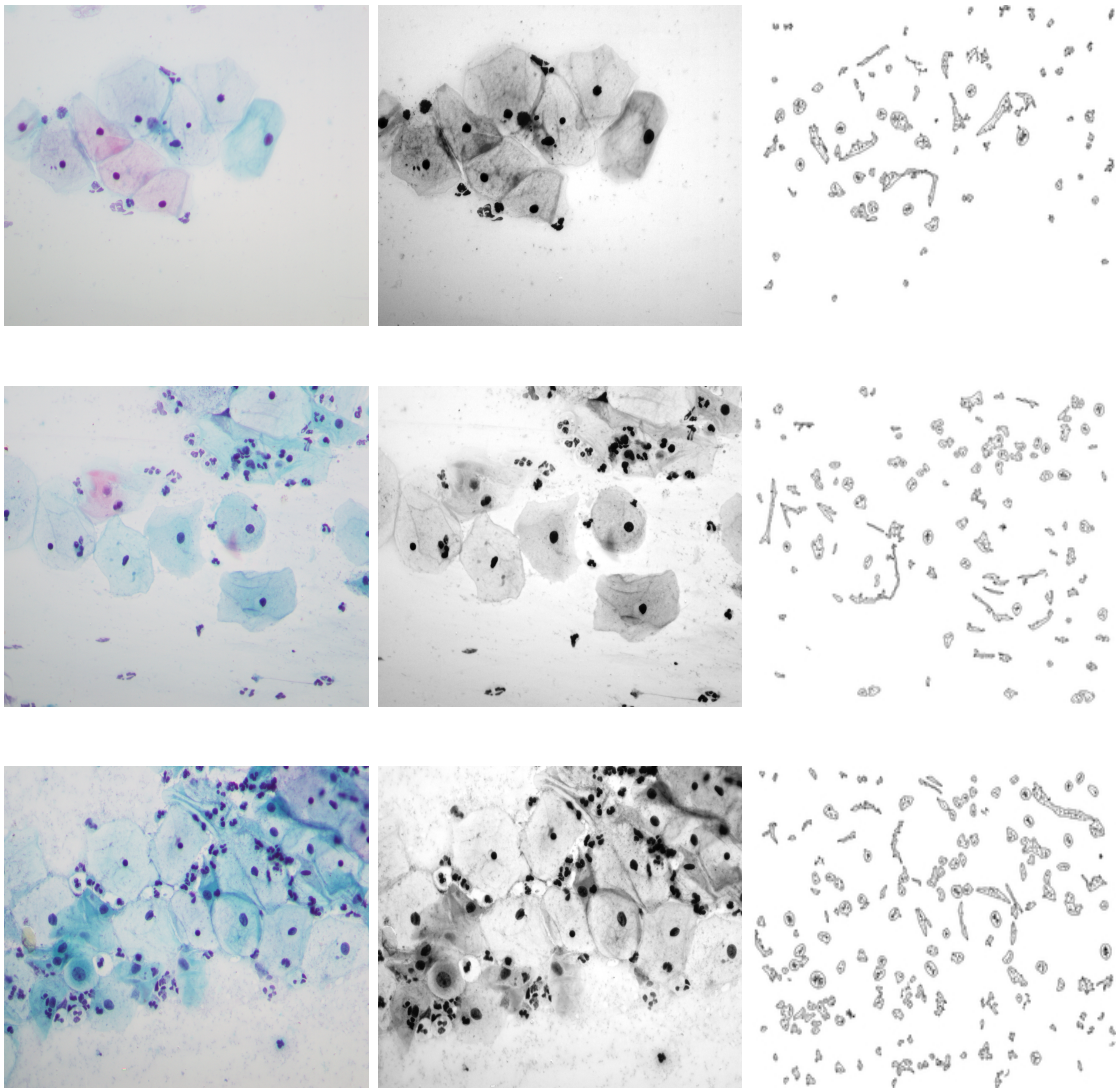
Figure 4.5: Data set of images for the UW/CSE algorithm of LBP texture

4.3 Image Processing

4.3.1 Image Processing Using the Adapted TB/MODS Algorithm

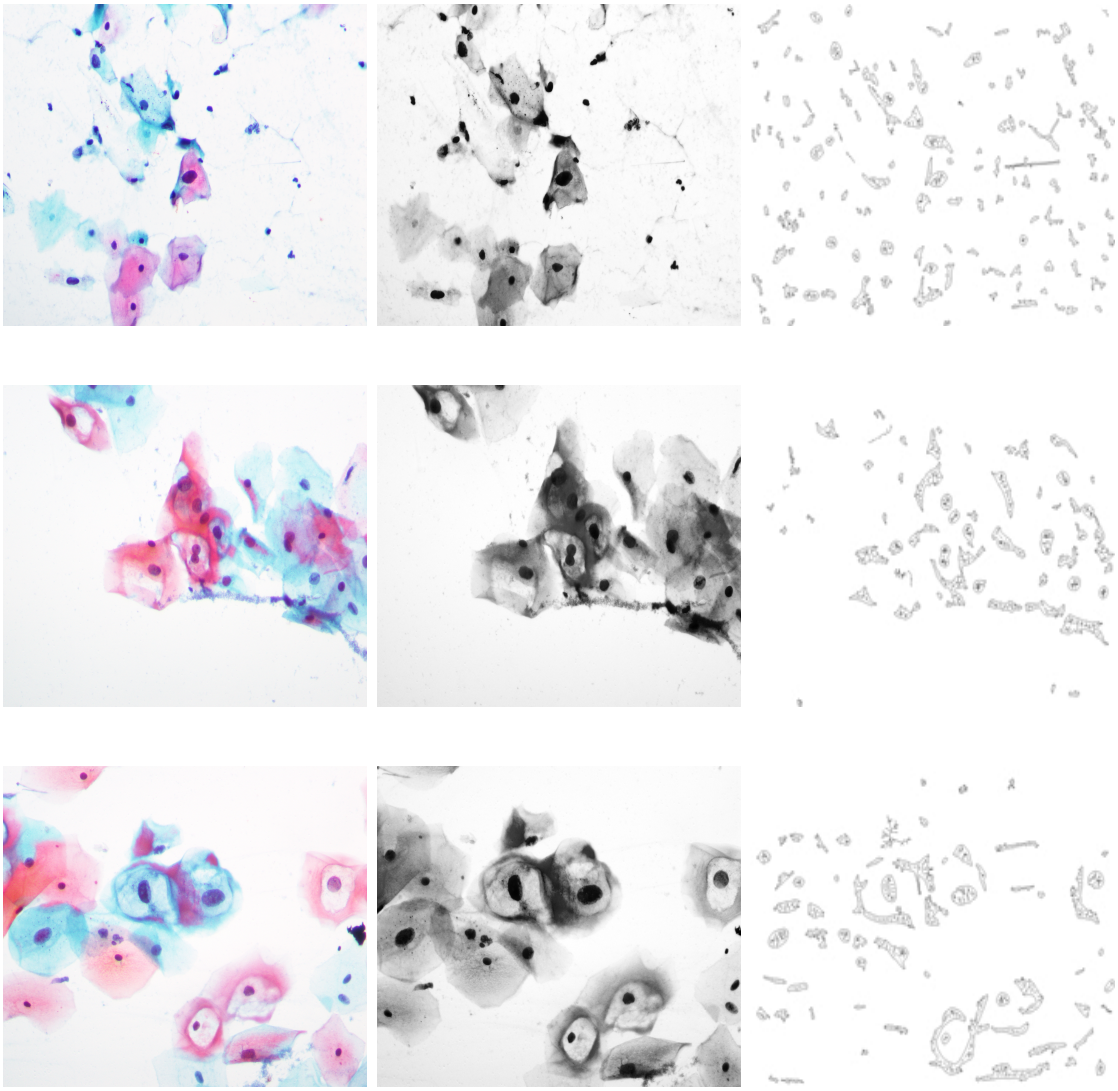
On the left branch of the tree, the proposed cervical cancer screening algorithm used the image processing from the adapted TB/MODS pattern recognition algorithm (mathematical morphology and skeletonization) for object segmentation. I followed the same methodology described in the preliminary results presented in Chapter 3 of Section 3.3. Figure 4.6 shows the image processing results of normal Pap images.

Figure 4.7 shows the image processing results of abnormal LSIL Pap images (low-grade abnormality), and Figure 4.7 shows the image processing results of abnormal HSIL Pap images (high-grade abnormality).



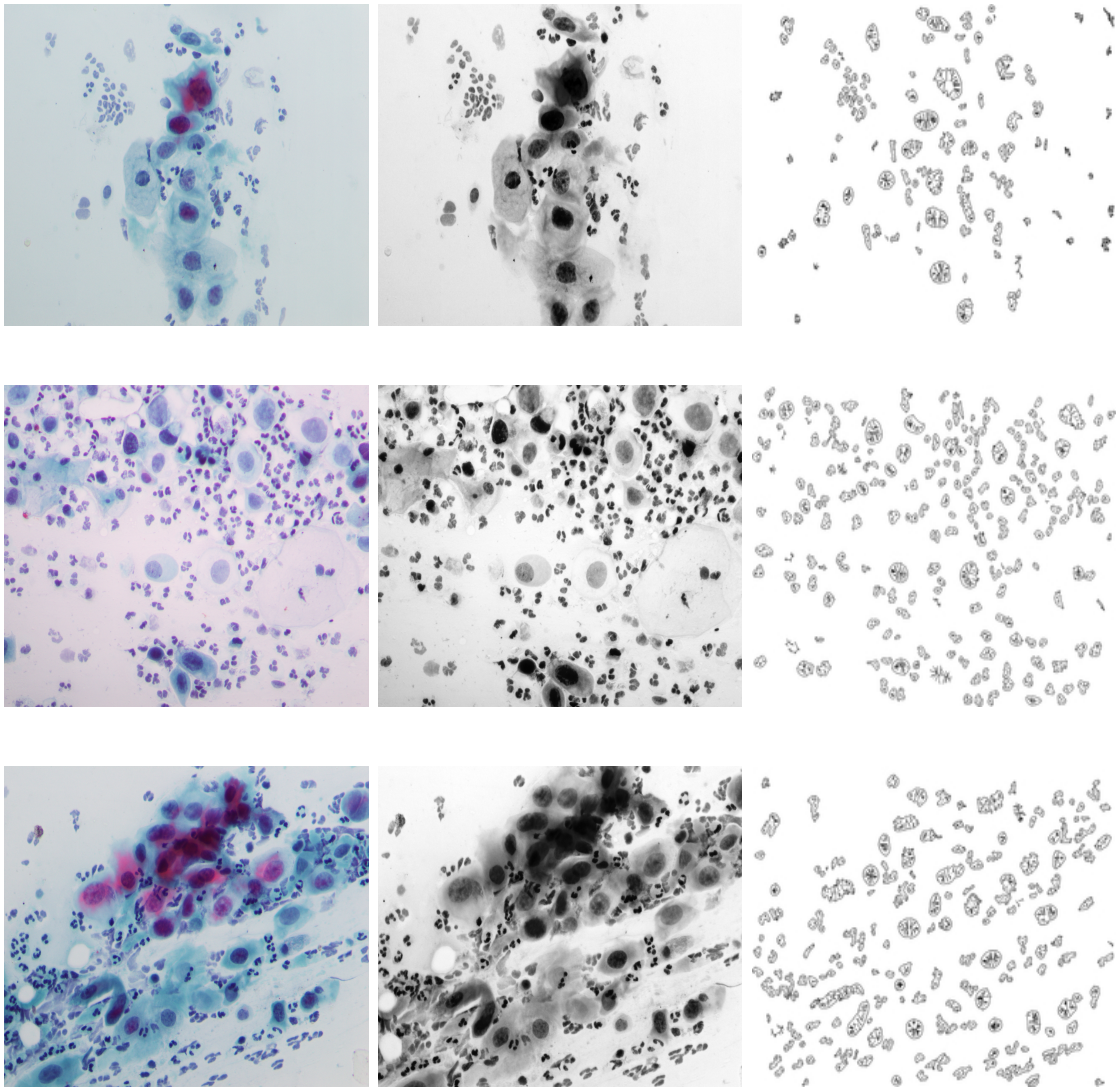
(a) Normal Pap image (b) Gray-scale Pap image (c) Edge-skeleton Pap image

Figure 4.6: Results of the image processing of normal Pap smears



(a) LSIL Pap image (b) Gray-scale Pap image (c) Edge-skeleton Pap image

Figure 4.7: Results of the image processing of abnormal LSIL Pap smears



(a) HSIL Pap image (b) Gray-scale Pap image (c) Edge-skeleton Pap image

Figure 4.8: Results of the image processing of abnormal HSIL Pap smears

4.3.2 Image Processing Using the UW/CSE Algorithm of LBP Texture

On the right branch of the tree, the proposed cervical cancer screening algorithm used the local binary pattern (LBP) texture feature from the UW/CSE pattern recognition algorithm for the whole image. The LBP texture image processing is very simple. First, a window size is defined. Then, each pixel is compared to each of its 8 neighbors. If the value of the center pixel is less than the value of its neighbors, a "0" is output. Otherwise, a "1" is output. This process provides an 8-digit binary number which is converted to a decimal number. Then, these outputs generate a 256-bin histogram representing the texture of the image [38]. There is no segmentation step in the LBP texture algorithm. See Figure 4.9 for LBP texture processing on normal Pap images. Figure 4.10 shows LBP texture processing on abnormal LSIL Pap images, and Figure 4.11 shows LBP texture processing on abnormal HSIL Pap images.

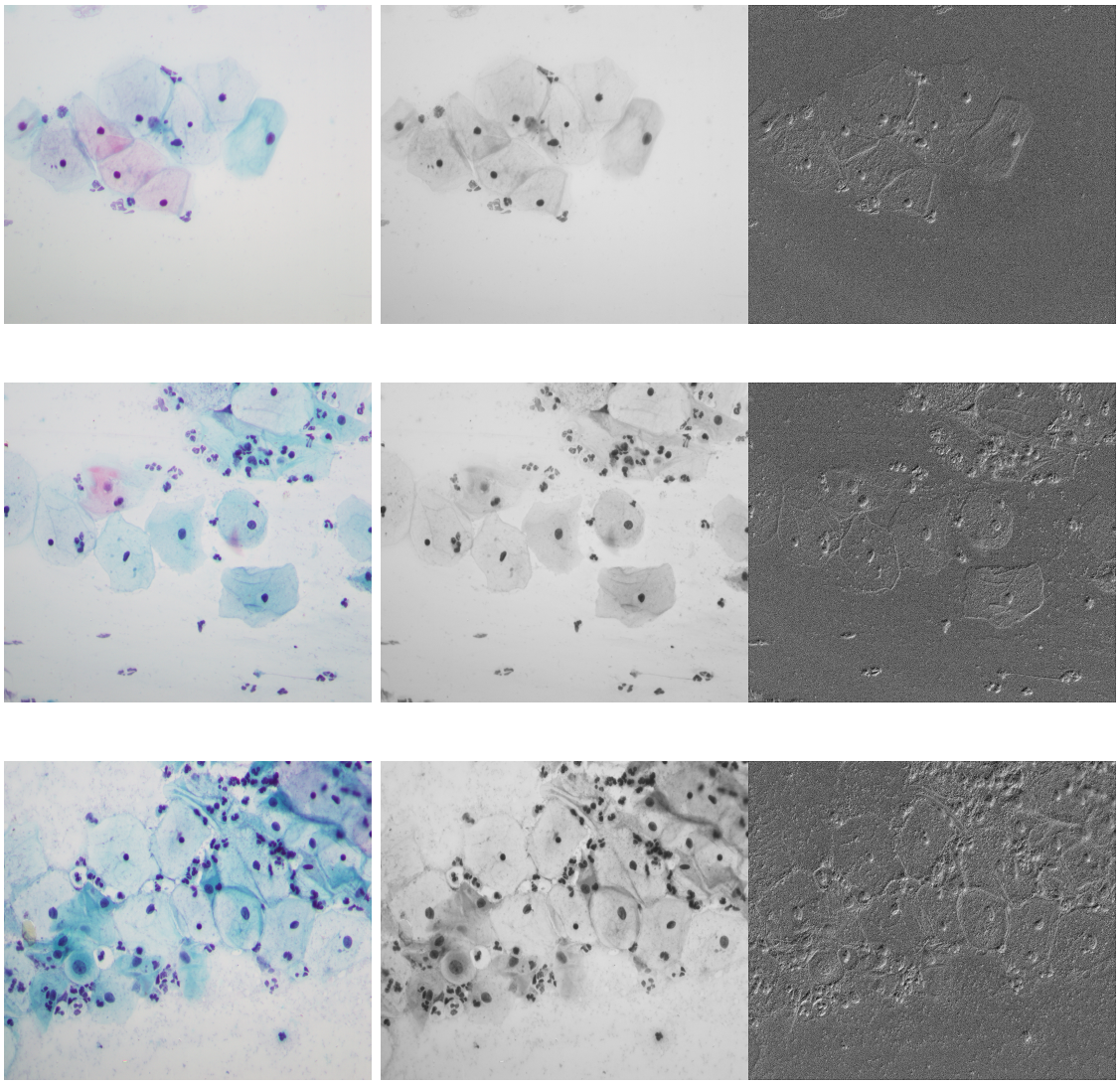
4.4 Feature Extraction and Pattern Recognition

4.4.1 Feature Extraction for the Object Classification: Nucleus vs. Non-Nucleus

The 54 features described in Appendix B [2] were extracted, after image processing, for the data set of 534 objects of nucleus and non-nucleus (267 nucleus and 267 non-nucleus). A CSV file was created using this data set. Then, different scripts automatically joined the nucleus and non-nucleus files into a single spreadsheet file with all 54 features of each 534 object. These features were used as feature descriptor for the object classification of nucleus vs. non-nucleus.

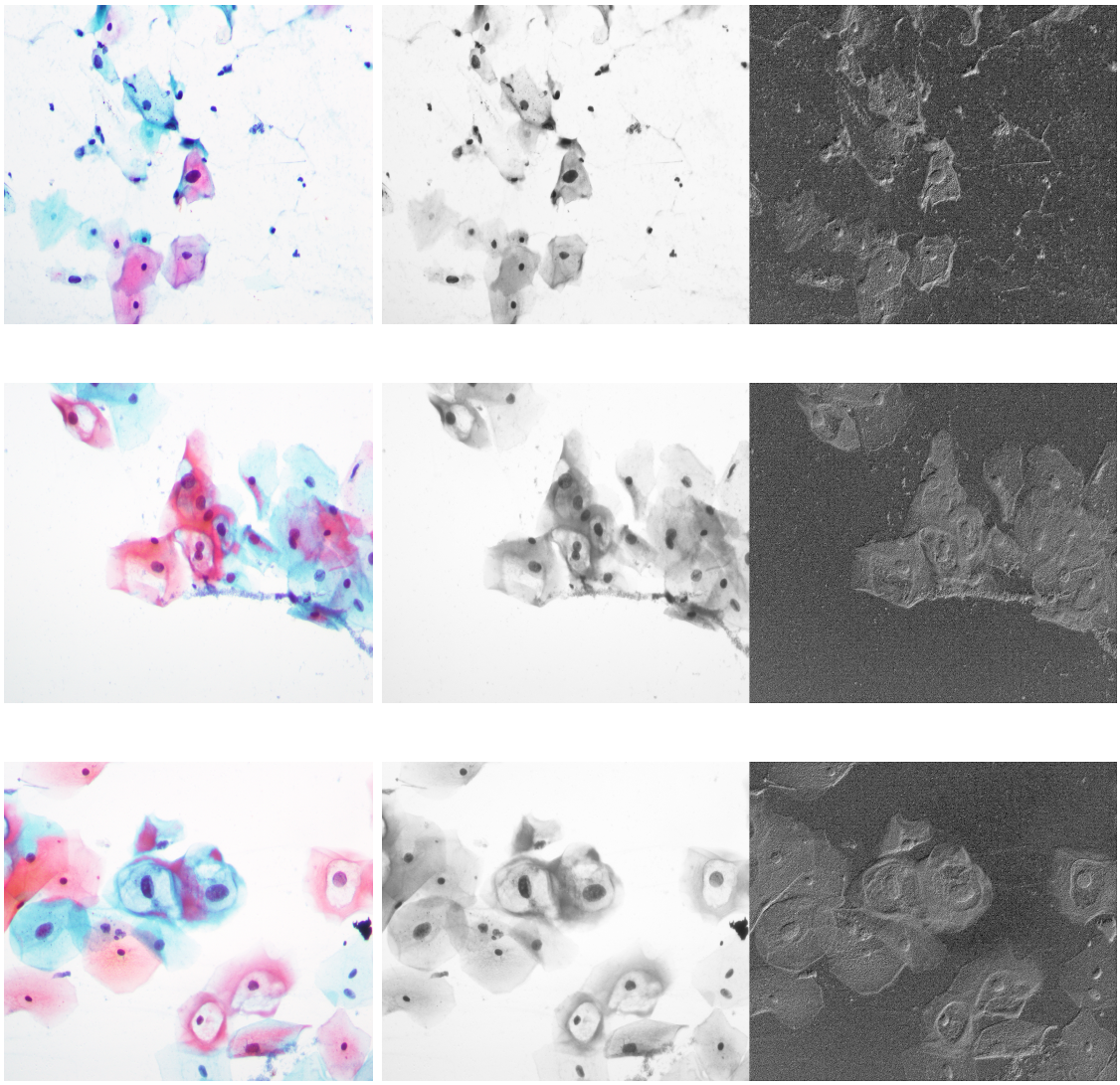
4.4.2 Feature Extraction for the Object Classification: Normal vs. Abnormal Nucleus

The algorithm extracted the same 54 features from the Appendix B [2] for the data set of 267 nucleus (150 normal, 42 LSIL and 75 HSIL). A CSV file was also created using this data set. The 54 features were used for the object classification of normal vs. abnormal nucleus that fed into the feature descriptor.



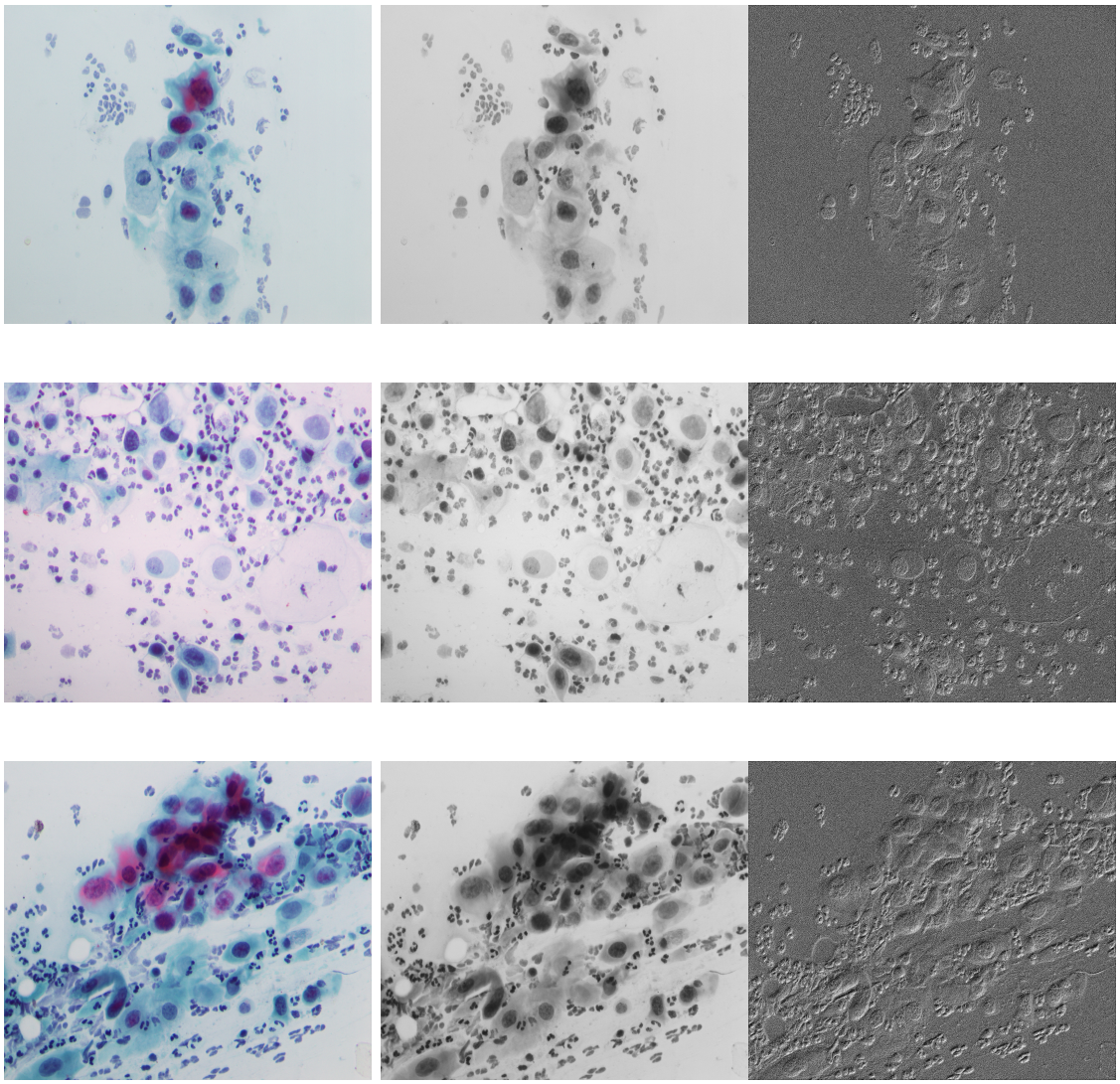
(a) Normal Pap image (b) Gray-scale Pap image and LBP texture Pap Image

Figure 4.9: Results of the image processing of normal Pap smears with LBP texture



(a) LSIL Pap image (b) Gray-scale Pap image and LBP texture Pap image

Figure 4.10: Results of the image processing of LSIL Pap smears with LBP texture



(a) HISL Pap image (b) Gray-scale Pap image and LBP texture Pap image

Figure 4.11: Results of the image processing of HSIL Pap smears with LBP texture

4.4.3 Feature Extraction for the Image Classification: Normal vs. Abnormal Image

This feature extraction used the data set of 213 images for the adapted TB/MODS algorithm, and used a different set of features compared to previous feature extractions for object classification.

The feature vector was composed of global geometric and illumination features plus a subvector f_i as shown on Table 4.1. The global geometric and illumination features are 3. The length of the subvector f_i is 25. There are four subvectors, giving a length of 100, for a total 103 dimensions. The global features are described on Table 4.2; and the feature subvector are described on Table 4.3 in detail.

Table 4.1: Feature vector

Type	**Length**	**Name**
Global geometric and illumination	3	number of objects, mean image-illumination and std dev image-illumination
Subvector f_i	25	f_1, f_2, f_3, f_4

Table 4.2: Global features

Name	**Description**
Number of objects	number of total objects in the image
Mean image-illumination	mean of the whole image-illumination
Std dev image-illumination	standard deviation of the whole image-illumination

Table 4.3: Feature subvector

Type	**Description**	**Value**
For each model m_i, feature subvector f_i includes:		
n_i	number of abnormal objects in image according to model m_i	n_1, n_2, n_3, n_4
$p_{i,j}$	12 highest probabilities of abnormal objects in image according to model m_i	$p_{i,1}$, $p_{i,2}$, $p_{i,3}$, $p_{i,4}$, $p_{i,5}$, $p_{i,6}$, $p_{i,7}$, $p_{i,8}$, $p_{i,9}$, $p_{i,10}$, $p_{i,11}$, $p_{i,12}$
$\mu_{i,j}$	mean of the 12 highest probabilities of abnormal objects in image according to model m_i	$\mu_{i,j}$ is $\mu_{i,1}$, $\mu_{i,2}$, $\mu_{i,3}$, $\mu_{i,4}$, $\mu_{i,5}$, $\mu_{i,6}$, $\mu_{i,7}$, $\mu_{i,8}$, $\mu_{i,9}$, $\mu_{i,10}$, $\mu_{i,11}$, $\mu_{i,12}$
*where m_i is m_1, m_2, m_3, m_4		

4.4.4 LBP Histogram Extraction

The LBP operator produces a 256-bin histogram representing the texture of image. The 256 values of the histogram are the feature vector for the image texture classification.

Chapter 5

CLASSIFICATION EXPERIMENTS AND RESULTS

This chapter presents classification experiments and results using machine learning algorithms employing the Weka software [21] to classify normal vs. abnormal Pap smear images assessing the accuracy, sensitivity and specificity as stated in Objective 3 of this dissertation. In addition, I will discuss the classification between nucleus vs. non-nucleus, and normal vs. abnormal nucleus using the Stata software [52].

The best image classification for normal vs. abnormal Pap smear images in the proposed cervical cancer screening algorithm has an accuracy of 98.12%, sensitivity of 98.39% and specificity of 97.75% using the most significant features of the two algorithms. I will describe the databases, classifiers, and classifications results using different training and testing sets (training set alone, 50%-training and 50%-testing, and 10-fold cross-validation) in detail below.

5.1 Databases

In the classification experiments, I worked with six databases. Four of them were described in Chapter 4, and the other two are a combination of those. These databases were stored in the CSV format and each one contained different sets of features. In Weka [54], the rows (images or objects) are termed instances, and the columns (features) are attributes. In Stata [52], the rows are named observations, and the columns are variables. The nomenclatures are different, but the content in the rows and columns are equivalent in both software packages.

5.1.1 Database of Objects: Nucleus and Non-Nucleus

This first database stores the 54 features of 534 objects from 162 images of the whole Cayetano Heredia data set as explained in Subsection 4.2.1. This database of objects contains 267 nucleus and 267 non-nucleus entries in CSV format that was used for algorithm training to classify nucleus vs. non-nucleus objects.

5.1.2 Database of Objects: Normal and Abnormal Nucleus

The second database only stores the information of the group of 267 nucleus objects from the previous one. This database was used for training to classify normal vs. abnormal nuclei.

5.1.3 Database of Images with Global and Object Models Features

The third database stores the global geometric and illumination features plus object models features generated by the object classification of normal vs. abnormal nucleus from the adapted TB/MODS algorithm. These features are described in Section 4.4.3 in detail. In total, this database contains 213 instances and 103 attributes plus one class attribute per instance.

5.1.4 Database of Images with the LBP Texture Features

The fourth database stores the LBP texture features generated by the UW/CSE algorithm. The texture features corresponds to the LBP histogram values as described in Chapter 4. In total, this database contains 213 instances and 256 attributes plus one class attribute per instance.

5.1.5 Database of Combined Features

The fifth database contains all features from the previous two databases: the global geometric and illumination features and object models features plus the LBP texture

features. In total, this data base contains 213 instances and 359 attributes plus one class attribute per instance.

5.1.6 Database of Significant Features

The last database stores the most significant features of the database of combined features after running the simple logistic regression classifier in Weka [21], and doing step-backward approximation (removing the least biological meaningful features). In total, this database contains 213 instances and 15 attributes plus one class attribute per instance.

5.2 Classifiers

There are different kinds of classifiers based on Bayes's theorem, functions, trees, rules and metaclassifiers. A summary description of the the machine learning algorithms used in Weka (stable version 3-6-9) [21] are shown on Table 5.1 [54].

5.3 Experiment 1: Classification Using the Adapted TB/MODS Algorithm

In this section, I present the classification results from the adapted TB/MODS algorithm: nucleus vs. non-nucleus, normal vs. abnormal nucleus, and normal vs. abnormal images using training sets in Stata [52].

5.3.1 Object Classification: Nucleus vs. Non-Nucleus

I built eight object models for the object classification of nucleus vs. non-nucleus using simple and multiple logistic regression. First, I used all 54 features (Appendix B) to perform a univariate analysis. Then, features with the highest Pseudo R2 and odds ratio were included, and the most highly correlated features were removed, keeping the most significant predictors. Filtered feature predictors were tested using

Table 5.1: Summary of Weka classifiers [54]

Classifier Name	**Classifier Type**	**Description**
Bayes Net	bayes	use bayesian nets based on set of random features and their conditional dependencies.
Naive Bayes	bayes	use the standard bayes' theorem of conditional probabilities.
Logistic	functions	build lineal logistic regression models. It uses a variant of the classic logistic regression.
Multilayer Perceptron (MLP)	functions	use the backpropagation neural network.
Simple Logistic	functions	build lineal logistic regression models based on attribute selection.
Sequential Minimal Optimization (SMO)	functions	use the sequential minimal optimization algorithm for support vector machine.
Adaboost & Decision Stump	metaclassifier	use Adaboost to boost the final result using as a base classifier the tree algorithm Decision Stump which construct one-level decision tree.
Random Forest	tree	build random forests.
Random Tree	tree	build a tree based on random features at each node.

multiple logistic regressions in a step-backward approximation. All these models were analyzed using Stata [52]. The models are described in Table 5.2

The eighth model identifies nucleus vs. non-nucleus with the highest sensitivity of 98.50% and the highest specificity of 98.50%. Table 5.3 describes its best six features according to the statistical analysis in detail. Three features were geometric features, and the other three features were illumination features.

5.3.2 Object Classification: Normal vs. Abnormal Nucleus

I built four object models for the object classification of normal vs. abnormal nucleus. The features of the four models are listed in Table 5.4. I applied the same methodology using simple and multiple logistic regression in Stata. The fourth model identifies normal vs. abnormal nucleus with the highest sensitivity of 87.93% and a specificity of 95.36%.

It is important to highlight that this model was applied to potential nuclei objects. Seven features were used to classify normal vs. abnormal nuclei. The best seven features are described in Table 5.5.

5.3.3 Image Classification: Normal vs. Abnormal Image

I built two image-models for the image classification of normal vs. abnormal image. I followed the same methodology using simple and multiple logistic regression. The best image model identifies normal vs. abnormal image with the highest sensitivity of 85.48% and the highest specificity of 94.38% as shown in Table 5.6. The features that best classify normal vs. abnormal Pap smear image are described in detail in Table 5.7.

Table 5.2: Results of the object classification - nucleus vs. non-nucleus

Model	Feature Name	No. Features	Sensitivity (%)	Specificity (%)
1	Circularity, max branch of the object, light refraction 8 in blue and avg of shape deviation	4	92.13	95.13
2	Circularity, max branch of the object and light refaction 8 in blue	3	89.51	97.00
3	Circularity, avg of object illumination, end-points amount and light refraction 22	4	96.25	98.12
4	Perimeter, circularity, avg of object illumination and prop of std dev of end-points	4	95.50	98.12
5	Ligth refraction 8 in the blue mask, avg of object illumination, max trans thickness and circularity	4	97.75	96.25
6	Thickness, sq waves number, light refraction 8 and dev of avg of object illumination in blue	4	97.37	98.12
7	Avg of object illumination, circularity, std dev of avg thickness, perimeter, end-points amount and dev of avg of object illumination in blue	6	96.62	98.50
8	Thickness, avg end-points, light refraction 8, dev of avg of object illumination, dev of avg of object illumination in blue and sq waves number	6	98.50	98.50

Table 5.3: Features of the best object classification: nucleus vs. non-nucleus

English Name	**Spanish Name**	**Type**	**Description**
Thickness	espesor	geometric	the average thickness of the object
Avg end-points	prompuntas	geometric	the average of the end-points of the object
Light refraction 8	bire8	illumination	brightness compared to the mean brightness of the image at 8%
Dev of avg of object illumination	bidesv	illumination	deviation of the average of the object illumination
Dev of avg of object illumination in blue	bidesv_b	illumination	deviation of the average of the object illumination in the blue mask
Sq waves number	factorforma_2	geometric	number of squared waves in the object

Table 5.4: Results of the object classification - normal vs. abnormal nucleus

Model	Feature Name	No. Features	Sensitivity (%)	Specificity (%)
1	Avg of object illumination in red, dev of avg of object illumination in red, waves number by avg of waves to length, light refraction 8 in blue, total area, avg of object illumination and light refraction 22 in blue	7	83.62%	97.35 %
2	Light refraction 8 in blue, light refraction 8, avg of object illumination in red, total area, dev of avg of object illumination in red, sq waves number and thickness	7	86.20%	94.70 %
3	Light refraction 18, avg of object illumination in red, dev of avg of object illumination in red, total area, thickness and sq waves number	7	85.34%	94.70 %
4	Total area, sq waves number, light refraction 8 in blue, light refraction 22 in blue, avg of object illumination, avg of object illumination in red and dev of avg of object illumination in red	7	87.93%	95.36 %

Table 5.5: Features of the best object classification: normal vs. abnormal nucleus

English Name	**Spanish Name**	**Type**	**Description**
Total area	area_total	geometric	area of the parallel rectangle to the axes that circumscribes the object
Sq waves number	factorforma_2	geometric	number of squared waves in the object
Light refraction 8 in blue	bire8 _b	illumination	brightness compared to the mean brightness of the image at 8% in the blue mask.
Light refraction 22 in blue	bire22_b	illumination	brightness compared to the mean brightness of the image at 22% in the blue mask.
Avg of object illumination	bimedia	illumination	average of the object illumination
Avg of object illumination in red	bimedia_r	illumination	average of the object illumination in the red mask
Dev of avg of object illumination in red	bidesv_r	geometric	deviation of the average of the object illumination in the red mask

Table 5.6: Image classification of normal vs. abnormal image

Model	**Feature Name**	**No. Features**	**Sensitivity (%)**	**Specificity (%)**
1	Mean image-illumination, std dev image-illumination, number of objects and number of positive objects from model 4 (n_4)	4	85.48%	94.38 %
2	Mean image-illumination, number of positive objects from model 3, number of positive objects from model 4 and mean probability of the two best objects from model 4	4	83.87%	93.25 %

Table 5.7: Features of the best image classification: normal vs. abnormal image

English Name	**Spanish Name**	**Type**	**Description**
Mean image-illumination	mediafoto	global illumination	mean of the whole image-illumination
Std dev image-illumination	dsfoto	global illumination	standard deviation of the whole image-illumination
Number of objects	numobj	global geometric	number of total objects in the image
Number of positive objects from model 4 (n_4)	numposi4_3	subvector	number of abnormal objects in image according to model 4

5.3.4 Summary

The best object model for nucleus and non-nucleus was able to identify nucleus with 98.50% sensitivity and 98.50% specificity in a database of 534 objects. The best object model for normal nucleus and abnormal nucleus was able to identify abnormal nucleus with 87.93% sensitivity and 95.36% specificity in a database of 267 nucleus. Finally, the best image model was able to identify abnormal Pap smear image with 85.48% sensitivity and 94.38% specificity in a database of 213 images. Figure 5.1 shows the summary of the results of the different training classifications in terms of true negative (TN), false positive (FP), false negative (FN) and true positive (TP).

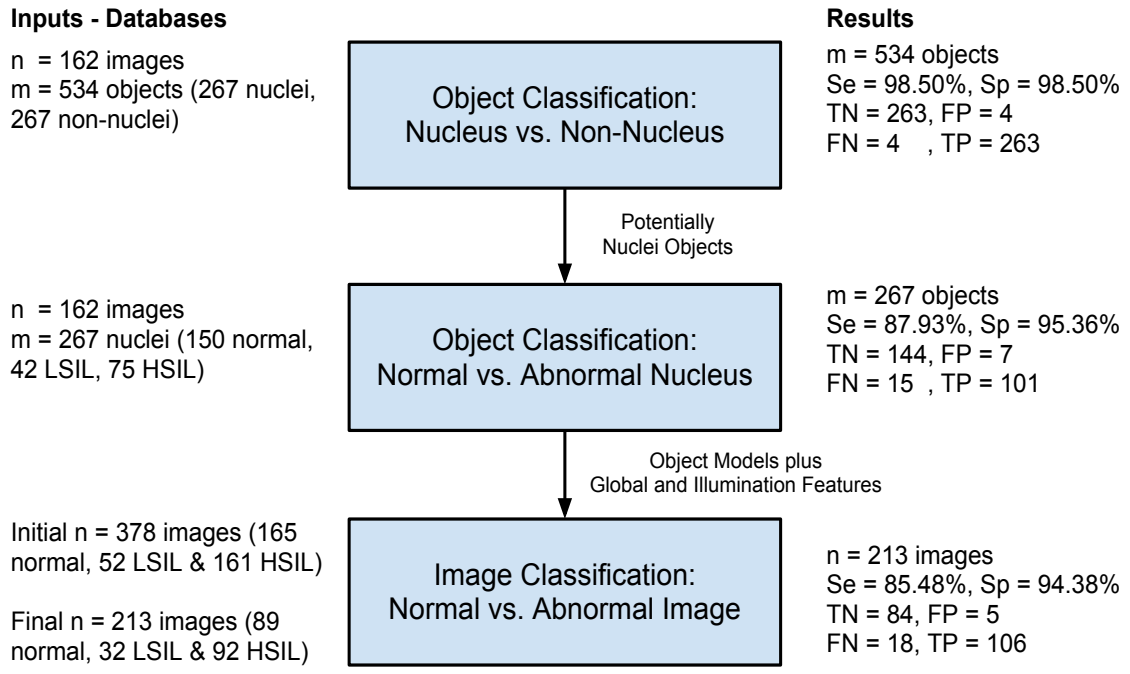


Figure 5.1: Experiment 1: Summary classification results

5.4 Experiment 2: Classification Using the UW/CSE Algorithm

In this section, I present the classification results from the UW/CSE algorithm of LBP texture for normal vs. abnormal images using Weka [21]. The UW/CSE algorithm classifies using 50% of the data for training set and 50% of data for testing set by default. The training set is built using the odd numbers of the data set, and the testing set is built using the even numbers. This was left unchanged for purposes of consistency. Table 5.8 shows the complete results. The best classification is given by MLP (multilayer perceptron - a neural network) with an accuracy of 94.33%, sensitivity of 95.16%, and specificity 93.18%.

Table 5.8: Texture classification using separate training and testing sets

Classifiers	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**TN**	**FP**	**FN**	**TP**
				(n=106)			
Bayes Net	83.01	83.87	81.82	36	8	10	52
Naive Bayes	83.96	80.65	88.64	39	5	12	50
Logistic	89.62	93.55	84.09	37	7	4	58
MLP	94.33	95.16	93.18	41	3	3	59
Simple Logistic	90.56	88.71	93.18	41	3	7	55
SMO	88.67	85.48	93.18	41	3	9	53
Adaboost+Decision Stump	86.79	80.65	95.45	42	2	12	50
Random Forest	86.79	85.48	88.64	39	5	9	53
Random Tree	85.84	83.87	88.64	39	5	10	52

5.5 Experiment 3: Classification for the Cervical Cancer Screening Algorithm

In this section, I present the classification results for normal vs. abnormal images for the proposed cervical cancer screening algorithm at different levels using 10-fold cross-validation from Weka [21].

Cross-validation is a validation method for assesing how the results of the classifier (machine learning algorithms) will generalize to different data sets. There are different types of cross-validation such as k-fold cross-validation and leave-one-out cross-validation.

The type of cross-validation that Weka runs by default is k-fold cross-validation. First, the data set is randomly split into k subsets of equal size. Then, the program uses k-1 subsets for the training set, and the remaining subset for the testing set. Second, Weka runs the process k times. Then, the results are averaged to achieve a final result. The standard method is 10-fold cross-validation which means k equal ten subsets (90% for training set and 10% for testing set) [54].

5.5.1 Pre-process

First, to get ready the database for 10-fold cross-validation, I used two filters of "Standardize" for standardizing the data values to the same scale; and, then the filter "NumericToNominal" for changing the last attribute (class attribute) from numeric to nominal. The last filter makes the class attribute categorical using 0 for normal and 1 for abnormal images.

5.5.2 10-fold Cross-Validation on the Database of Images with Global and Object Models Features

I ran different machine learning algorithms with the database of images with global and object models features using 10-fold cross-validation. See Table 5.9. The best

classification for normal vs. abnormal Pap images using just global and object models features was obtained with simple logistic with an accuracy of 88.73%, sensitivity of 91.94% and specificity of 84.27%. The classification results using 10-fold cross-validation has better sensitivity than the classification results using just training in Stata. It is important to emphasize that expert pathologists look for higher sensivity in cervical cancer screening using Pap smears.

Table 5.9: 10-fold cross-validation using global and object models features

Classifiers	**Accuracy** (%)	**Sensitivity** (%)	**Specificity** (%)	**TN**	**FP**	**FN**	**TP**
				(n=213)			
Bayes Net	86.38	87.10	85.39	76	13	16	108
Naive Bayes	84.97	91.13	76.40	68	21	11	113
Logistic	79.81	86.29	70.79	63	26	17	107
MLP	84.50	88.71	78.65	70	19	14	110
Simple Logistic	88.73	91.94	84.27	75	14	10	114
SMO	85.44	88.71	80.90	72	17	14	110
Adaboost+Decision Stump	87.79	87.90	87.64	78	11	15	109
Random Forest	83.09	84.68	80.90	72	17	19	105
Random Tree	79.81	83.87	74.16	66	23	20	104

5.5.3 10-fold Cross-Validation on the Database of Images with LBP Texture Features

I also ran different machine learning algorithms with the database of images with LBP texture features using 10-fold cross-validation. See Table 5.10. The best classification for normal vs. abnormal Pap images using just LBP texture features is obtained with MLP with an accuracy of 97.18%, sensitivity of 98.39% and specificity of 95.51%.

The classification results using 10-fold cross-validation is better than the classification results using 50% training set and 50% testing set.

Table 5.10: 10-fold cross-validation using texture features

Classifiers	**Accuracy** (%)	**Sensitivity** (%)	**Specificity** (%)	**TN**	**FP**	**FN**	**TP**
					(n=213)		
Bayes Net	85.91	85.48	86.52	77	12	18	106
Naive Bayes	85.44	83.06	88.76	79	10	21	103
Logistic	90.14	89.52	91.01	81	8	13	111
MLP	97.18	98.39	95.51	85	4	2	122
Simple Logistic	91.54	91.94	91.01	81	8	10	114
SMO	90.61	86.29	96.63	86	3	17	107
Adaboost+Decision Stump	94.36	93.55	95.51	85	4	8	116
Random Forest	93.89	95.97	91.01	81	8	5	119
Random Tree	88.26	87.90	88.76	79	10	15	109

5.5.4 10-fold Cross-Validation on the Database of Combined Features

In this classification, I used the database of combined features that stores all the information of the database of images with the global and object models features, and the database of images with the LBP texture features. I also ran the machine learning algorithms using the standard 10-fold cross-validation. See Table 5.11. MLP is again the best classifier with an accuracy of 95.30%, sensitivity of 95.97% and specificity of 94.38%. The accuracy, sensitivity and specificity was higher when using both sets of features than using just the global and object models features, but less than using just the LBP texture features.

Table 5.11: 10-fold cross-validation using combined features

Classifiers	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**TN**	**FP**	**FN**	**TP**
				(n=213)			
Bayes Net	91.07	91.94	89.89	80	9	10	114
Naive Bayes	89.67	90.32	88.76	79	10	12	112
Logistic	93.42	95.97	89.89	80	9	5	119
MLP	95.30	95.97	94.38	84	5	5	119
Simple Logistic	93.42	93.55	93.26	83	6	8	116
SMO	91.54	91.13	92.13	82	7	11	113
Adaboost + Decision Stump	94.36	95.16	93.26	83	6	6	118
Random Forest	94.36	95.97	92.13	82	7	5	119
Random Tree	79.34	79.84	78.65	70	19	25	99

5.5.5 10-fold Cross-Validation on the Database of Significant Features

To obtain better classification results using the features of the previous databases, I used the most significant features of its database of combined features taking the features from the simple logistic regression classifier in Weka [21], and doing the step-backward approximation (removing the least biologically meaningful features). This gave the best combined results. The best classification for normal vs. abnormal Pap images using the most significant features has an accuracy of 98.12%, sensitivity of 98.39% and specificity of 97.75%. See Table 5.12. MLP continues to be the best classifier. This classification results was using 15 features (1 global feature, 7 object models features, and 7 LBP texture features). See Table 5.13 and Table 5.14 for details of these features.

Table 5.12: 10-fold cross-validation using significant features

Classifiers	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**TN**	**FP**	**FN**	**TP**
					(n=213)		
Bayes Net	90.14	90.32	89.89	80	9	12	112
Naive Bayes	89.67	90.32	88.76	79	10	12	112
Logistic	92.95	93.55	92.13	82	7	8	116
MLP	98.12	98.39	97.75	87	2	2	122
Simple Logistic	93.89	95.97	91.01	81	8	5	119
SMO	94.36	94.35	94.38	84	5	7	117
Adaboost + Decision Stump	94.83	96.77	92.13	82	7	4	120
Random Forest	93.89	95.16	92.13	82	7	6	118
Random Tree	89.67	92.74	85.39	76	13	9	115

5.5.6 Conclusions

The type and size of training and testing data sets can provide different perfomance results. Furthermore, the classification methods can behave in different ways according to the nature and size of the sets of features. Thus, the size of the training set is very important. The bigger the training set, the better the results. The best classification result for normal vs. abnormal Pap images has an accuracy of 98.12%, sensitivity of 98.39% and specificity of 97.75% in the proposed cervical cancer screening algorithm using the most significant features of the two algorithms with the MLP classifier. It should be noted that the LBP texture features alone have a very good perfomance, but the combination of the key features of both algorithms improves the overall perfomance of the proposed cervical cancer screening algorithm.

Table 5.13: The most significant features of the best image classification

English Name	**Spanish Name**	**Type**	**Description**
Number of objects	numobj	global geometric	number of total objects in the image
Number of positive objects from model 1	numposi3_3	object model	number of abnormal nuclei according to the first model of the object classification of normal vs. abnormal nucleus (model 1)
Number of positive objects from model 3	numposi4_2	object model	number of abnormal nuclei according to the third model of the object classification of normal vs. abnormal nucleus (model 3)
Number of positive objects from model 4	numposi4_3	object model	number of abnormal nuclei according to the fourth model of the object classification of normal vs. abnormal nucleus (model 4)
Probability of the best object from model 4	prob4_3_1	object model	The highest probability of abnormal nuclei according to the fourth model of the object classification of normal vs. abnormal nucleus (model 4)

Table 5.14: The most significant features of the best image classification - continued

English Name	**Spanish Name**	**Type**	**Description**
Probability of the second best object from model 4	prob4_3_2	object model	second highest probability of abnormal nuclei according to the fourth model of the object classification of normal vs. abnormal nucleus (model 4)
Probability of the third best object from model 4	prob4_3_3	object model	third highest probability of abnormal nuclei according to the fourth model of the object classification of normal vs. abnormal nucleus (model 4)
Mean probability of the nine best objects from model 4	prob4_3_me_9	object model	average probability of the nine highest probabilities of abnormal nuclei according to the fourth model of the object classification of normal vs. abnormal nucleus (model 4)
Texture 24	T24	texture	histogram value generated using the LBP operator
Texture 46	T46	texture	histogram value generated using the LBP operator
Texture 71	T71	texture	histogram value generated using the LBP operator
Texture 74	T74	texture	histogram value generated using the LBP operator
Texture 131	T131	texture	histogram value generated using the LBP operator
Texture 188	T188	texture	histogram value generated using the LBP operator
Texture 218	T218	texture	histogram value generated using the LBP operator

Chapter 6

CONCLUSION, DISCUSSION AND FUTURE WORK

This Chapter presents the final conclusions and discusses the positive implications of this work in a real environment that lead to my original contributions to this research area. I also discuss the study limitations and challenges, and how I overcame them. Finally, I will describe the future work for this proof of concept imaging system for automated cervical cancer screening in Peru.

6.1 Conclusions

Cervical cancer in Peru has the highest incidence and the second highest mortality rate of cancers among women, making it one of the countries with the highest risk of cervical cancer death in the world [18]. Thus, screening techniques such as the conventional Pap smear can still be a key approach in this environment.

The implementation of the proposed cervical cancer screening algorithm for automated screening of normal vs. abnormal Pap smears may help to reduce the incidence and mortality rates of cervical cancer, especially in remote poor-settings in Peru in which there is a lack of expert pathologists and delayed follow-ups. The final version of the system classifier should have an extremely large training set (thousands of Pap images) before it is implemented in the real settings in Peru. In addition, the LBP texture algorithm should be explored in more detail due to its robustness with respect to illumination changes and its computational simplicity with real images. Furthermore, more texture features focused on the local segmentation of the nucleus can be added to the feature descriptor. It is important to mention that expert pathologists are looking for higher sensitivity during Pap smear screening to be able to recognize

all the suspicious abnormal cases. The proposed cervical cancer screening algorithm obtains the highest sensitivity of 98.39%, specificity of 97.75% and overall accuracy of 98.12% classifying normal vs. abnormal Pap images compared to the expert pathologists' ground truth, thus providing a promising system for cervical cancer screening in Peru and other developing countries.

6.2 Discussion

The overall performance of the proposed system can be compared to the manual conventional cytology in rural Peru showing that our proposed system could have a great impact in cervical cancer screening in the remote-poor settings in this country. Almonte *et al.* [1] in 2007 reported that conventional Pap smear screening in the rural Amazonian region in Peru had a low sensitivity of 42.54% and specificity of 98.68% for detecting carcinoma *in situ* or cervical cancer in a population of 5,435 women that suffers from lack of health infrastructure and expert pathologists. We can also compare our proposed system to the U.S. automated cervical cancer systems that have FDA approval. Barroeta *et al.* [3] analyzed the perfomance of the ThinPrep Imaging System in 111,080 Pap tests performed in 2007. ThinPrep showed a sensitivity of 99.95% to detect cervical abnormalities compared to manual screening. Colgan *et al.* [11] evalued the performance of the BD FocalPoint system compared to manual screening in 10,233 abnormal slides in Ontario-Canada. They reported that the BD FocalPoint had a sensitivity of 88% for detecting LSIL or worse and 83.8% for HSIL or worse showing no significant difference from the manual screening in Ontario-Canada. We can infer that the performance of our proposed system is on track in terms of overall accuracy, sensitivity and specificity especially compared to the manual screening (conventional citology) in rural Peru. In addition, the proposed system should be tested with a huge data set as shown in the evaluation studies of the commercial software to be implemented in a real setting.

An important consideration is that collaborative work needs to be conducted with

the end users that will manage the intake of the Pap smears and formats for reporting the Pap smear images in the real-system implementation. Also, there is a need for ensuring standardization of images during collection to optimize functionality of the proposed algorithm. The end users of the proposed system can be cytotechnologists or midwives in the remote areas to overcome the barrier of the lack of expert pathologists in the underserved regions in Peru.

The coverage of the Pap smear screening in the remote and/or poor regions in this country could be extended for the introduction of the proposed system. Paz-Soldan *et al.* [43] reported that the Pap smear coverage (aged 18 to 29 year olds) in urban Peru is just 30.9 %, which is higher compared to the coverage in the highlands and the jungles. Another advantage of the proposed system can be the use of the digital image collection for educational coverage in cytopathology training for health professionals across the country. An extension of the educational component can be the promotion of cervical cancer prevention through the Pap tests. Paz-Soldan *et al.* [44] also suggested starting to promote the Pap smear tests among women in the remote areas in close collaboration with the authorities in Peru. In general, the proposed work presents promising implications in terms of cervical cancer prevention in this country, especially in its underserved populations in the remote regions.

6.3 Contributions

The main contributions of this research work are:

- Two digital collections of anonymous Pap smear images in Peru. These images can be used not just for pattern recognition algorithm purposes, but also for educational purposes, such as cytopathology training for health professionals in Peru.

- The proof of concept of an automated cervical cancer screening system in Peru.

This prototype could be extended to other types of cancers that use cytopathology screening techniques.

- The design of a high sensitive classifier using the most significant features for cervical cancer screening. This methodology can be used for other classification challenges such as Tuberculosis diagnosis.

6.4 Study Limitations and Challenges

One limitation in the initial phase was that the majority of the laboratories in this country only conduct conventional Pap smears instead of Liquid-Based Cytology (LBC), which is easier to analyze, because the cells are separated. However, this dissertation addressed this limitation using the lastest classification techniques in machine learning and modern computer vision technology that can handle group of cells in a conventional Pap test, as can be seen from the results in Chapter 5.

Another limitation of this study was that there is not a standard Pap test procedure in all laboratories in Peru. We can see that in the collection of data sets of Pap smears: the Ainbo data set and the Cayetano Heredia data set. Both data sets have different stain procedures in their slides, and the digital images were acquired with different magnifications. In addition, the Ainbo data set has only ground truth at slide level, not at image or cell level. Thus, I used the Cayetano Heredia data set in the final experiments due to its expert's ground truth and origin from a real hospital setting in Peru. I also made sure that the expert's ground truth of the data sets followed the U. S. Bethesda system as a standard input for the algorithm.

Another limitation in the initial phase was that the number of pre-processing images for the adapted TB/MODS algorithm in the preliminary studies and nucleus classification was only used for training sets due to time and funding constrain. However, the adapted algorithm can generate object model features that improves the overall specificity of the proposed system. The use of the nucleus classifiers in the

overall system with the true data set of whole images, and the fact that the nucleus features were selected as most important for classification proves the utility of the nucleus classifiers.

In the preliminary studies, the related limitation contained the distinct experimental conditions on the evaluation of two pattern recognition algorithms. The TB/MODS algorithm used only a training data set, whereas the UW/CSE algorithm used small training and testing data sets. Thus, I ran the following experiments under the same conditions using the same size and type of data set (213 images) and classifiying the whole Pap image. Finally, the other limitation was that the adapted TB/MODS algorithm could not handle all the provided images due to illumination changes caused by too many blood cells and obscuring inflamation. In clinical practice, pathologists usually ask for a second specimen in this case. However, I addressed this challenge using only the remaining images (213 images) in the image classification of normal vs. abnormal Pap smears.

6.5 Ethical Aspect

I received an IRB-exemption for this dissertation at University of Washington (UW IRB-Exempt status nr. 41965) and at Universidad Peruana Cayetano Heredia (UPCH IRB-Exempt SIDISI code 59215) because this research involves the receipt and analysis of anonymous Pap images. The data sets were de-identified and cannot be linked to any specific individuals.

6.6 Future Work

Cervical cancer in Peru is a key public health issue severely affecting Peruvian women's health [39]. There are still many actions to take to fully overcome this health problem, from the proper intervention of the government authorities to the many efforts of non-governmental organizations, academia and research. Thus, we can start introducing this proof of concept system to automatically screen Pap smears images in a real

setting in Lima-Peru. We should focus on three main components: 1) algorithm optimization, 2) user interface development, and 3) pilot implementation with the experts.

First, on the algorithm optimization, it is very important to add global filters to handle illumination changes, and to appropriately categorize the inadequate or indeterminate Pap smears images (e.g. images with too many blood cells) in the adapted TB/MODS pattern recognition algorithm. It is also key to add another classification category such as "inadequate/indeterminate images". These changes will allow the adapted TB/MODS pattern recognition algorithm to be robust enough to handle the whole spectrum of Pap smear images. On the other branch of the algorithm, we can add texture features of nuclei to see the chromatin activity and/or the presence of nucleoli. Futhermore, we can test other novel computer vision techniques such as Gabor filters [19] to obtain other texture features in the Pap smears. Then, the classifiers should be trained with at least the current 966 images of the Cayetano Heredia data set.

Second, on the user interface development, we should perform an iterative system development life cycle. The specifications and the design should be based on the current workflow of cervical cancer screening in a real setting in Lima-Peru. We will use open source tools such as the LAMP approach (Linux, Apache, MySQL and PHP) to develop the user interface to facilitate a cost effective approach.

Third, on the pilot implementation, we should start our first implementation with the real setting in Lima-Peru such as the Hospital Nacional Cayetano Heredia due to the established connections and availability of the experts. The goal of the pilot implementation will be to evaluate the usability and the acceptability of the proposed cervical cancer screening system according to the end users (expert pathologists or senior medical residents). We can conduct different usability workshops.

In summary, with this study I demonstrated that the overall performance of the proposed system compared to the experts' review has the highest sensitivity of 98.39%,

specificity of 97.75% and overall accuracy of 98.12% in classifying normal vs. abnormal Pap smear images; this could greatly impact the coverage of the cervical cancer screening throughout Peru. Thus, the proposed system may help to reduce the high incidence and mortality rates of cervical cancer in this country.

BIBLIOGRAPHY

[1] M. Almonte, C. Ferreccio, J. L. Winkler, J. Cuzick, V. Tsu, S. Robles, R. Takahashi, and P. Sasieni. Cervical screening by visual inspection, HPV testing, liquid-based and conventional cytology in Amazonian Peru. *International Journal of Cancer*, 121(4):796–802, Aug 2007.

[2] A. Alva, F. Aquino, D. Requena, C. Olivares, A. Gutierrez, L. Caviedes, J. Coronel, P. Sheen, S. Larson, R. H. Gilman, D. Moore, and M. Zimic. Morphological characterization of Mycobacterium tuberculosis in a MODS culture for an automatic diagnostics through pattern recognition. *Universidad Peruana Cayetano Heredia*, 2012. Manuscript submitted for publication.

[3] J. E. Barroeta, M. E. Reilly, M. M. Steinhoff, and W. D. Lawrence. Utility of the Thin Prep imaging system in the detection of squamous intraepithelial abnormalities on retrospective evaluation: Can we trust the imager? *Diagnostic Cytopathology*, 40(2):124–127, Feb 2012. doi: 10.1002/dc.21516.

[4] P. H. Bartels, W. Abmayr, M. Bibbo, G. Burger, H. J. Soost, J. Taylor, and G. L. Wied. Computer recognition of ectocervical cells: Image features. *Analytical and Quantitative Cytology*, 3(2):157–64, Jun 1981.

[5] C. Bergmeir, M. García-Silvente, and J. M. Benítez. Segmentation of cervical cell nuclei in high-resolution microscopic images: A new algorithm and a web-based software framework. *Computer Methods and Programs in Biomedicine*, 107(3):497–512, Sep 2012. 10.1016/j.cmpb.2011.09.017.

[6] J. Bernsen. Dynamic thresholding of grey-level images. *Proocedings of the 8th International Conference on Pattern Recognition*, pages 1251–1255, Oct 1986.

[7] M. M. Blas, I. E. Alva, P. J. Garcia, C. Carcamo, S. M. Montano, R. Mu nante, and J. R. Zunt. Association between human papillomavirus and human T-lymphotropic virus in indigenous women from the Peruvian amazon. *PLoS ONE*, 7(8):e44240, 2012. doi:10.1371/journal.pone.0044240.

[8] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'REILLY, first edition, 2008.

[9] F. Bray, J. S. Ren, E. Masuyer, and J. Ferlay. Estimates of global cancer prevalence for 27 sites in the adult population in 2008. *International Journal of Cancer*, 132(5):1133–1145, Mar 2013. doi:10.1002/ijc.27711.

[10] M. E. Celebi. *Fourier: An open-source image processing and analysis library written in ANSI C*. Louisiana State University at Shreveport. Available: `http://www.lsus.edu/emre-celebi`, last accessed in May 2013.

[11] T. J. Colgan, N. Bon, S. Clipsham, G. Gardiner, J. Sumner, V. Walley, and C. M. McLachlin. A validation study of the FocalPoint GS Imaging System for gynecologic cytology screening. *Cancer Cytopathology*, 121(4):189–196, Apr 2013. doi: 10.1002/cncy.21271.

[12] Dimac Imaging Company. CHAMP: Citology and histology analysis modular package. Official web site. Available: `http://www.dimac-imaging.com`, last accessed in May 2013.

[13] HOLOGIC: The Women's Health Company. About the ThinPrep Imaging System. Official web site. Available: `http://www.thinprep.com/info/why_pap_test/thinprep_imaging_system.html`, last accessed in May 2013.

[14] N. Dell and W. Brunette. CSE/EE 577: Medical image analysis - course research project. Technical report, University of Washington. Autumn 2011.

[15] M. Desai. Role of automation in cervical cytology. *Diagnostic Histopathology*, 15(7):323–329, Jun 2009. Mini-Symposium: Cervical Cytology.

[16] BD Diagnostics. BD FocalPoint GS Imaging System. Official web site. Available: `http://www.bd.com/tripath/labs/fp_gs_system.asp`, last accessed in May 2013.

[17] J. H. Eichhorn, T. A. Brauns, J. A. Gelfrand, B. A. Crothers, and D. C. Wilbur. A novel automated screening and interpretation process for cervical cytology using the internet transmission of low-resolution images: A feasibility study. *Cancer Cytopathology*, 105(4):199–206, Aug 2005.

[18] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin. GLOBOCAN 2008 v2.0, Cancer incidence and mortality Worldwide: IARC CancerBase no. 10 [internet], 2010. Available: `http://globocan.iarc.fr`, last accessed in May 2013.

[19] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biological Cybernetics*, 61(2):103–113, Jun 1989.

[20] Program for Appropriate Technology in Health (PATH). RHO Cervical cancer. Official website. Available: `http://www.rho.org/about-cervical-cancer.htm`, last accessed in May 2013.

[21] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009. Available: `http://www.cs.waikato.ac.nz/ml/weka`, last accessed in May 2013.

[22] Hospital Nacional Cayetano Heredia. Department of Clinical Pathology and Pathological Anatomy. Official web site. Available: `http://www.hospitalcayetano.gob.pe/index.php?option=com_content&task=view&id=534&Itemid=`, last accessed in May 2013.

[23] Universidad Peruana Cayetano Heredia. QUIPU: Andean Global Health Informatics Research and Training Center. Official web site. Available: `http://www.andeanquipu.org/`, last accessed in May 2013.

[24] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard. Pap-smear benchmark data for pattern classification. *Proceedings in NiSIS 2005: Nature inspired Smart Information Systems (NiSIS), EU co-ordination action*, pages 1–9, 2005. Albufeira, Portugal.

[25] A. Kale. Segmentation and classification of cervical cell images. Master's thesis, Department of Computer Engineering at the Bilkent University, 2010. Ankara, Turkey.

[26] J. S. Lee, W. I. Bannister, L. C. Kuan, P. H. Bartels, and A. C. Nelson. A processing strategy for automated papanicolaou smear screening. *Analytical and Quantitative Cytology and Histology*, 14(5):415–425, Oct 1992.

[27] J. S. Lee, L. C. Kuan, M. Rosenlof, D. Brancheau, P. H. Bartels, and A. C. Nelson. Object analysis and decision strategy in an automated pap smear prescreening system. *Electronic Imaging*, 90:220–223, Oct-Nov 1990.

[28] J. S. Lee and A. C. Nelson. Stanley F. Patten, Jr., M.D., Ph.D. and the development of an automated Papanicolau smear screening system. *Cancer Cytopathology*, 81(6), Dec 1997.

[29] P. Locht. A system for automated screening for cervical cancer, 2003. Presentation in the Medical Vision Day at the Technical University of Denmark. Available: http://www2.imm.dtu.dk/visiondag/VD03/medicinsk/medvision.html, last accessed in May 2013.

[30] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[31] S. Luciani and J. Winkler. Cervical cancer prevention in Peru: Lessons learned from the TATI demostration project. Technical report, Pan American Health Organization, 2006. Washington. Available: http://www.rho.org/files/PAHO_PATH_TATI_report_2006.pdf, last accessed in May 2013.

[32] N. A. Mat-Isa, M. Y. Mashor, and N. H. Othman. An automated cervical precancerous diagnostic system. *Articial Intelligence in Medicine*, 42(1):1–11, Jan 2008.

[33] W. Niblack. *An Introduction to Digital Image Processing*. Prentice Hall, first edition, 1986. Englewood Cliffs, New Jersey.

[34] American Society of Cytopathology. *NCI Bethesda System website atlas*. Available: http://nih.techriver.net/, last accessed in May 2013.

[35] Institute of Electrical and Electronics Engineers. IEEE Xplore. Official web site. Available: http://ieeexplore.ieee.org/Xplore/guesthome.jsp, last accessed in May 2013.

[36] Ministry of Health Peru. *National plan of gynecologic cancer prevention: Cervix and breast cancer 1998-2000*. Edition in Spanish. Available: http://bvs.minsa.gob.pe/local/MINSA/1174_MINSA1411.pdf, last accessed in May 2013.

[37] US National Library of Medicine - National Institutes of Health. PubMed. Official web site. Available: http://www.ncbi.nlm.nih.gov/pubmed, last accessed in May 2013.

[38] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[39] WHO/ICO Information Centre on HPV and Cervical Cancer (HPV Information Centre). Human Papillomavirus and related cancers in Peru. Summary Report, Sep 2010. Available: www.who.int/hpvcentre, last accessed in May 2013.

[40] WHO/ICO Information Centre on HPV and Cervical Cancer (HPV Information Centre). Human Papillomavirus and related cancers in World. Summary Report, Nov 2010. Available: `www.who.int/hpvcentre`, last accessed in May 2013.

[41] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on Systems, Man and Cibernetics*, 9(1):62–66, Jan 1979.

[42] S. F. Patten, J. S. Lee, D. C. Wilbur, T. A. Bonfiglio, T. J. Colgan, R. M. Richart, H. Cramer, and S. Moinuddin. The AutoPap 300 QC System multicenter clinical trials for use in quality control rescreening of cervical smears. *Cancer Cytopathology*, 81(6):337–342, Dic 1997.

[43] V. A. Paz-Soldan, F. H. Lee, C. Carcamo, K. K. Holmes, G. P. Garnett, and P. Garcia. Who is getting Pap smears in urban Peru?

[44] V. A. Paz-Soldan, L. Nussbaum, A. M. Bayer, and L. Cabrera. Low knowledge of cervical cancer and cervical pap smears among women in Peru, and their ideas of how this could be improved. *International Quarterly Community Health Education*, 31(3):245–236, 2011. doi: 10.2190/IQ.31.3.d.

[45] J. E. Pérez-Lu. *Time flowchart spent by a woman to receive her Pap smear in Callao-Peru*. Universidad Peruana Cayetano Heredia, 2010. Data collected at the Regional Health Directorate in Callao-Peru (DIRESA Callao). Unpublished raw data.

[46] M. E. Plissiti, C. Nikou, and A. Charchanti. Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):233–241, Mar 2011. doi:10.1109/TITB.2010.2087030.

[47] M. E. Plissiti, C. Nikou, and A. Charchanti. Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images. *Pattern Recognition Letters*, 32(6):838–853, 2011.

[48] W. Rasband. *ImageJ: Image Processing and Analysis in Java*. Research Services Branch, National Institute of Mental Health. Official website. Available: `http://rsb.info.nih.gov/ij/`, last accessed in May 2013.

[49] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, Feb 2000.

[50] T. Schilling, L. Miroslaw, G. Glab, and M. Smereka. Towards rapid cervical cancer diagnosis: automated detection and classification of pathologic cells in phase-contrast images. *International Journal of Gynecological Cancer*, 17(1):118–126, Jan-Feb 2007.

[51] J. Sherris, S. Wittet, A. Kleine, J. Sellors, S. Luciani, R. Sankaranarayanan, and M. A. Barone. Evidence-based, alternative cervical cancer screening approaches in low-resource settings. *International Perspectives on Sexual and Reproductive Health*, 35(3):147–152, Sep 2009.

[52] StataCorp. Stata Statistical Software: Release 10, 2007. College Station, Texas: StataCorp LP.

[53] G. L. Wied, P. H. Bartels, M. Bibbo, J. Puls, J. Taylor, and J. J. Sychra. Computer recognition of ectocervical cells. Classification accuracy and spatial resolution. *Acta Cytologica*, 21(6):753–764, Nov-Dec 1977.

[54] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems, third edition, 2011.

Appendix A

TIME FLOWCHART OF PAP SMEAR SCREENING IN PERU

The public health research group at the Universidad Peruana Cayetano Heredia in Lima-Peru collected data showing that the time delay for women to receive their Pap results is around 3 to 9 months at one health network of the Ministry of Health in Callao-Peru [45]. The time flowchart is shown below [45].

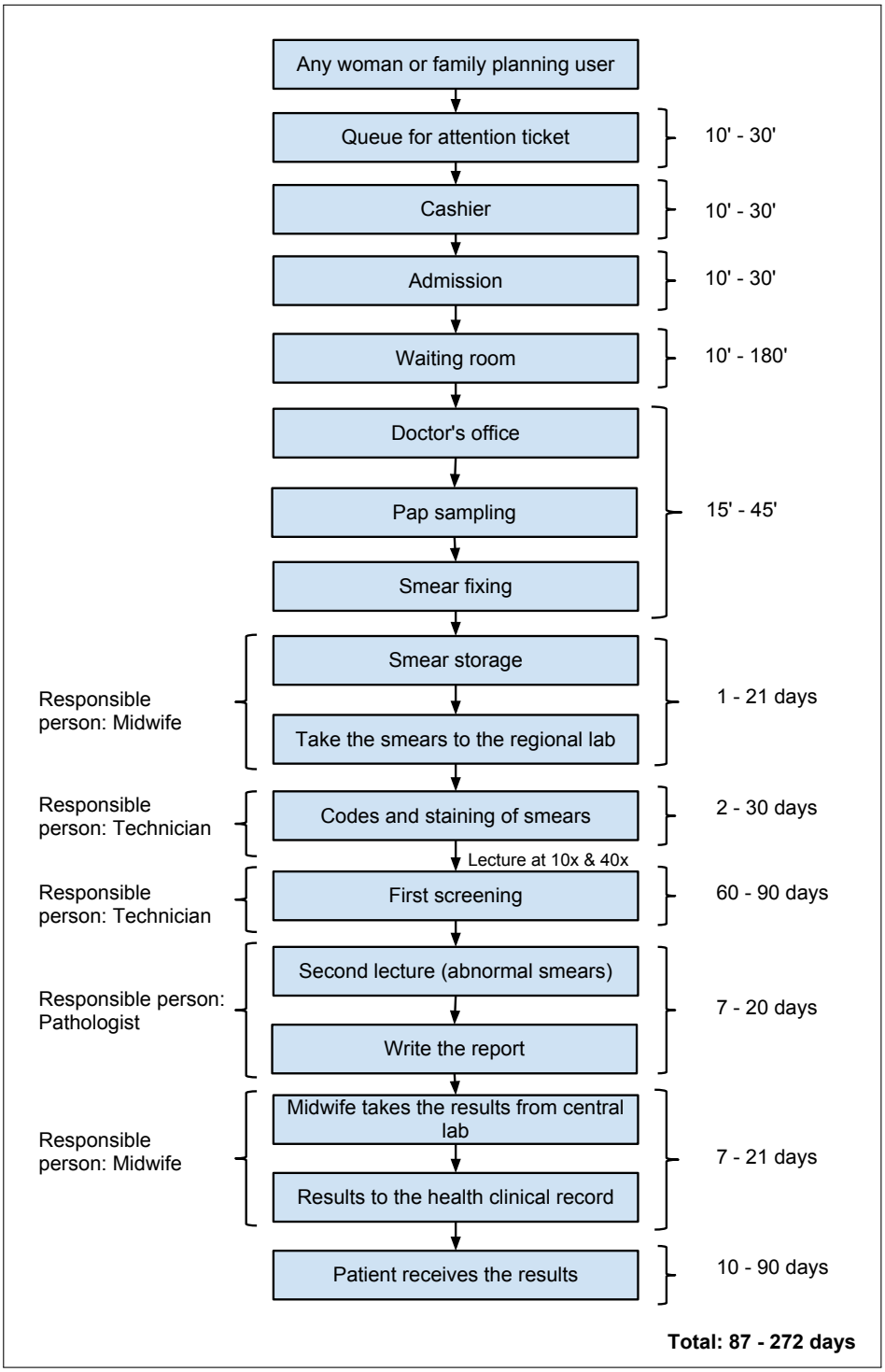


Figure A.1: Time flowchart spent by a woman to receive her Pap smear result in Peru [45]

Appendix B

FEATURES OF THE TB/MODS PATTERN RECOGNITION ALGORITHM

The features of the TB/MODS pattern recognition algorithm developed in the Bioinformatics Unit at the Universidad Peruana Cayetano Heredia are shown in the following tables [2].

Table B.1: List of features of the TB/MODS pattern recognition algorithm [2]

English Name	**Spanish Name**	**Type**	**Description**
Image name	numfoto	global	the name of the image
Length	longitud	geometric	length of the object
Perimeter	perimetro	geometric	perimeter of the object
Circularity	circularidad	geometric	circularity of the object
Max trans thickness	maxi_espe	geometric	maximum transverse thickness
Prop thickness to length	maxi_espe_re	geometric	proportion of maximum transverse thickness to the length
Area	area bact	geometric	area of the object
Total area	area total	geometric	area of the parallel rectangle to the axes that circumscribes the object

Table B.2: List of features of the TB/MODS pattern recognition algorithm [2] - continued

English Name	**Spanish Name**	**Type**	**Description**
End-points amount	cant puntas	geometric	amount of end-points of the skeleton of the object
End-points	puntas	geometric	width of the end-points of the object
Avg end-points	prom puntas	geometric	average of the end-points
Std dev of end-points	prom puntas des st	geometric	standard deviation of the previous feature
Avg prop of end-points	puntas_re	geometric	average proportion of the end-points to its length
Prop of std dev of end-points	prom puntas re des st	geometric	proportion of the standard deviation of the previous feature to its length
Thickness	espesor	geometric	average thickness of the object
Std dev of avg thickness	espesor desv stan	geometric	standard deviation of the average thickness of the object
Prop of avg thickness	espesor_re	geometric	proportion of the average thickness to length
Std dev of prop of avg thickness	espesor_re desv stan	geomtric	standard deviation of the previous feature

Table B.3: List of features of the TB/MODS pattern recognition algorithm [2] - continued

English Name	**Spanish Name**	**Type**	**Description**
Avg of shape deviation	prom for	geometric	average of the shape's deviation of the object to a straight line
Std dev of avg of shape deviation	d_sta2_for	geometric	standard deviation of the previous feature
Min size of object wave	tam p min ond	geometic	minimum size of the object wave
Waves num by avg of the object wave	formaprom onda	geometric	number of waves by the average measurement of the object waves
Prop of max size wave to length	tam p max ond_re	geometric	proportion of maximum size wave to length
Prop of min size wave to length	tam p min ond_re	geometric	proportion of the minimum size wave to length
Waves number by avg of waves to length	formaprom onda re	geometric	number of waves by average of relative size waves to length
Max curvature of the class	max k class_k	geometric	curvature maxima of the classes
Min curvature of the class	min k class_k	geometric	curvature minima of the classes
Max curvature	max kkk	geometic	maxima curvature

Table B.4: List of features of the TB/MODS pattern recognition algorithm [2] - continued

English Name	**Spanish Name**	**Type**	**Description**
Min curvature	min kkk	geometric	minimum curvature
Avg curvature	promedio curvatura	geometric	average of the curvature
Waves number by max size of waves	factor forma	geometric	number of waves by maximum size of waves
Waves number by max size sq wave	factor forma 2	geometric	number of waves by maximum size squared wave
Waves number by max size cubed wave	factor forma 3	geometric	number of waves by maximum size cubed wave
Waves number by max size waves 4	factor forma 4	geometric	number of waves by maximum size of waves factor 4
Sq waves number	factor forma_2	geometric	number of squared waves
Sq waves number by max size sq waves	factor forma2_2	geometric	number of squared waves by maximum size squared wave
Cubed waves number by max size sq wave	factor forma3_2	geometric	number of cubed waves by maximum size squared wave
waves factor 4 by max size sq wave	factor forma4_2	geometric	Number of waves factor 4 by maximum size squared wave
max branch of the object	Tam_max_ram	geometric	Size of the maximum branch of the object

Table B.5: List of features of the TB/MODS pattern recognition algorithm [2] - continued

English Name	**Spanish Name**	**Type**	**Description**
Max branch of the object divided length	tam_max_ram_re	geometric	previous feature divided by the length
Perimeterlength	perimetrolongitud	geometric	perimeter divided by the length
Light refraction 8	bire 8	illumination	brightness compared to the mean brightness of the image at 8%
Light refraction 10	bire 10	illumination	brightness compared to the mean brightness of the image at 10%
Light refraction 12	bire 12	illumination	brightness compared to the mean brightness of the image at 12%
Light refraction 15	bire 15	illumination	brightness compared to the mean brightness of the image at 15%
Light refraction 18	bire 18	illumination	brightness compared to the mean brightness of the image at 18%
Light refraction 20	bire 20	illumination	brightness compared to the mean brightness of the image at 20%
Light refraction 22	bire 22	illumination	brightness compared to the mean brightness of the image at 22%
Avg of object illumination	bi media	illumination	average of the object illumination
Dev of the avg of object illumination	bi desv	illumination	deviation of previous feature

Table B.6: List of features of the TB/MODS pattern recognition algorithm [2] - continued

English Name	**Spanish Name**	**Type**	**Description**
Light refraction 8 in red	bire 8_r	illumination	brightness compared to the mean brightness of the image at 8% in the red mask
Light refraction 10 in red	bire 10_r	illumination	brightness compared to the mean brightness of the image at 10% in the red mask
Light refraction 12 in red	bire 12_r	illumination	brightness compared to the mean brightness of the image at 12% in the red mask
Light refraction 15 in red	bire 15_r	illumination	brightness compared to the mean brightness of the image at 15% in the red mask
Light refraction 18 in red	bire 18_r	illumination	brightness compared to the mean brightness of the image at 18% in the red mask
Light refraction 20 in red	bire 20_r	illumination	brightness compared to the mean brightness of the image at 20% in the red mask
Light refraction 22 in red	bire 22_r	illumination	brightness compared to the mean brightness of the image at 22% in the red mask
Avg of object illuminatio in red	bi media_r	illumination	average values of illumination object in the red mask
Dev of avg of object illumination in red	bi desv_r	illumination	deviation of the previous feature

Table B.7: List of features of the TB/MODS pattern recognition algorithm [2] - continued

English Name	**Spanish Name**	**Type**	**Description**
Light refraction 8 in green	bire 8_g	illumination	Brightness compared to the mean brightness of the image at 8% in the green mask
Light refraction 10 in green	bire 10_g	illumination	Brightness compared to the mean brightness of the image at 10% in the green mask
Light refraction 12 in green	bire 12_g	illumination	Brightness compared to the mean brightness of the image at 12% in the green mask
Light refraction 15 in green	bire 15_g	illumination	Brightness compared to the mean brightness of the image at 15% in the green mask
Light refraction 18 in green	bire 18_g	illumination	Brightness compared to the mean brightness of the image at 18% in the green mask
Light refraction 20 in green	bire 20_g	illumination	Brightness compared to the mean brightness of the image at 20% in the green mask
Light refraction 22 in green	bire 22_g	illumination	Brightness compared to the mean brightness of the image at 22% in the green mask
Avg of object illumination in green	bi media_g	illumination	Average of the object illumination in the green mask
Dev of avg of object illumination in green	bi desv_g	illumination	Deviation of the previous feature

Table B.8: List of features of the TB/MODS pattern recognition algorithm [2] - continued

English Name	**Spanish Name**	**Type**	**Description**
Light refraction 8 in blue	bire 8_b	illumination	brightness compared to the mean brightness of the image at 8% in the blue mask
Light refraction 10 in blue	bire 10_b	illumination	brightness compared to the mean brightness of the image at 10% in the blue mask
Light refraction 12 in blue	bire 12_b	illumination	brightness compared to the mean brightness of the image at 12% in the blue mask
Light refraction 15 in blue	bire 15_b	illumination	brightness compared to the mean brightness of the image at 15% in the blue mask
Light refraction 18 in blue	bire 18_b	illumination	brightness compared to the mean brightness of the image at 18% in the blue mask
Light refraction 20 in blue	bire 20_b	illumination	brightness compared to the mean brightness of the image at 20% in the blue mask
Light refraction 22 in blue	bire 22_b	illumination	brightness compared to the mean brightness of the image at 22% in the blue mask
Avg of object illumination in blue	bi media_b	illumination	average of the object illumination in the blue mask
Dev of avg of object illumination in blue	bi desv_b	illumination	deviation of the previous feature
Mean image-illumination	media foto	global	mean of the whole image-illumination
Std dev image illumination	ds foto	global	standard deviation of the whole image-illumination

VITA

Mabel Karel Raza Garcia earned her BS degree in Informatics Engineering at the Universidad Peruana Cayetano Heredia in 2004 and her MS in Biomedical and Health Informatics at the University of Washington in 2009. She is currently a PhD candidate and Fogarty fellow in Biomedical and Health Informatics at the University of Washington.