FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Cervical Cancer Automated Screening Module - CervCancerScreening

José Paulo Soares Ferreira

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Luís Teixeira

Supervisor: Maria João Vasconcelos

Supervisor: Lucília Pinheiro

June, 2016

Cervical Cancer Automated Screening Module - CervCancerScreening

José Paulo Soares Ferreira

Mestrado Integrado em Engenharia Informática e Computação

June, 2016

Abstract

Despite all the medical advances, cancer remains one of the leading causes of death all over the world. Many are fatal but many others can be treated. Even though there are drugs and a broad panoply of treatments, most of the deaths are caused by the lack of early diagnosis and, sometimes, even the lack of information. This kind of problems escalate even more in developing countries, for obvious reasons, as is the case of cervical cancer. Cervical cancer is a type of cancer that if diagnosed early on can be treated and non-fatal. In developing countries, like African countries, cervical cancer is the second leading cause of female deaths. Since most of these women do not have gynecological surveillance, there are higher chances that the cancer grows to a state that can no longer be treated. Also, most of the equipment that could screen this kind of cancer is not within the reach of developing countries.

Taking all this in consideration, this project aims to create an innovative and, at the same time, low-cost solution to respond to these current issues. This tool aims to pose as a fast and trustworthy alternative and at the same time being cost accessible to developing countries, compared to the already marketed solution. The tool would be used to assist the medical staff in order to achieve a faster first overview of the patient's cervical cancer situation. Nevertheless, the ultimate decision comes from a doctor.

This solution attempts to fulfill these goals by deploying an Android tool which combines an image processing module with a machine learning one. In the first module, the application aims to enhance, detect and segment all the valuable objects within a sample, gathered during a pap-smear or thin-prep exam. These valuable objects are the cells nuclei. By extracting their features information, it is possible to make decisions about their nature. Foremost, during the machine learning module, a Nuclei/Non-Nuclei classifier is deployed, so the non-nuclei objects are filtered from the remaining. With sensitivities higher than 70%, this classifier can successfully split the objects into these two groups. Then, the nuclei are classified regarding their morphology, which will provide evidence about the existence of cervical cancer. Even thought this classifier is extra sensitive to any morphology's abnormalities, the sensitivity rate is higher than the previous classifier, being above the 75%. The final step of the deployment of this tool was to port this implementation into an Android application.

Resumo

Apesar de todos os avanços médicos, o cancro continua a ser uma das maiores causas de morte por todo o Mundo. Embora exista medicação e uma vasta panóplia de tratamentos, a maior parte das mortes ocorre devido ao diagnóstico tardio ou por falta de informação. Este tipo de problemas é mais frequente em países em vias de desenvolvimento. No entanto, se diagnosticado atempadamente, o cancro do colo do útero pode ser tratado e não ser fatal. Nos países em desenvolvimento, como em vários países Africanos, este tipo de cancro é o segundo principal responsável pela morte das mulheres.

Geralmente o entrave para um acompanhamento ginecológico regular nestes países é a falta de equipamento adequado. Tendo em consideração os factos anteriores, o projecto desenvolvido pretende suprimir estes problemas criando uma solução inovadora e, ao mesmo tempo, de baixo custo. O objectivo desta ferramenta é que seja acessível aos países em desenvolvimento, quando comparada com as soluções no mercado, mas ao mesmo que forneça resultados fiáveis. Esta ferramenta poderá ser usada de forma a ser uma ajuda para os médicos conseguirem dar uma resposta rápida sobre a situação actual do paciente, relativamente ao cancro do colo do útero.

A solução tenta satisfazer estes objectivos aliando, numa só aplicação Android, o processamento de imagem ao machine learning. No primeiro módulo o objectivo é melhorar, detectar e segmentar todos os objectos, contidos nas amostras, que sejam relevantes para a detecção do cancro do colo do útero. Estes objectos, neste caso, são os núcleos das células das amostras recolhidas durante os exames ginecológicos. Através da extracção da sua informação, é possível tomar decisões relativamente à natureza dos núcleos. Primeiramente, no módulo de machine learning, é implementado um classificador que avalia cada objecto quanto à sua probabilidade de ser núcleo ou não. Este classificador é particularmente importante pois filtrará muitos dos objectos que não representam qualquer utilidade para a aplicação, tendo uma sensibilidade de mais de 70%. Posteriormente, os restantes objectos são classificados quanto à sua morfologia. O resultado desta classificação revelará a presença, ou não, de cancro do colo do útero. Apesar de este classificador ser bastante sensível a qualquer anomalia dos núcleos, a sensibilidade do classificador é ainda maior que o anterior, atingindo valores de mais de 75%. Como esta aplicação não foi criada originalmente para Android, o último passo foi fazer esta transacção entre plataformas.

Acknowledgements

Firstly, I would like to thank my Father, Mother and Brother to the unconditional support they gave me throughout my whole life, including these crucial five years. Even though, sometimes, it seemed impossible to believe in my success, they always did and consequently made this journey a huge success.

Also, my girlfriend by making this journey to be even more successful, who always put up with my anguishes and helped me during this whole project. My friends who also made a huge part of this whole experience.

Lastly, I would like to thank my supervisors Luís Teixeira, Maria Vasconcelos and Lucília Pinheiro for answering all my e-mails and giving me all the support I needed to develop this project successfully.

José Paulo Soares Ferreira

“Cap ou pas Cap?”

Jeux d’enfants (2003)

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

Fraunhofer AICOS	Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions
HPV	Human Papillomavirus
HSIL	High-grade squamous intraepithelial lesion
SVM	Support Vector Machines
TCGFE	Texture-Color-Geometry Feature Extraction
GPS	Global Positioning System
OpenCV	Open Source Computer Vision
OS	Operating System
ROI	Region of Interest
Adaboost	Adaptive Boosting
XML	Extensible Markup Language
NDK	Native Development Kit

Chapter 1

Introduction

1.1 Context

Nowadays every aspect in our life is affected by technology, everything we use as some kind of it. Related to that situation, increasingly more fields start to be affected by it, even medicine. Due to this "invasion", there is a need to develop solutions to help this very important field to get up-to-date and discard old methods and practices that can be replaced by something new and technological. One of the goals of this project is exactly, to somehow help the medicine field to get improved. Since this is one of the main concerns, this project is being developed with the Fraunhofer AICOS (Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions). This center is located in the UPTEC building, in Porto.

As the name may indicate, the main goal of this association is to improve the living standards of people in general by creating intuitive, useful and most of all, innovative solutions. A result derived from that goal is the concern to progressively introduce technology in the everyday life of those who are not used to it, like the senior citizens and those who live in low income countries. Like mentioned before, Fraunhofer gives special attention to the medicine field, always trying to create solutions to help, mainly, developing countries. This concern comes from the lack of all kinds of wherewithal of those countries. What Fraunhofer AICOS tries to do is to ease up the job of the doctors that work in those countries by creating new tools that perform the screening of diseases that affect them. Related to this problem, the project that will be developed will help the screening of the Cervical Cancer.

1.2 Project

The developed tool aims to provide a trustful and low-cost solution to identify and screen cervical cancer. This tool will screen liquid based cytology samples that were collected during gynecological exams, like pap-smear and thin-prep test. Although these exams and the developed tool have the same goal (i.e. to screen the cervical cancer), the difference between them is that this project aims to facilitate the doctor's time consuming work. This will be achieved by highlighting all the objects in the sample that show a high risk of indicating cervical cancer. This way, the screener just needs to look at those objects in order to do some kind of diagnose. Since the screener needs to evaluate several images just to conclude about one sample, by using this tool this process would take considerable less time to perform. At the same time there would be fewer errors, due to the fact that this would be less exhausting compared to the old techniques where no objects were highlighted. This project, as already hinted, will be using several image processing methods including a C++ library called OpenCV. This library is a well established one, including several known and important algorithms which are used in this kind of projects. Firstly, in order to achieve better results in later stages, the image will be enhanced so all the relevant attributes of it are improved for the segmentation phase. Then all the features which were augmented are extracted in order to be used in the machine learning module. This module will be split into two different classifiers, one followed by the other. The last of the classifiers will be the one that will give the final results regarding the cervical cancer screening process.

Even though it seems like this project will try to replace the doctor's expertise, its only and main goal is to facilitate and help to perform this tiring and difficult job. With this tool a first screening of the cervical cancer could be done and be submitted for final evaluation by a surgical pathologist.

1.3 Motivation and Objectives

One of the motivations and objectives that this project has, is ease up the doctor's work during the cervical cancer screening. Still, the main objective is to create a low-cost solution in order to be used to help low income countries. Almost every marketed solution, as we will see during this document, have high selling prices, making impossible for these countries to acquire them. Although the cervical cancer is not deathly in developed countries, given its treatable nature, in the developing countries this is one of the leading causes of death [CTUA14]. That happens due to the lack of early diagnosis and gynecological screening solutions. Consequently, the women only find out that they have cervical cancer already in later stages, being impossible to cure. These reasons create the need of producing a low-cost solution so the women of these countries can also have gynecological supervision. Since most of the times the samples gathered in developing countries need to be sent to another country that has the specific equipment, with this solution this need would be almost suppressed. This way only the samples from the exams that revealed a high risk of having cervical cancer would be sent. Money and time would be spared by doing this, two things of which these countries have shortage of, in these situations.

1.4 Structure and Organization of the Document

Besides this introduction, this document has more 6 chapters which will approach every aspect of the development of this project.

The chapter 2 main goal is to provide a background context about the problem we are trying to tackle. This means that we will talk about the virus that can cause cervical cancer and the cancer itself. Some of the anatomy where this type of cancer appears will be shown in order to define some concepts referred during the chapter. Lastly, we will talk about how the cervical cancer is identified by the screeners.

Afterward, in the chapter 3, all the technological concepts will be talked about. Not only the concepts but some of the algorithms that are used during the development. Firstly, since a big part of this project revolves around image processing, we talk about what is an image, a digital one, and some of its processing techniques. The image processing will be split into 3 different layers, which each one will have their own techniques and algorithms to treat their problems. In the end the languages and operating system that the project was developed with will also be approached, as well as the marked solution regarding the screening of the cervical cancer.

In the chapter 4, it will be described each stage of the implementation of the project. The first, the segmentation, where is going to be talked about the experiments made during its implementation. These experiments, as will be described, were the kick-starter for the actual implementation of it, sub-chapter that will follow the one of the experiments. Then the dataset used will be fully described, as an introduction to the next phase which will be the machine learning one. About it, the experiments and algorithms used will be shown in order to justify its implementation.

In the chapter 5, we will explore deeply how this tool was ported to the, already developed, Android Application. Some images with use cases will also be shown.

Finally, in the chapter 6, a conclusion will be made about this whole document and about all the work done. Some possibilities and guidelines about future work, regarding some unsolved or encountered issues during this implementation, will also be presented.

Chapter 2

Background Context

This chapter has the goal of letting know some relevant concepts about this project. For a better understanding of the problem it is crucial to talk about the virus that can cause the cervical cancer and, obviously, about the cervical cancer itself. We will go deeply into the stages of the cervical cancer and the commonly used exam for screening this kind of cancer, the Papanicolaou Test. Lastly, some sample images will be shown in order to explain how the cervical cancer is detected with optical microscopy.

2.1 Human Papillomavirus

2.1.1 What is it?

In order to understand the magnitude of the Human Papilloma Virus and how it can cause cervical cancer we should start from the beginning, from which family this kind of virus comes from. The Papillomaviridae family has a lot of strains associated to it. The causes of being infected by one of this virus can go from to be asymptomatic, where the carrier does not actually experience anything, to actually have some effects on the carrier. Although there are some strains that just give some small benign tumors (also called papillomas or warts), which do not represent a death risk but can be highly infectious, there are a lot of them that can lead to many kinds of cancer. Those strains are the ones that were not resolved spontaneously by the body and persisted through time, because normally the immune system can kill the virus by itself [HRDP06]. As stated before there are some strains that can affect Humans, these are called Human Papilloma Virus. Since these virus are easily spread it is very common for either men or women to get infected by one strain through their lifetime. The spreading of this virus it is, mainly, made by sexual contact between persons, even if the person does not seem infected by it. Getting infected with this virus can occur in many ways but the most common way is by sexual contact, so this is the way who gets more attention in terms of prevention. Being infected with HPV can also occur by blood transfusions and during labor. Another way to get infected by it is by sharing objects with a person with warts.

2.1.2 General Symptoms

Usually, being infected by the HPV does not carry any risks because it never gets to have an actually effect in the person or it resolves itself, but there are some strains that can lead to cancer, namely types 16, 18, 31 and 45 which are the most critical ones that a person can be infected with. Besides the actual cancer, the most common manifestations of the disease are the warts and papillomas. Usually these warts appear in the genital area but they can also appear in the throat, lips, mouth and tongue. There are several types of warts: common, plantar, genital and the flat ones [Hpv]. Common warts usually appear on the hands, fingers and elbows and look like rough, raised bumps. Plantar warts just appear on the feet of the carrier, which look like as hard, grainy growths. Genital warts, as the name says, appear on the genital area, which can look like a small bump, cluster of bumps or even stem-like protrusions. These are highly infectious during intercourse. Mostly, the ones referred before affect adults but the flat warts generally affect younger people. Commonly they are found on the face, neck or areas who have been scratched and look like lesions which are darker than the normal skin color.The other kind of manifestation of the Human Papilloma Virus is more deathly, which is the actual cancer. These manifestations only occur when the HPV is not threatened early on, which takes years to actually develop into cancer. If not diagnosed the cancer can only be noticed in a very late stage, when the treatment is very difficult. Since the HPV usually spreads across people through sexual contact, also these cancers affect where people make those contacts, namely in the cervix, vulva, vagina, penis, anus and oropharynx (the last two, both on men and women).

2.1.3 Prevention

Since the HPV is mostly transmitted by sexual contact, the best way to prevent to get infected with this disease is to abstain of any sexual activity, including oral, anal, and vaginal sex. Obviously this is not a realistic option for most of the adults so there are other ways to prevent this disease. Another way is to get vaccinated against HPV, although the person is only protected against two types of high risk HPV. To improve the protection given by the vaccine, it is recommended for either the girls or boys who take it to be vaccinated before becoming sexually active. Specifically for the men, being circumcised is a good way to prevent being infected with this virus since clinicians say there is a lower risk of the virus to be housed in the foreskin of the penis. Medically related there is not much more a person can do, so it is all about the person lifestyle. Using condom during the intercourse it is highly recommended since it lowers the contact in critical areas. Even using condom a person can still get infected since the bodies still touch themselves, but it is not as critical as not using condoms. Waiting to have a sexual relationship until a person is older is also a kind of prevention since starting early exposes the person to more partners increasing the risk of getting in contact with someone who has this disease. This raises another problem that is even if a person does not start early her sexual life, having a high amount of sexual partners can be also very risky. There are studies that show that if a person wait at least eight months before having sex it can reduce the probability of getting infected since any HPV infection that is present can disappear in that amount of time.

2.1.4 Diagnosis

Diagnosis is a very important aspect of this disease since deaths only occur if the Human Papilloma virus does not get noticed before turning into cancer. The prevention methods cited above are the best way to not get infected with this disease but if in fact a person get it is advised to do regular tests, so the HPV does not evolve into a state that can no longer be treated. This kind of problems happens where the equipment or the resources are not easily obtained, like in developing countries. When, in fact, there is equipment capable to test if there is any trace of the HPV, the death rate is very low compared with the developing countries.

Women continue to be more affected by this virus than men, but for either one there are some tests that can be performed. Papanicolaou smear is the most common one. Although Pap smear test is always associated with women, it can also be performed on men. What happens is that the clinician collects a sample of cells from the cervix or anal canal of the woman, or on another hand the anal canal of the man. Then the clinician analyses the images of the sample and check for cellular abnormalities on it. There is also the HPV DNA test, that shows, by molecular analysis, the type of virus strain present in the sample. If in either of the exams the result is positive, and it is a strain that can lead to cancer, the next step is to perform a colposcopy, in case there are some traces in either the vulva, vagina or cervix, or an anoscopy, in case there are some traces in the anal canal. These exams are performed by cutting a small tissue sample and then analyzed further. The DNA test is most of the times performed on older women since the immune system of younger ones is stronger and most of the times can clear the HPV alone. In addition to all these tests, there is also the vinegar (acetic acid) solution test. This helps to identify if there is any cervical lesions that are not visible since the skin turns into a white color in the infected zone.

2.2 Cervical Cancer

2.2.1 What is it and how it is related to the HPV?

As stated before, the Human Papillomavirus can be the leading cause of many cancers. More concretely the HPV is the biggest responsible of the Cervical Cancer, since it causes 99% of them. Firstly, a cancer is a disease where happens an unusual growth of the human cell, so the Cervical Cancer is a disease where cells grow out of control in the cervix. The cervix is the lower part of the uterus and is the connection between the uterus and the vagina. This cancer is one of the most common among women, and the seventh overall, with an estimated 530 000 cases per year, which 85% of them occur in developing countries [SGKC+13]. Because of this there is an increasing need of information in this kind of countries, since this is a deathly cancer if not diagnosed early on.

2.2.2 Symptoms

One of the biggest problems about this cancer is the absence of symptoms in the early stages of this disease. This gives room for the cancer to develop into critical stages, where the symptoms get noticed. Usually in the beginning the first symptom is unusual bleeding from the vagina and usually it occurs after sexual intercourse. In some cases there is a cervical mass which indicates the presence of a growing cervical cancer. Also, if there is some kind of discomfort during sex and unpleasant or unusual vaginal discharge, it can suggest that a person has cervical cancer in an early stage. If there is vaginal bleeding before or after the expected monthly time, it can also be a red flag. When the cervical cancer gets in advanced states the symptoms get more noticeable and dangerous for the carrier. The body gets weaker and symptoms like tiredness, lack of energy, weight loss and loss of appetite appear. Since the vaginal bleeding is common in the reproductive system of the women, it does not mean that a person has cervical cancer. Although it is advised to schedule a session with a doctor if a person thinks it is an unusual bleeding.

2.2.3 Stages

Before talking about the cervical cancer stages, the vagina's anatomy is shown for a better understanding of the concepts talked about. The most important part, related to this project, is obviously the cervix where the cancer is firstly developed and then proliferated to the rest of the vagina and sometimes, even, the rest of the body.

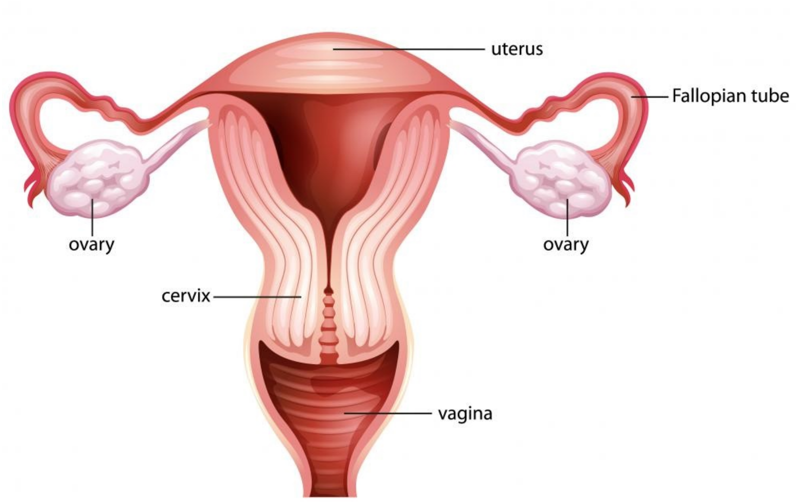
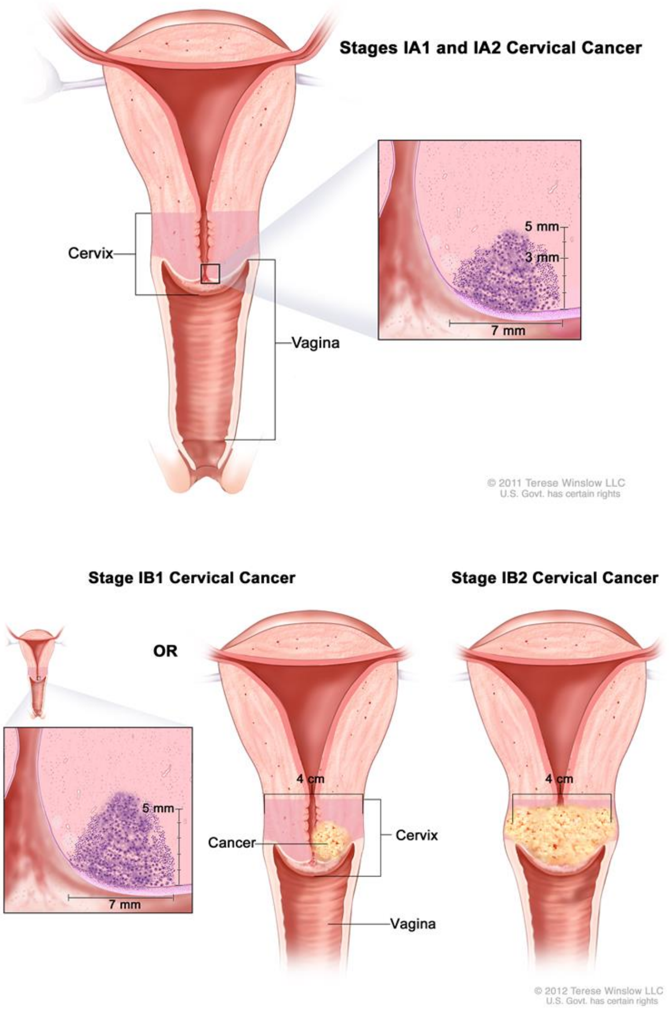


Figure 2.1: Vagina Anatomy

The cervical cancer has five main anatomic stages, each one with their specific symptoms and characteristics. The first stage is the stage 0 or the Carcinoma in situ. In this stage some abnormal cells get noticed in the cervix. These cells appear in the cervical epithelium and, at that stage, there is no invasive cancer.

The next stage, stage I, the cancer is already developed but can only be found in the cervix area. This stage can be divided in two sub-stages: stage I A and stage I B. In the stage I A the cancer can only be observed by microscope in the cervix tissues. This stage also has two sub-stages which are based on the size of the tumor. In the stage I B the cancer is bigger and in some sub-stages can be observed without microscope. Again these stages are based on the size, but in bigger dimensions than the previous one.

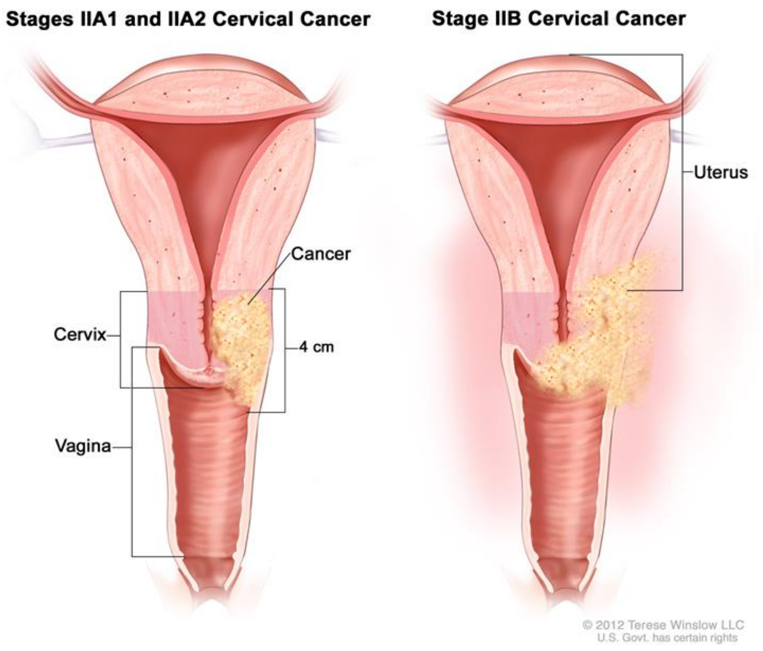

Stages IA1 and IA2 Cervical Cancer
Cervix
Vagina
5 mm
3 mm
7 mm
Stage IB1 Cervical Cancer
Stage IB2 Cervical Cancer
OR
5 mm
7 mm
4 cm
Cancer
Cervix
Vagina
4 cm

© 2011 Terese Winslow LLC
U.S. Govt. has certain rights

© 2012 Terese Winslow LLC
U.S. Govt. has certain rights

Figure 2.2: Stage I A and I B of the cervical cancer

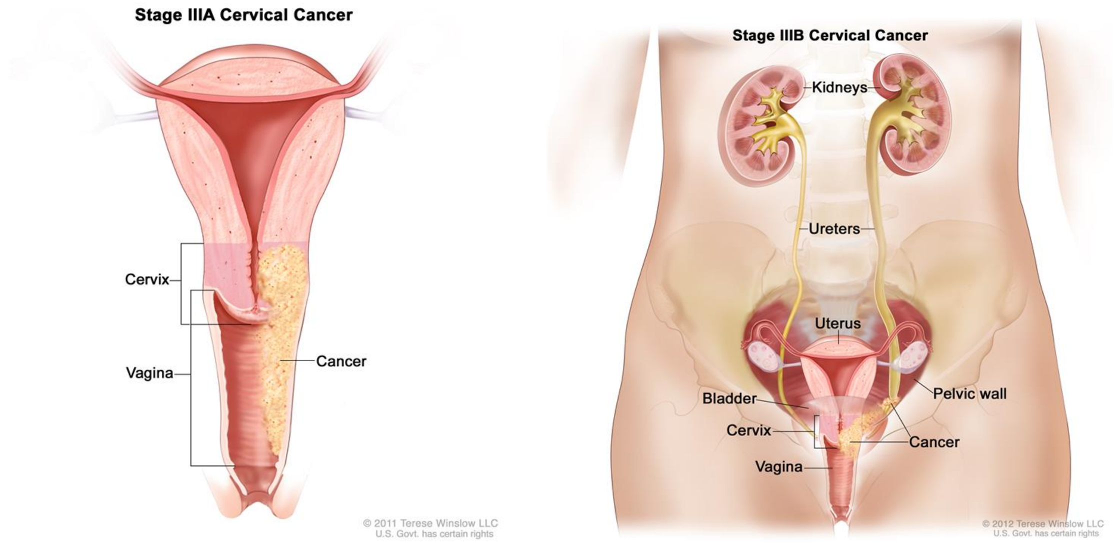
In stage II the cancer keeps spreading and gets beyond the uterus but not to pelvic area. Also in this stage the cancer did not spread to the lower parts of the vagina. Based on how the cancer has spread there is two different stages: stage II A and stage II B. The cervical cancer in the first one did not spread to the tissues around the woman uterus. In the second the cancer has spread beyond the cervix and to the tissues around the uterus. Compared to stage I, the cancer can be seen without microscope on both stages, which means that the cancer keeps getting bigger.


Stages IIA1 and IIA2 Cervical Cancer
Stage IIB Cervical Cancer
Cancer
Cervix
4 cm
Vagina
Uterus

© 2012 Terese Winslow LLC
U.S. Govt. has certain rights

Figure 2.3: Stage II A and II B of the cervical cancer

Stage III reflect involvement of other body organs besides the ones in the reproductive system. As well in this stage there are two different sub ones based on the spreading of the cancer. In the stage III A the cancer has spread to the remaining part of the vagina but not into the pelvic area. Stage III B is the one that can affect the kidneys blocking one or both ureters, swelling them or even make them stop. Also in this stage the pelvic area can also be affected.



© 2011 Terese Winslow LLC
U.S. Govt. has certain rights

© 2012 Terese Winslow LLC
U.S. Govt. has certain rights

Figure 2.4: Stage III A and III B of the cervical cancer

Lastly, in the stage IV the cancer gets so big and dangerous that can affect all the human body. Both sub-stages are based on where the cancer as reached. Stage IV A means that the cancer has spread into all the vagina and started to spread into nearby organs, like kidneys, bladder and rectum. Stage IV B means that the cancer has spread to any other organ, even lungs and bones.

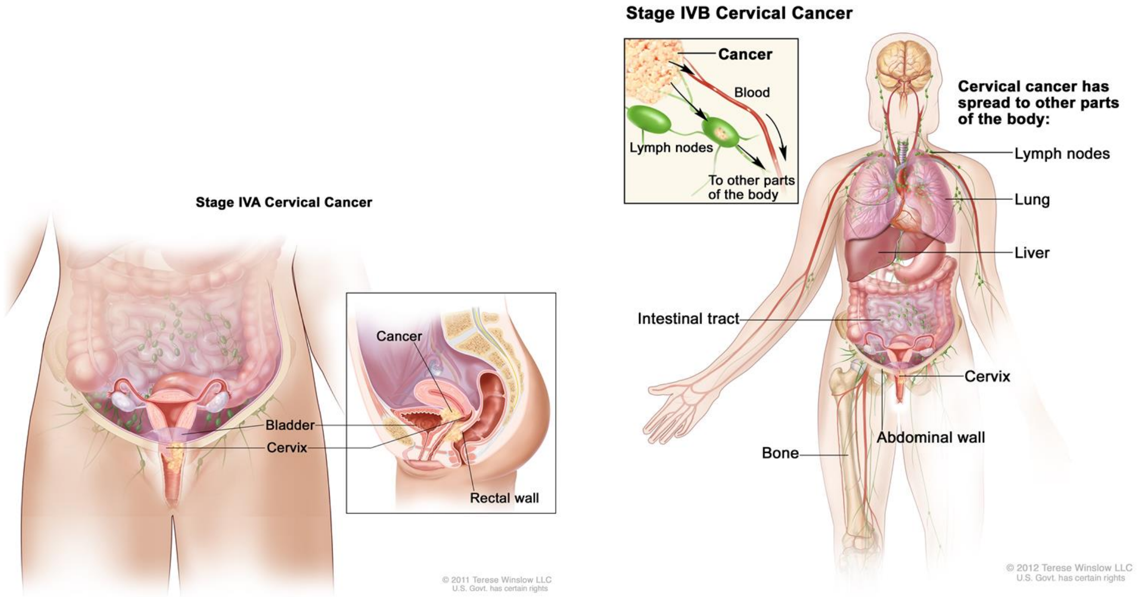

Stage IVA Cervical Cancer
Cancer
Bladder
Cervix
Rectal wall
© 2011 Terese Winslow LLC
U.S. Govt. has certain rights
Stage IVB Cervical Cancer
Cancer
Blood
Lymph nodes
To other parts of the body
Cervical cancer has spread to other parts of the body:
Lymph nodes
Lung
Liver
Intestinal tract
Cervix
Abdominal wall
Bone
© 2012 Terese Winslow LLC
U.S. Govt. has certain rights

Figure 2.5: Stage IV A and IV B of the cervical cancer

2.2.4 Treatment

Treating the cervical cancer can be possible in most of the stages besides the last one if it is on the stage IV B, where the cancer has spread all over the human body. In stage 0, since the cancer is not invasive there are many ways to treat and remove the neoplastic tissues. Every method is based on removing the tissues by cryosurgery, laser surgery or by loop electrosurgical excision procedure. For the next stages there is a big down-side which is deciding if whether or not a woman wants to continue to be fertile. Before the stage I B, the cancer can be treated without recurring to chemotherapy or radiotherapy, so in stage I A1 and I A2 if the person wants to keep the possibility to have children either chooses to perform a radical trachelectomy with removal of pelvic lymph nodes or a cone biopsy to remove the cancer, if it is small [Can]. Beyond these two stages, as mentioned before there is a need of using more invasive methods since the cancer is in a later stage. Besides the radio and chemotherapy, the standard treatment is a radical hysterectomy with removal of lymph nodes in the pelvis. In even later stages the recommended treatment is the chemoradiation and radioterapy since it is more invasive than the ones mentioned before. Lastly, and as mentioned in the beginning, the stage IV B is untreatable since the cancer is far too spread but for relieve reasons the radiation therapy can be performed.

2.3 Papanicolaou Test

The Papanicolaou Test, or Pap Test, is as mentioned before the main screening method of the cervix. The goal of this test is trying to find pre-cancerous or even cancerous cells in the cervix, that can lead to the cervical cancer. If there is any trace of cancerous cells the person needs to take further exams to know in which stage the cervical cancer is, since she needs treatment. Is said that performing this kind of exam reduced 80% of the deaths that could be caused by it.

To take the cervical sample for cytological analysis, the woman lays down in a gynecological bed and put her feet in the stirrups in the way that the woman as her legs open. Then the clinician places, carefully, a speculum inside the vagina, just to open it a little, in order to be possible to take a look inside, namely to see the vagina and the actual cervix. This way is also possible to scrap a little of the cervix to take the said sample. Although the clinician can actually see if there is cancer or not, if the cancer is well developed, what will give more insides about the patient situation is the sample taken from the cervix. When the sample is taken, a surgical pathologist will take a look at it on an optical microscope in order to see if there are any traces of atypical cells or carcinomas. Most of the times, this last step is done manually which takes longer than using an automatic machine. Sometimes the cervical cytology is done on a liquid base and, in this case, a HPV DNA test can be performed. There is no extra work taking another sample, because the same sample that is used to screen cervical cancer can be used to test HPV. Sometimes there are false negatives in this kind of test since it is basically done "by eye" and because the cervical cancer is takes a very long time to develop. Regarding the sample acquirement, there is a liquid based cytology test called ThinPrep pap test, which will be the ones that will provide the samples to this project. With this method, the samples will be easier to read, and minimizes obscuring blood, mucus, and non-diagnostic debris, enabling increased accuracy for both manual assessment and computerized assessment of the cells [LCy].

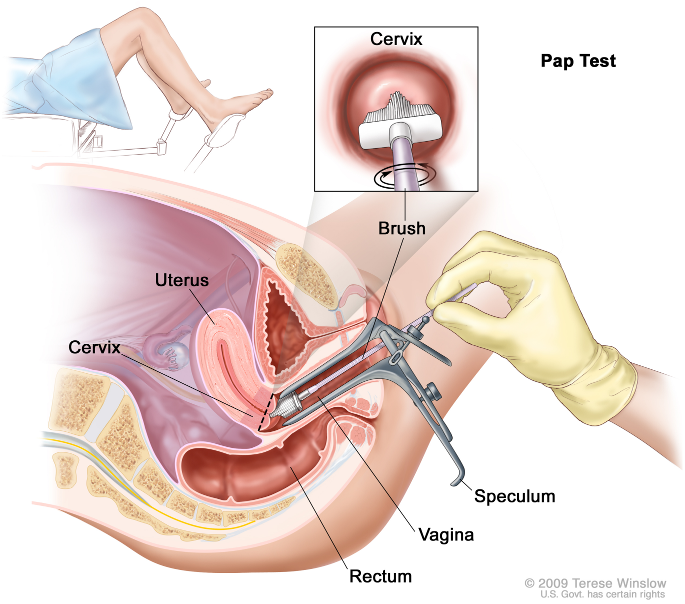


Figure 2.6: How a Papanicolaou test is performed

All women are advised to have a gynecological observation with the assessment of a cervical cytology, either a Pap smear or a liquid based exam, as soon as the active sexual life begins. Even if the women take the HPV vaccine they should continue to do this kind of exam since that vaccine does not cover all types of cervical cancer. After the first exam has took place, the person should go regularly to a gynecological observation for a cervical cytology [Pap]. An exception to these advise, are people that made a Hysterectomy. This procedure is the removal of the uterus, cervix, ovaries, fallopian tubes and other surrounding structures. Naturally, if a person does not have the cervix, there is no need to test for cervical cancer. Also, women who are above 65 years old and did not have a positive test in the last 10 years can stop taking this exam.

2.4 Cervical cancer identification

The cervical cancer screening is something that do not exist, in some countries, or that takes to long and it is deprecated. In order to analyze the gathered sample during the Papanicolaou or liquid based cytology, the clinician uses an optical microscope to visualize the sample and perform the manual inspection. During this procedure, the clinician must constantly unfocus and focus the image in order to be able to see the upper and lower layer of the sample. The aforementioned must happen so the clinician can visualize all the details from every cell. This procedure is done to multiple images, from the same sample, so there is a high confidence level about the existence, or not, of abnormal cervical cells.

There are some factors that can spoil the sample evaluation and make them impossible to gather information. If 75% of the cells are obscured by blood or by inflammatory cells, the sample does not satisfy the minimum conditions for being able to be screened. In the case that the percentage is lower than 75% but higher than 50%, that condition must be taken into account since that sample can give wrong results. Another factor that must be taken into account is the number of cells in the sample. The minimum admissible criteria is approximately 150 cells in a 4x microscopic field. A liquid based cytology should have an estimated minimum of at least 5000 well-preserved squamous cells in the whole sample (7-9 cells in a 40x microscopic field).

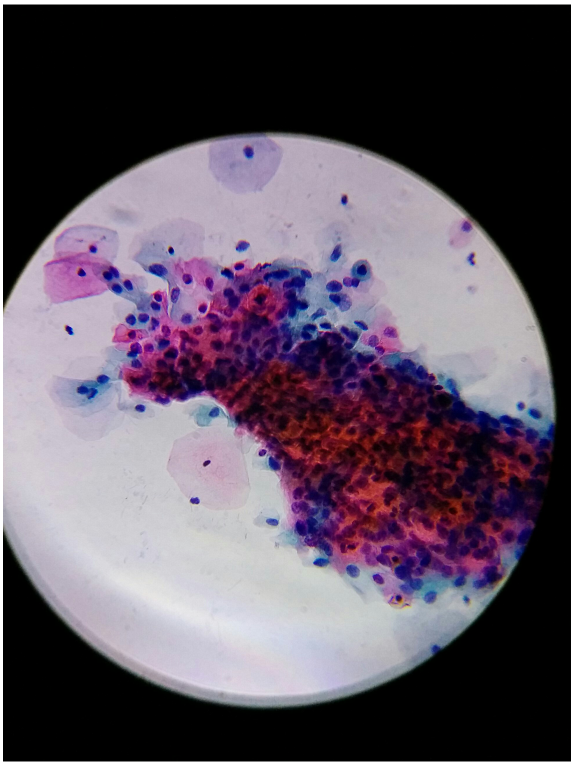


Figure 2.7: Image of the objects that can exist in a gynecological test sample

Regarding the cell identification itself we can observe four types of cells in the samples. Firstly, the superficial and intermediate squamous epithelial cells are the most noticeable in all the images taken from samples. On the figure 2.7, we can identify these four kinds. Starting with the smaller objects, we have the scattered blue ones. These blue objects are the inflammatory cells, which are very common in any sample gathered during cervical sampling. These kind of cells do not give any information about the cervical cancer existence, so they are ignored during the screening process. Then we can notice the big group of cells, which are several endocervical epithelial cells gathered in one place.

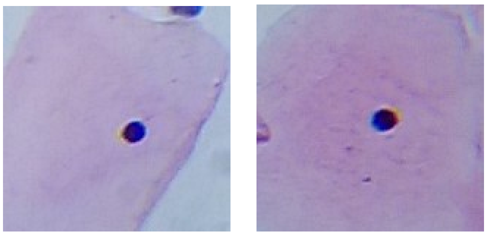


Figure 2.8: Two images that represent two normal cell's nuclei

The big blue ones represent the first stage of the squamous epithelial cells, which are called basal and intermediate squamous epithelial cells. These cells, while growing up, turn themselves into the pink ones, which are the superficial squamous epithelial cells. Since the blue ones are younger than the pink, usually their nuclei are bigger compared to the subsequent. This is an important aspect to emphasize because, one of the aspects to pay attention when looking for cells abnormalities is the size of the nuclei. So, even though, sometimes the blue cells have big nuclei that does not mean cervical cancer.

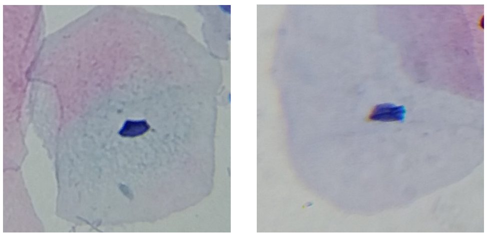


Figure 2.9: Two images that represent two abnormal cell's nuclei

Other aspects that can give away the existence of cancer are the morphology of the nuclei and their color. On the figure 2.8 we can have an idea of what is a normal nucleus. A normal nucleus is characterized by being almost a perfect circle, without any morphological abnormalities or being extremely big. Also, these two nuclei do not have a strange color. On the other hand the nuclei on the figure 2.9 represent two abnormal nuclei. Compared to the first two nuclei, it is visible the different morphology they have. Instead of having an uniform and circular aspect they look like a trapezium, with a lot of irregularities in its morphology. Also if a nucleus display a very dark and unusual color, compared to the ones presented, there are strong possibilities that the cell is abnormal. Although these kind of lesions indict atypical cervical cells, they do not represent high grade ones. Instead, they are characterized by, besides the aspects mentioned before, a low quantity of cytoplasm. An example of this kind of lesion is represented on the figure 2.10.

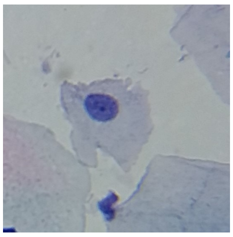


Figure 2.10: Image representing a high grade lesion

Accordingly with what was said in the beginning of this section, it is noticeable why the screening by eye detection is prone to errors. In order to identify cervical cancer, the surgical pathologists must pay attention to all the details of each and every nuclei. Obviously there are many kinds of lesions, much more perceptible than the ones shown. Nevertheless, as shown in the figures 2.9 and 2.10, the clinicians need to have a tremendous level of concentration in order to not neglect these less distinct ones.

Chapter 3

Literature Review

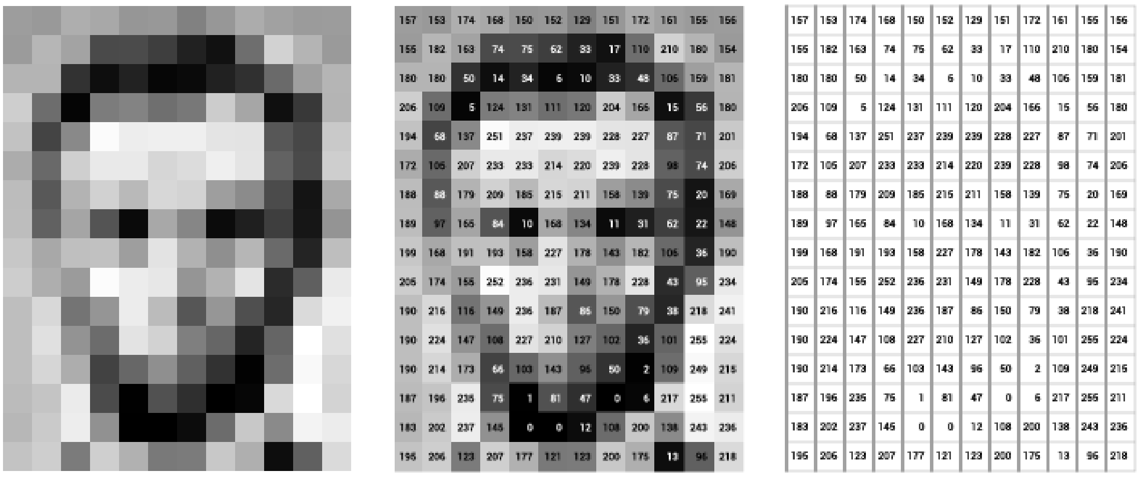
The goal of this chapter is, after given the background details of this project, to explain what will happen technologically wise. Since this project is based on images, a short introduction about what is a digital image will be given. Then we will go thoroughly about the image processing, which is split into two different steps. For both of them we will explain their utility as well as some of the important and most used techniques and algorithms. Whereas that image processing is not enough for learning anything about the images, we then will cover the machine learning topic. In the end, we will be reviewing some of the technologies used as well a section about related work in the field.

3.1 Image Processing

3.1.1 Digital Image

By definition an image, in computer science standards, is an optically formed duplicate or other reproduction of an object formed by a lens or mirror. A digital image that is stored in a device is a rectangular array composed by elements called pixels. The array is organized as a matrix so it has columns and rows. The size of this matrix will be the size of the image. Another important aspect of a digital image is its resolution, that is the spatial scale of the image pixels. The resolution of an image can be measured, mainly, by ppi and dpi, that are pixel per inch and dots per inch respectively. The first represents the quality of the image on a screen and the other when printed.

Only with what we have defined we do not have an actual image because just by only defining the pixel array and the resolution we would only have a shape. The true essence about this is the intensity value that each pixel has, creating the image itself, as we may see in the image 3.1. If all pixels have the same value we create an uniform image, an image with only one color. On the other hand when we have pixels with different intensity values we create a colorful image or a black and white image. The black and white images, also known as grayscale images, has its pixels taking values from the darkest gray, which is black, to the lightest gray, which is white. Colorful images have its pixels taking the darkest and lightest from the three main colors Red, Green and Blue. Colorful images are also known as RGB images, which is the name commonly used in the image processing field.



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Figure 3.1: Representation of a digital image

3.1.2 Defining Image Processing

Image processing takes a big share in the development of this project. Image processing is the method of applying mathematical operations to it and change some of its characteristics. There are two kinds of image processing, analog and digital, but regarding this project we will explore mostly the digital variant.

We are more concern on defining digital image processing than the first one. As the time went by, the digital cameras and processing took over the analog ancestors due to theirs wider range of applications. This new processing method acts directly over the digital images, applying algorithms and operations which can successfully change an image. Digital image processing is not always related to complex operations. The first steps on this new field were enhancing the light of an image, which is the most commonly used operation. Together with the all the computer's technology breakthrough, the image processing evolved into a more complex and useful field. Due to this whole evolution, the digital image processing allied itself with the machine learning field in order to learn about the world from information gathered from images. So, besides the machine learning, this field can be split into two stages: the pre-processing and the feature extraction, or segmentation. Machine learning can also be considered one of the steps of machine learning but it will be treated as a different topic, ahead in the document.

First off, the main goal of the pre-processing is to increase the reliability of the feature extraction stage. Many algorithms and techniques can be applied in order to enhance or reduce image details. Then, after all the pre-processing is done we can start the feature extraction. Also in this step we have several algorithms to achieve the desired results. Using the enhanced image from the last step, we segment the image and extract all the characteristics that will give valuable information for the machine learning process. Then, finally, when all the information extraction is done we will start the machine learning. Both of the steps, regarding the image processing, will be explained up next in detail.

3.1.3 Pre-Processing

As mentioned before, this will be the first step and one of the main pillars of the image processing. During this first stage we will enhance and omit details from the image. The output will be another one with some of the object's characteristics emphasized. This will be useful in the segmentation phase so we can easily identify what we want, and also in the feature extraction and machine learning [low].

There are several pre-processing techniques, which involve enhancement as well as noise reduction methods. As we will see, some of the noise reduction techniques also enhance some of the desired objects characteristics. Enhancing an object can simply be described as highlighting the said object. Unlike the former, noise reduction can be much more complex. Reduction the noise from an image is one of the goals of this stage. By definition, noise means unwanted signals in the information we are receiving, either in sounds or images. In the image processing field, noise can be defined by the existence of pixels that possess brightness or color values that do not correspond to the reality. Noise can also be objects that simply were not supposed to be there.

Therefore, both of this set of methods are applied in order to suppress or attenuate some of the problems images have. Some of these methods are going to be explained in detail up next.

3.1.3.1 Image Enhancement Algorithms and Techniques

The main goal of these techniques is to make some of the objects stand-out compared to others. This is particularly useful for the segmentation phase since this way we can easily identify the objects to extract.

Regarding the actual algorithms techniques, one that is vastly used and trustful is the Mean Shift Filtering. This technique tries to homogenize the color of a certain local in an image, taking in account the color values of the surrounding pixels. What happen is that all the shades that an object could have will be dismissed and the color will be the mean of the most dominant values [PLP09]. Considering that there is an actual function on the OpenCV library for implementing this algorithm, this technique is easily applied and so, in the case of this project, commonly used, due to the cells color features.

Another set of techniques that has great importance in the field of the image processing are the edges detection algorithms. These methods aim to detect, as the name says, the points which delineates the edges of an object. These methods calculate the results, based on the premise that the brightness changes sharply or discontinues in the edges. The most commonly used technique is the Canny Edge Detection. This first step of this method is to smooth the whole image, using a blurring algorithm. These kind algorithms and techniques will be explained in the next subsection. The reason for doing this is to suppress any tiny elements that can spoil the edge detection of an object. Then, the value that was referred before is calculated. This value is called intensity gradient. The next step is to suppress all the non-maximum values that the image matrix has. This must be done because, as said previously, the higher the numbers are, the higher is the probability that we are dealing with object's edges. Finally, a double threshold and a simple filter is applied in order to find the calculated edges. The filter will eliminate all the lines that are not connected with the others because they do not close, thus not representing an actual object.

3.1.3.2 Denoising Algorithms and Techniques

As the name points out, these algorithms aim to remove or attenuate the noise in an image. This is something that always should be done, unless the image is 100% clean, because even a tiny object can spoil the whole image processing. Even though there are a lot of algorithms and techniques, the ones that will be mentioned during the section are the ones that are commonly used.

Unlike the similar algorithm referred in the previous section, the non-local means algorithm calculate the new value of a pixel based on all the pixels in an image. This is a quite interesting algorithm since it could easily eliminate an object from an image if it is totally different from the rest. As said, this algorithm uses the mean of all image pixels and checks how close are they to each one of them, changing it accordingly. Comparing to local mean algorithms, this one achieve better results in terms of denoising instead of enhancement because it takes in account all the values, thus checking if an object is similar with most of the image content. This denoising technique is particularly useful when we really do not want to lose object details. Which is one important characteristic considering that we are trying to identify cells and theirs nuclei.

Another technique which is also focused on preserving objects details, is the total variation denoising. This kind of methods is particularly useful because all the characteristics of the desired objects are preserved. The total variation value is calculated based on how the frequency of an image changes. Thus, the premise of this algorithm is that when we are dealing with a very noisy image, we are also dealing with a high total variation value. The goal of this technique is to remove the pixels that are causing such high variation in order to stabilize it. Obviously this will not remove all the noise an image has but it can be used to complement another denoising technique.

Finally, and probably the must widely used techniques, the low-pass filtering. The basic idea behind this filtering is for each given pixel we calculate the mean of the eight surrounding pixels and give it that value. The result of implementation of one of these techniques is a blurry image. One of the must used low-pass filtering techniques is the Gaussian Blurring. Instead of calculating the new pixel value by the mean, the new value is calculated by a Gaussian function. An X and Y maximum distance is given to the function in order to use that amount of pixels to calculate the new value. The advantage of using this blurring method it because of the linear property it has, thus spending very few time computing the new values. The problem is that this way a lot of detail is lost during the application of this method, which sometimes it is a characteristic that is not possible to allow.

3.1.4 Segmentation and Feature Extraction

The segmentation and feature extraction, are highly dependent on the last stage of the image processing. Firstly, considering that we are going to segment an image, it must be known which objects, or regions, are going to be segmented. We know this by just segmenting the objects which are enhanced, that is why the previous step is very important. The main goal of this stage is extracting what is important in the image and not really understanding what were dealing with. The most common used method for object segmentation is the thresholding. As we will confirm during this document, the OpenCV library has many forms of thresholding in order to segment objects in an image. The most common one is the normal threshold. A threshold can be defined as turning a grayscale image into a binary one. Ideally, the objects we want to segment will be black and the background will be white. Some values can be defined in order to limit which objects will be recognized as such, like the intensity of its pixels. Using the same premise, thresholding by color is also possible. This method is used when we are dealing with objects that have clear and distinct colors. If this does not happen, the threshold will contain a lot of errors and the morphology of the objects will not be well defined.

Thresholding also has its own disadvantages. When an image has a lot of different brightness levels, the threshold is not computed very well due to the different pixels characteristics levels. To overcome this problem local or adaptive thresholding are applied. These methods crop the image into small pieces and apply the normal threshold into each one of them, so the brightness values of the pixels coincide with each other.

After we gather all the useful objects or details from the image, we can go through the next step which is the feature extraction. In order to understand about the image we must know the characteristics of the objects we enhanced and consequently segmented. It must be held in consideration that during the development of this project it was not used any special feature extraction techniques, due to the fact that an already developed library was used. This library, as will it also be approached later, it is called TCGFE and it was developed by Fraunhofer [RCEc16]. This library extracts exactly 152 characteristics regarding the objects Texture, Color and Geometry, all of which are related to this project. Also, as any other feature extraction technique, a mask of the object is fed and its original image so the algorithm knows the exact contours of the object we want to extract the features. This extraction will be used for later stages, namely the machine learning process. During this stage, it is also usual and common practice to count how many objects we are dealing with in order to be used, also in later stages, for performance calculations.

3.2 Machine Learning

This is the stage that will reflect all the work done with the image processing. If both of the steps from the previous module can accomplish a high performance, this one will too. The ultimate goal of the machine learning is to understand and find patterns in the extracted data during the segmentation, which will create a statistical-based model in order to make prediction and decisions. An example of machine learning is given a set of characteristics and traits, we must find a person's age. Even though, his example is not correlated with this project but it is based in the same idea. In this section it will be presented some of the most used algorithms as well as some optimization and model validation techniques [hig].

3.2.1 Feature Selection

Most of the machine learning techniques aim to improve the output given. Feature selection also has the same goal. Usually, during the feature extraction we try to get all the information that is possible. Unlike the former, the feature selection aims get rid of attributes or features that are either confusing our classifier or are simply useless to it. That is because having a lot of information does not mean that the classifier will be a good one. Another advantage of removing useless information is the time we spare during the training and classification of new data. With less information the time spent is much less, besides the simplification of the used dataset. During the classification of new data, there are some techniques that can be used in order to find out which are the most important and relevant features that lead to a successful classification. These techniques are split into three main groups: Filter Methods, Wrapper Methods and Embedded Methods [DHS09].

The first one is based on a ranking system. The filter methods uses a ranking in order to keep or discard features from the dataset. These methods do not consider relationships between features. Although this is a good property, due to its robustness to overfitting and computation time, sometimes some useless features are also ranked as useful because its disregard of relations with the others. Compared to the others, the filter methods do not use a classification model in order to achieve results. Thus, in terms of computation time this is the best solution.

Compared to the former, the wrapper methods considers the feature selection a search problem. That being said, in these methods, also, a ranking system is created in order to find the best combination of features. Multiple combination of features are tried and evaluated by a classification model. They are discarded, or not, according to the score given by the model.

Also, for the last methods, a ranking system is created in order to find the best combination of features. All the combinations are tried and consequently evaluated. Unlike the previous one, the embedded methods are used during the implementation of the classifier, which spares computation time but at the same time it is impossible to generalize it and use them on different classifiers.

3.2.2 Model Validation Techniques

Validation is something that always should be done for the machine learning models. These techniques calculate how accurate and how well adapted is our model for evaluating new data that will be fed with. Instead of testing our model with some data, we can apply one of these techniques in order to have an idea of how will our classifiers behave [PAH+09]. Generally it is not good practice to test the model with data that was used for training, due to the fact that it will produce a model which will be overoptimistic and can easily overfit new data. That being said cross-validation techniques can be used to evaluate these models. There are two kinds of cross-validation: Exhaustive and Non-Exhaustive cross-validation [Cro]. One method of each kind will be explained next.

3.2.2.1 Exhaustive Cross-Validation

Regarding this method, the idea is to learn and test each and every possible combination of the dataset. This can be, sometimes, impossible to calculate, depending on the number of entries of the dataset. The most common technique is leave-p-out cross-validation. For this technique we leave p combinations of n entries to use as the validation set and the remaining as the training set. We must do this for each combination of n there is. Thus, if we have an excessively large dataset it is impossible to apply this kind of technique. In order to overcome this problem, non-exhaustive methods were created, which will be explained in the next sub-section.

3.2.2.2 Non-Exhaustive Cross-Validation

As said before, these kind of techniques always split the original data in order to evaluate the dataset. In the case of the non-exhaustive evaluation models, the data is only split once, compared to the first method in which all the combinations were tried. Although there are many sub-types, the mostly used is the k-fold cross-validation. Firstly, this method subdivides the dataset into k equal, or almost equal, folds. K-1 of these folds are used to create a model and the remaining fold is used as the testing data. We apply the same method for which one of them, in order to use all folds as the testing data. This method is commonly used, due to its easy implementation and good outputs it provides. An example of this technique is explained in the figure 3.2.

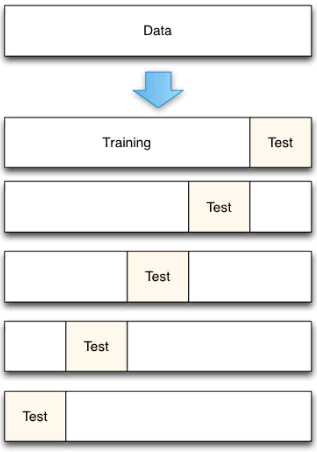


Figure 3.2: Example of the cross-validation method.

3.2.3 Classification Algorithms and Techniques

The whole goal of the machine learning is to predict something about new data. In order to do this, we must apply classification algorithms, which are also known as classifiers. These classifiers are trained using a previously given dataset and only then, after being trained, they can predict something. This learning process can be either supervised or unsupervised [Mac]. In the first one, the classifier is trained using an already labeled dataset. On another hand, an unsupervised learning process uses unlabeled data and the classifier is trained using clustering methods. There are also some algorithms which use at the same time unsupervised and supervised methods in order to train a classifier. Some of the classifiers used during the development of this solution will be described in the next sub-sections.

3.2.3.1 Decision Tree

Decision trees are usually the most used classification method, due to its easy implementation nature. This tree-shaped classification model, calculates the probability of the events outcomes and its cost or utility. Although this is not a feature selection technique, it only uses the most valuable or useful features which gives an actual classification. The tree is grown by recursively partitioning the feature values and deciding which way leads to a well defining classification. This goes on until the partitioning of the values can not separate the elements or does not add up any value to the evaluation. The features that are chosen to take part on the decision tree are chosen in order to increase the information gain of the model, thus being a greedy model. Although, as said, this is a very easy to read and implement model, it has its own disadvantages as well. In decision trees, sometimes, the calculation can get very complex and not-worthy due to the countless possible outcomes that a leaf origins. Also, the dataset used for this model must be well balanced or else some features/classes will be biased due to the fact that some of them dominate. Also, if not well balanced the decision tree can overfit new coming data [KS08].

3.2.3.2 Support Vector Machine

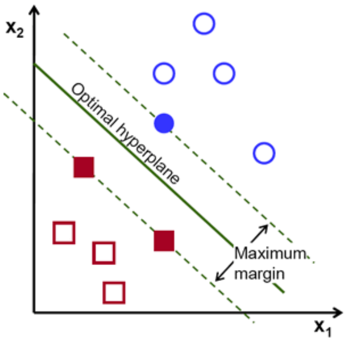


Figure 3.3: Support Vector Machine graphic.

Support Vector Machine, or SVM, is also a supervised learning algorithm. The idea behind this non-probabilistic linear classifier, is creating a model which can separate the two different kinds of data and successfully include a new example in one of those groups. This separation is called hyperplane [SC08]. This line is calculated in order to maximize the separability between both of the different kind of data. The problem is that sometimes we just can not separate linearly the dataset, so we must use a softer approach in order to admit some errors. The soft hyperplane, or soft margin, wins over the hard one since it can be easily generalized, which is normally the goal of a good classifier. Another problem of admitting zero errors, is that instead of classifying new data we are just overfitting them into the model. An example of these kind of classifiers can be seen in the figure 3.3.

3.2.3.3 Adaptive Boosting

Adaptive Boosting, also known as Adaboost, is a powerful machine learning concept which combines several weak classifiers into one. This combined tool is far more capable of providing good results compared to each one of the weak classifiers. This kind of classifier often outperforms most of the already strong classifiers, like SVM. The idea behind Adaboost is having several steps and weak classifiers result's weighting in order to achieve a good result [Ada]. Considering that weak classifiers are easy to implement and so, do not require a lot of computational resources, we can tweak each new result on behalf of the previously misclassified results. This will result into a better output each iteration, allied to very low performing times. There are four types of Adaboost, which will be briefly explained: Real, Gentle and Logit [Opeb]. Each one of them works better with different times of data. Normally the most used types are the Real and the Gentle Adaboost. The Real Adaboost type, use ranking predictions in order to pick results, working mostly with categorical data. Gentle Adaboost often ignores, or put less value on, data which is apart from the rest (i.e. outlier data). This often produces good results with regression data, like the Logit Adaboost [Log].

3.2.3.4 K-Means Clustering

Clustering, in the machine learning field, is an unsupervised learning technique which groups entries from an unlabeled dataset that are similar to each other. The K-Means Clustering method is based on the same premise. It groups data which are similar in terms of features/characteristics and groups them into K groups. This calculation is done by picking, firstly, the two entries that are the farthest apart. Then the algorithm start to group them regarding the proximity of the features to the two first picked examples. Whenever it is added one example to the cluster, the mean of that cluster is calculated in order to take account the new added value. This process is done until we group every single example from the dataset into the k clusters [KMN+02].

3.3 Technologies Review

Some aspects about the technologies that will be used are worth to be explored, due to the innovative concept of this tool. Some insights about the language and library used during the development of this project will be given, as well as an introduction about the mobile operation system that this tool is going to run.

3.3.1 C++ and OpenCV Library

All the image processing covered above will be done in C++ with the help of a library called OpenCV. C++ is a well established and general-purposed programming language. This is a robust and object oriented language so it makes sense to use it in this project. Since this language has been around for many years many libraries have been developed, which in turn are also well established. OpenCV is one of them, which is one of the most used image processing libraries. OpenCV is free for any purposed and supports almost any platform, being this a plus since the project is going to be ported as a mobile application. Another reason to use this library is because of the countless optimized algorithms that it have implemented, which are used for image denoising and feature extraction. This library is optimized to be used with C/C++ language, so it can take advantage of multi-core processing for high performance solutions [Opea].

3.3.2 Smartphones and Mobile Operating Systems

Considering that we are trying to develop a tool that is both low-cost and easy to use, the solution is to run it on a mobile system. The reason behind this is based on the fact that the smartphones combine features from many gadgets, like computers or cameras, into one single machine that is totally portable. With a smartphone we have a lot of features, like Wi-Fi or GPS, that give freedom and more opportunities for the developers to create new solutions. The most important feature of the smartphones, at least for this project, is the camera. This feature gives the possibility of taking images directly and use the tool automatically, instead of transferring them to another machine that is capable of doing the job.

Regarding the actual mobile OS, there are two main systems that are used: Android and iOS. The choice of one of these, lied upon which had the biggest market share compared to other systems. Thus, as a first step on deploying the developed tool, it would make sense to do it for Android smartphones since they have considerable more users than the others operating systems, as shown in the figure 3.4. This mobile operating system was created and it is owned by Google. The success behind this OS it is, firstly, because being open source and due to its easy licensing policy. These characteristics caught the attention of a lot of developers since they could create their own new and portable solutions without having much trouble. Even though smartphones

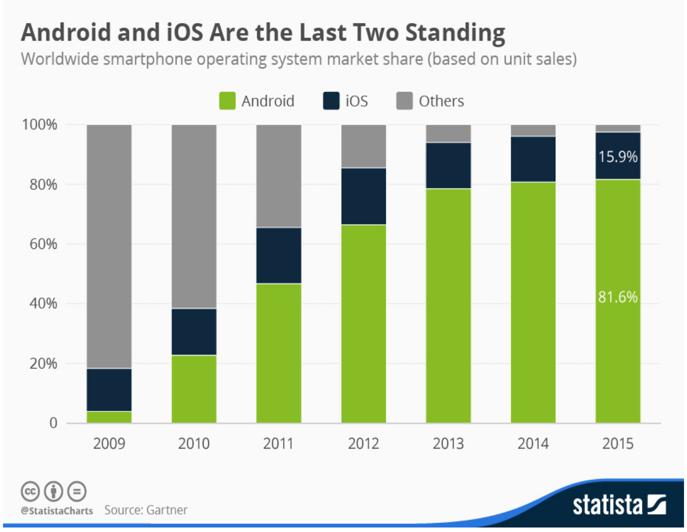


Figure 3.4: Android share in the mobile OS market [AMS]

specifications are getting better, they do not have the same capabilities of an actual computer. Thus, when creating a smartphone application we need to pay attention to its complexity or we risk to not be able to run it properly.

3.3.3 Marketed Cervical Cancer Automatic Screening Solutions

The creation of most of the cervical cancer automatic screening devices started in the seventies but it only reached a prototype phase in the mid eighties [CCS], so these solutions are around for a while now. Before there was only empirical solutions to find out about the existence of cervical cancer or primordial computers that could do only basic operations, returning a lot of false alarms.

Even after the computer turned out to be interactive there were some systems that could not be interactive yet. The first one to be interactive was the PAPNET system [KLS+94]. This system firstly would process low resolution objects, by an algorithmic classifier, and only after the high resolution ones, by a neural network classifier [BM14]. The result of this process was a ranked classification of the most abnormal cells that would in turn be classified as normal or suspicious. This classification was made by a clinician, making sure that the normal ones were well classified and also checking if the suspicious ones really needed further investigation or not. Other system that is similar to the last one was the AutoPap 300 [CBB+03]. The difference between them is that the last one uses strobed illumination and do not have the clinician classification step, suppressing the interactiveness of the system. Then more systems appeared like the ThinPrep Imaging System [QNR+09]. This system had two novelties that was the way the sample was prepared, in turn giving better results due to its cleanness. The other one was the measurement of DNA in the cells, since the cancerous ones would have more DNA than the normal ones. Lately because some companies merging, the FocalPoint system was developed [BM14]. This system was an extended version of AutoPap 300 but with a new sample preparation technique called SurePath that improved a lot the quality of the results [CCS]. This system would also rank each sample as normal or instead ranked by its abnormality.

3.4 Related Work

3.4.1 MalariaScope

As seen in the figure 3.5, the MalariaScope does not look like the normal high-end microscopes. That is because it is a low-cost microscope prototype developed by Fraunhofer. All the images gathered for this project, either for the dataset or the testing set, are obtained with the help of this equipment. In the image on the left, it possible to notice an adapter. We use this adapter to attach a smartphone in order to take images from the sample we are observing. That is the main reason why this microscope is used for this kind of projects. Also, since Fraunhofer tries to develop low-cost solution, these intents should go through this kind of equipment as well.

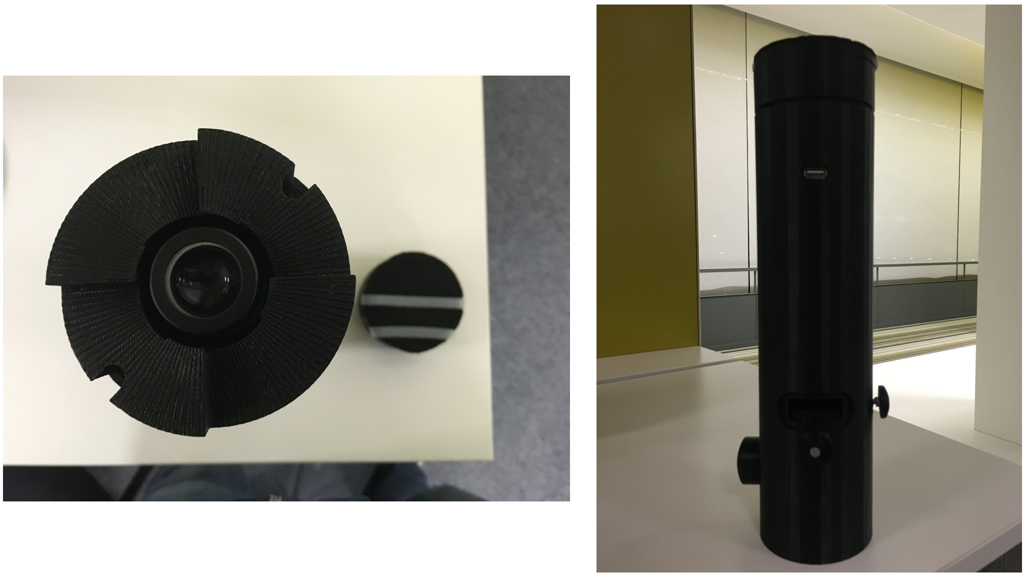


Figure 3.5: Malaria Scope developed by Fraunhofer

3.4.2 Autofocus in Image Processing

One of the most important characteristic of an image, either digital or not, is how well focused it is. A focused image can be defined as an image that is clear and sharply defined. This is a very simple explanation but it can provide an overview of what we are going to talk about in this chapter. Either it is for processing purposes or not, this characteristic that images possess is very important. Obviously when we are talking about image processing, this takes bigger attention since we need to have a well focused image to actually process it. Related to the project that will be developed, the focusing degree of an image is imperative in order for the system to be able to recognize tiny objects on the image, which without an image properly focused is something that is not achievable. Not only on this project, but generally this is a very important characteristic in biological and biomedical fields. Due to the technological progress, also the degree of how well focused is an image can be achieved by automatic means. Since the screening of diseases are starting to be also automatic, these two technologies can be allied to reach even better results.

Although it was already mentioned that autofocus technologies are widely used in the biological and biomedical fields, this was first used due to the digital camera “boom”. These systems, firstly, were relied on only one sensor but during the time more sensors were added to determine the correct focus of an image. Also, algorithms are used on the digital camera’s autofocusing systems. These algorithms are used, for example, to focus on moving objects or people by using light, speed and acceleration metering. Basically for this kind of systems there are two kinds of detection in order to achieve the right degree of focus: phase detection and contrast detection. In the last kind of detection method, what happens is that the system tries to focus by measuring the contrast between the pixels. This can achieve the right focus because it increases as the intensity between the pixels also increases. The contrast detection method is not particularly good in systems that need to take moving images since it does not use any distance measurement at all. So, because of this, it does not know if the object is moving away or closer even if there is a loss of contrast on it, when taking moving pictures. Comparatively, the phase detection method is a bit more complex than the previous one. In this method the incoming light, as in image, is split into two pairs of images so the light is redirected into micro-lenses that are placed in the opposite sides of the lens. This creates a rangefinder, in order to give the distance between the object and the camera. Then each separated image is compared in terms of light intensity so the separation error between them, if it exists, is used to provide how the lens needs to move in order to focus it and if the object is in the front or back focus. These two methods are considered as a passive autofocus since they both work with contrast sensors but there is also active autofocus that use signals in order to illuminate or estimate distances between the lens and the object. For project reasons the passive AF is more interesting, since we do not need to calculate the distance between the camera and the object because we are using a microscope, and because active focusing is somehow deprecated.

Regarding the software component of this kind of systems there is a wide amount of algorithms in order to achieve good focusing levels, or also to improve it with already implemented systems. Driven by trying to find the best algorithms to achieve good autofocus results, Yu sun et al. [SDN05], performed a study using the 18 most used algorithms in order to find it, since the standard used technique, AutoCorrelation, was not the best for all the cases found. The algorithms used in the study can be split in three kinds: Derivative-based algorithms, Statistics-based algorithms and Histogram-based algorithms.

Firstly, the derivative-based algorithms are based on the fact that neighboring pixels in images have an intensity level. The calculation of these levels can reveal how sharp and well focused an image is because of how the intensity changes, higher the changes, higher the sharpness of the image. Statistics-based algorithms are less sensitive to the noise comparatively with the last method and they can decide either an image is focused or not based on the calculation of the variance or the correlation. An image histogram is the graphical representation of the distribution of the lightness in an image. The Histogram-based algorithms use this kind of data to analyze the distribution and frequency of image intensities in order to identify focused and defocused images. Still regarding the study made by Yu sun et al., these algorithms were ranked by evaluating each focus curve, provided by each algorithm. Five individual criteria were used in order to rank each image set for each algorithm: accuracy, range, number of false maxima, width and noise level. These criteria were all summed up in a total score in order to compare it with the optimal level computed manually. The results of this study revealed that the Normalized Variance algorithm, from the Statistics-based algorithms group, was the best in most of the cases. Specifically in noisy images and low-pass filtered images (images which were reduced in frequency levels) this kind of algorithm revealed the best overall scores. On the other hand, when the images were segmented and sub-sampled in order to increase execution and performance speed, gradient-based focus algorithms achieved the best results. Since in this project, firstly, we will segment the images that are given, this can give some insides of each algorithms we need to use, namely the last ones that were referred.

Chapter 4

Implementation

The main goal of this chapter is to go through the implementation of the developed solution. In order to be well understood we will go deeply into the image processing module as well as the machine learning process. Before talking about the actual implementation, the section 4.1 about the dataset used, and created, is presented. This sub-chapter makes sense, considering that without a good and well built dataset the machine learning would not work. Then we will describe how the segmentation was done, step that helped to create the dataset and where we extract the information used for the machine learning module. Finally, the machine learning process is described and simultaneous some of the results that were achieved during this process. A simple diagram is shown in order to understand better all this line of work.

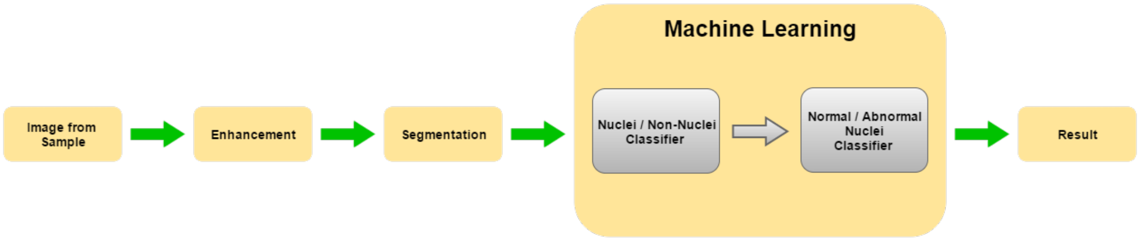


Image from Sample
Enhancement
Segmentation
Machine Learning
Nuclei / Non-Nuclei Classifier
Normal / Abnormal Nuclei Classifier
Result

4.1 Dataset

Considering that we are dealing with very sensitive issues we must implement the most accurate techniques. This also concerns the dataset being used. In order to have a complete and accurate dataset we must collect a considerable amount of images from the samples. As a consequence we will have a plethora of nuclei images, which will provide enough information for the classifiers perform a trust and successful classification. Another reason for having a populous dataset it is because the Android component of this project. Since the images will be captured, mostly, by smartphones, we will have a lot of discrepancies between the images captured due to the different hardware components they have. To counteract this challenge, the dataset created it is composed by images taken by multiple smartphones, more specifically with: HTC One M8, LG

Nexus and Samsung S5. By using these three smartphones, it is possible to have a wide range of different values which represent the same type of nuclei in the samples. Assuming that we would only use one kind of smartphone to build the training set, analyzing nuclei from images taken by another kind of smartphone would probably be ignored or bad classified since there would not exist any terms of comparison.

For the creation of this dataset two kinds of sample were used. One of them was a negative sample, which did not have cervical cancer at all, and the other one was positive, which had evidence of this cancer with either high grade and low grade lesions. Regarding the actual numbers of the training set, it was created using a total of 150 images, which 66% of them were taken from the positive sample and 34% from the negative. The reason behind this resided in the lack of abnormal nuclei. Since there exists more normal nuclei than abnormal, it was necessary to have more images from the positive sample so there was enough knowledge about the abnormalities a nucleus can have in order for the classifier know how to work properly. Although the abnormal characteristics are only used for the second classifier, it was decided that we should take in account this challenge in order to construct from the beginning a dataset that could be used for both situation, with little differences between them.

Justified the need of a robust dataset we can go through its content. This dataset is created with the help of the TCGFE library [RCEc16], used to find 152 characteristics about each nucleus. Also, we classify new samples based on the nuclei characteristics that are found through this library. Since most of them play a crucial role during the classification module, they are all used during this phase, in both of the classifiers. Although this seems quite a lot of characteristics, various experiments were made in order to check if they were actually useful. This library is composed by three components: Texture, Color and Geometry. At the first glance, for example, it was easy to say that we could fully classify the abnormalities of the nucleus based on its geometry but that is not entirely true. What was found in these experiments was that, although geometry attributes are the most used ones (i.e. in the abnormal/normal classification), the others components play their role as well. This happens because machine learning is so impressive that can find patterns on values that are not so easy to identify right away.

All these characteristics did not have the same value if there was not enough examples. As said before it has been taken 150 images, total, from both positive and negative, which originated 3919 examples. Out of this number, 1741 of it represents inflammatory cells and objects which not represent valuable information for the second classifier and 2178 the actual nuclei. We can consider that these numbers are well balanced since we are dealing with very large numbers and because a difference of roughly 400 examples is not considerable. It is possible to confirm these numbers in the table 4.1. The number of examples is something to be taken in account since this way either the classifier can be successful or not. With very few examples we could barely classify new examples since we, probably, did not have information to classify examples that were very differently from the characteristics we already had. On another hand, if we had an even larger amount of examples we could be just doing over fitting, which technically it is not classifying an example. Over fitting happens when we have a large amount of examples which a new can be compared to. When the classification process starts in these cases, the classifier will probably just find an identical example and will not try any new classification at all. This is a problem since we are not really sure if the classifier can actually classify or is just trying to find identical examples. When the second option happens, when we feed the classifier a totally different and new example to classify the classifier would struggle to give a good result since it was not well trained.

The second dataset, although it is represented in a different file, is based on the first one. Considering that we only need the nuclei information to classify other nuclei as to their morphology, it is ignored all objects which do not fit in this assumption. Thus creating the second dataset, which is a filtered version of the first. As we can see in the table 4.1, now we only have 2178 examples of which 1384 are normal nuclei and 794 are abnormal nuclei. Another difference between these two datasets was the difficulty on the labeling process. For the first one, it was very straightforward, since we only had to classify the objects as nuclei or inflammatory cells, which is not that hard in naked eye. On the other hand for the second classifier, this process was a bit more complicated, which originated a big step back. Considering that we have a lot of examples, even though it is a filtered version, we need to label them one by one, which is very time consuming. The classification for the second is about how abnormal is the morphology of the nucleus. Although a lot of information was read about this, the labeling did not match the accuracy of an actual surgical pathologist. Thus, when we used the dataset for classifying the abnormalities of the nuclei the results were not very consistent. This forced an urgent meeting with a surgical pathologist which could evaluate the accuracy of the labeling. After the correction of the dataset, we used this dataset for classifying the nuclei.

As a final remark about the dataset, all the data used is normalized and only then fed to the classifiers. Both situation normalized and not, were tried in order to find out if it really made the difference in the output, which it did. Normalizing the data means that we fit the data within a specific range of values normally, and used in this case, the unity, where the values range from zero to one. The problem of not normalizing the data is that, in most of the cases, we have difference characteristics or fields with different scales and units. What happens is that we can not fairly compare all of them with each other if they have different units because of the different range of values they have. With normalization this problem cease to exist since they fit from zero to one, or any other range we choose, which can give the possibility of a fair comparison and consequently a better analyses of what fields are the most important. Ultimately this also gives better results in the machine learning process, at least for this project.

Table 4.1: Object counting in the created dataset

	Negative Sample	Positive Sample	Total	Percentage
N^o of Inflammatory Cells and other Objects	1049	692	1741	44,4%
N^o of Normal Nuclei	711	673	1384	35,3%
N^o of Abnormal Nuclei	0	794	794	20,3%
			3919	100%

4.2 Image Processing

One of the most important aspects of this project is the segmentation component. This will be a crucial variable on the successfulness of this solution. Since the project aims to develop a tool to screen cervical cancer in developing countries, a lot of hold-backs were set in the beginning. First of all, one of the reasons that the pre-processing is important is because how the samples are stored in first place. Due to the fact that developing countries do not have the latest technologies, not all the samples are stored in the most "clean" way. Compared to other solutions that already exist, this do not help the screening of the cervical cancer. Because of this reason the first thing to do with the images is to remove most of the noise they have. This is a particularly meticulous job since the cells nuclei are very small compared to all objects in the image, therefore they are very sensitive to all image modification operations made. This reason led to a lot of image processing experiences being made in order to find out the best solution that at the same time could remove the noise of the image and in the other hand to keep the cells nuclei on it. Despite being said that the way the sample are stored cause a lot of noise, this is not the only reason of it. Since one of the goals is to not rely on complex machinery, the images are taken by a smartphone. This cause a noticeable discrepancy between the images in terms of light, color and resolution. To solve this problem there was a need to take a considerable amount of images with different smartphones in order to find a solution that was uniform across all of them. Regarding the image processing implementation itself, firstly, we will go through the experiments made. The intention is to show why there was a need of performing them and how they affected and decided the subsequent work. Later and concluding this topic, the resultant work and its implementation will be presented.

4.2.1 Image Processing Experiments

As mentioned before, this was a very important step in the implementation of this module because this whole set of experiments helped to refine the whole image processing. The success of this implementation depends on how the image looks like and what we can extract from it. Given this nature, the first and natural approach was a "trial and error" methodology in order to give some insights of which tools were needed to perform what was aspired. This then evolved into a more methodical approach compared with the initial one, specifying and outlining a course. Both this distinct approaches culminated into the implementation of the actual image processing module of this project.

One of the first experiments made was detecting the whole cells by color, as we my see on the set of figures 4.1. At the first sight this was an approach to consider since the cells have distinct colors (red and blue). After testing and comparing the results this idea was discarded immediately. The reason behind this, and as reported before, was because of the different color and light levels of each image. Since this OpenCV's function depends on pre-set values of colors, this approach was very hard to uniformise in order to work for every sample. Because of the variance of the color levels in the images sometimes the algorithm would get what was pretended and in other cases, where the image was darker or lighter, would not. This was a very important approach for setting the course of the project because, at outset, it was avoided pre-set values function in order to not commit to just one kind or specific color and light levels images.

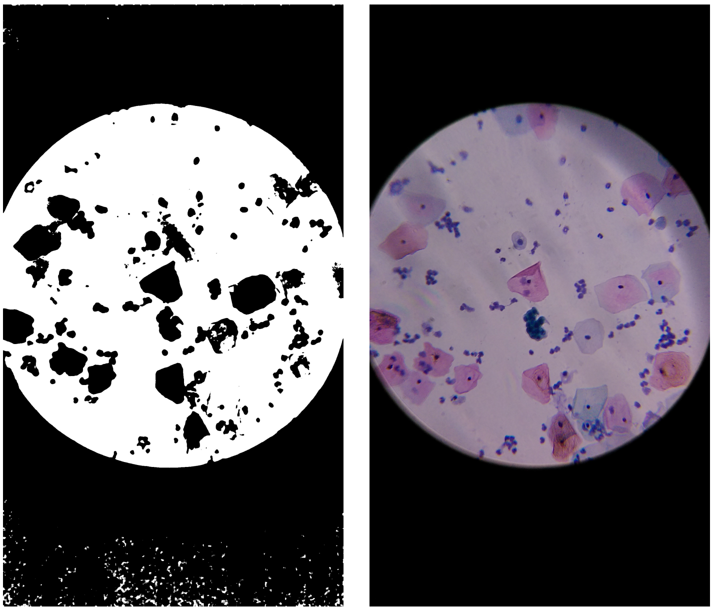


Figure 4.1: Object detection by color. On the left said the result of the segmentation by color of the right one.

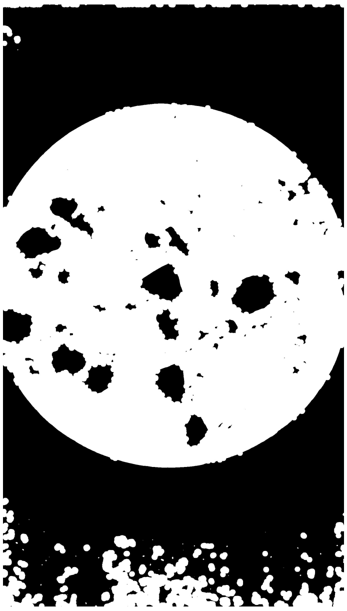


Figure 4.2: Object detection by whole. Resultant threshold of the same picture above shown.

Another time consuming experiment was trying to identify the cells as a whole. Again, at the first sight this seems to be the natural approach but the time spent on it unveiled that it was not the optimal one. This led to a considerable amount of variants which were also discarded. The attempt of the whole cell identification was composed by two steps. First, the image was enhanced in order to reveal the morphology of the cells and ignoring the little details of it. Then the contours were found based on the output of the last step and each cell's ROI was created. Although this approach was discarded, satisfactory results were achieved. The problem was that in some cases the noise that persisted in the image was too considerable and was negatively influencing the results. Then if we tried to remove it through morphological operations we ended up removing many details from the image, as we can compare the figure 4.2 with the left side figure of 4.1. This experiment originated what would be the implemented solution for the image segmentation.

Trying to improve the last approach, color manipulation techniques were used in order to enhance the morphology of the cells. The output of this experiment revealed that subtracting the red channel to the blue one enhanced the blue colors (i.e. enhancing inflammatory cells and nucleus, as well as big blue cells) and subtracting the green channel to the red enhanced the red colors (i.e. enhancing big red cells). Even though this kind of results were achieved it was not enough since some cells were still being ignored, creating missed opportunities of identifying abnormal cells. Nonetheless, this was a breakthrough since it was found that instead of subtracting color channels we should only use the green one for a better identification process. As we may see in the set of figures 4.3, the green channel at the same time enhances the small objects and ignore unnecessary objects ones. This was an easier approach than subtracting channels, so this triggered the realization that for this project we should try approaches that are, at the first glance, easier and if that does not work we should move on to a more complex one. This helps in terms of optimization levels since the computer does not spend time doing useless tasks.

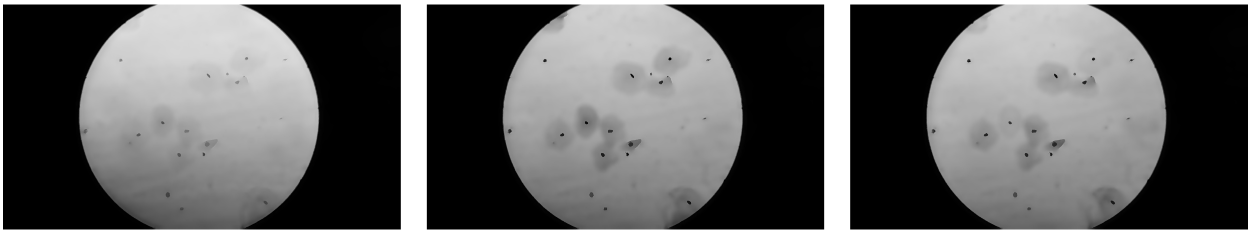


Figure 4.3: Blue, Green and Red channel, respectively, from an image

At the same time another experiment was being made. This one, instead of trying to enhance the identification, had the intention of ignoring objects that had nothing to do with the sample itself (e.g. black/grey spots or hairs). Since most of this object's blurriness level were higher comparing to inflammatory cells and big cells the thought was that we could filter this kind of objects by it. In fact, it was possible to filter these undesired objects but at the same time the algorithm was filtering nucleus as they fitted in the blurriness range. Since there was not any kind of pattern, this idea was dismissed since we could not afford to ignore any nucleus. Also, as it was found later, the blurriness level of the nucleus was not substantial in order to distort the image in a way that we could not identify the most important characteristics from it. A function was created in order to calculate this blurriness level. The function would use the sobel function in order to calculate the variation from the light to the dark on the x axis and y axis. Higher levels would mean more image quality, so to calculate the blurriness of an image we add these two values. Consequently, if the value was too low the image was really blurry and would be discarded. As said before, this function is not applied at any point of the solution since there was not a pattern on the blurriness levels.

4.2.2 Implementation of the image processing module

As for the image processing component itself, the previous work not only led to its implementation but also the improvement of it. As stated, the morphology of the big cells would not give us important information so an approach that naturally occurred was trying to identify smaller objects, such as inflammatory cells and nucleus, and completely ignore the bigger ones. The results of this approach were so promising that they made the first one to be completely discarded (i.e. trying to identify bigger cells). This time, on most of the samples, we could identify nearly all of the nuclei on it. Although the inflammatory cells were also in this set due to their similarity to the nucleus, the main goal was not to discard any of the objects in the sample that could be the key of finding cervical cancer. By virtue of this approach's success, the identification of the smaller objects was the path that has been followed. A key feature of the OpenCV library that led to the successful implementation of this module was a function which calculated the mean color around a certain point. Considering that we want to identify the nucleus, which are very little objects comparing to the whole image, we want that their color level is the most homogeneous possible. This way the borders between each object is so well defined that when we apply edges detection algorithms we can extract exactly what we want. All this image quality we want to accomplish is achieved by the previous pointed out function. The work behind the curtain is that each pixel acquires the color of the mean of the pixels which are at a certain range of it. Also this function only takes in account the ones that are within a defined range so it does not use them to miscalculate the mean. With the help of both this function and by only using the green channel, as seen in the experiments sub-chapter, we can smooth the image and at the same time enhance the nucleus visibility, as we can see by comparing both images on the figure 4.4. Since this action is commonly used in this project (i.e. mean shift filtering the image and then using the green channel from the output image) I created a function that automatically do this at the same time to avoid duplicated code, thus improving the project itself and to spare time writing the same code over again.

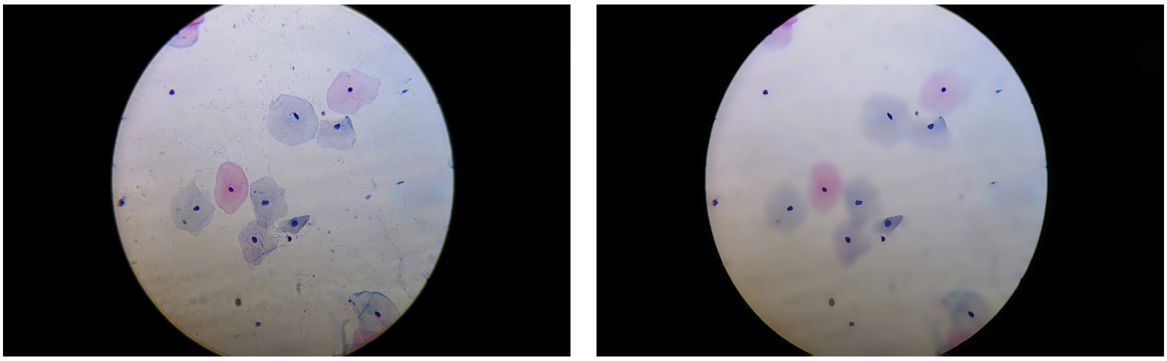


Figure 4.4: Original image on the left and on the right the result of mean shift filtering.

This previous function defines the first phase of the image processing module, where the image is improved in order to enhance all the relevant objects for the cervical cancer screening. The next step of the implementation implicates using the output of the previous one (i.e. the enhanced image) and identify all the enhanced objects. Again, it is used openCV library functions to perform the edge detection in order to find out the objects contours. To achieve this we calculate the image derivatives in x and y directions using the scharr function. The decision behind the use of this function resides in the fact that there was not another one that could achieve the pretended results. Although there is another function that can perform this kind of calculations (i.e. sobel), the scharr function not only can detect the objects of the images being used but also, theoretically, can achieve better results comparing with the first one. In order to continue the edge detection we then convert the derivatives to unsigned 8 bit so we can combine both of them into one single image. Since this kind of functions create very noisy images we must pass it through a function which blurs the image, so non relevant contours are found. Resulting of all this is a good identification of the morphology of the small objects. This pre-processing before actually detecting the edges it is important in order to ignore all the bigger objects morphology and sticking with the aspects which are important for this project. Lastly and after we have the enhanced image, it is used the well known canny edge detector. With all the work done before, this operator easily finds the small objects contour's. As we may see in the figure 4.5, some of the contours lines are not connected so we perform a morphological transformation. Comparing the left image on the figure 4.6 with the last one we can surely detect the differences and why this was made. Even though we could simply use the canny edges output, some of the objects could not be detected on later stages of the image processing module since, as referred before, some of the lines are not connected. To finish

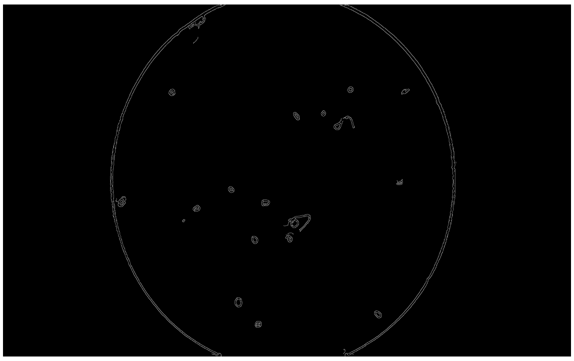


Figure 4.5: Detected edges from the original image

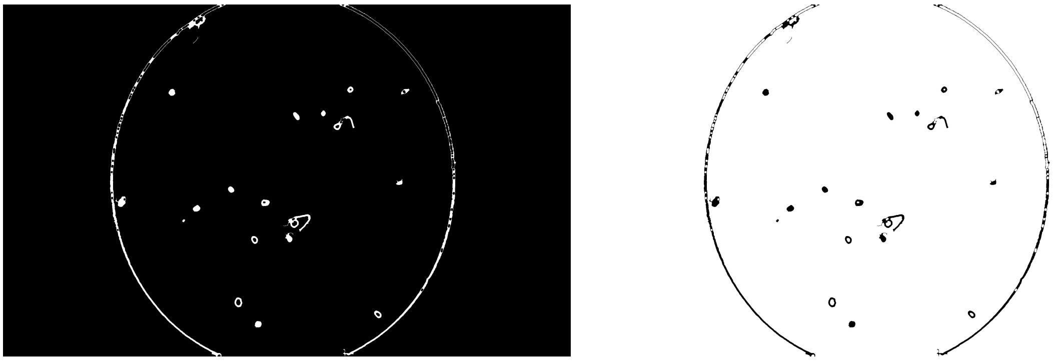


Figure 4.6: On the left the edges after being applied a morphological operation. On the right the same image after being applied the bitwise not operation.

the edge detection phase, we then perform a bitwise not operation with the purpose of displaying the objects as small black blobs, as seen in the right image on the figure 4.6.

Settled the last stage, we can now use the detected edges image in order to draw the contours of the objects. The result of this will be the segmentation of each small object, which will be used to find their own specific characteristics. In a way to prepare for the classification phase there is a sub-step embodied in this one, which will be described as well. As referred, for this step we feed the output of the edge detection and use it to have the actual contours of the small black blobs of the image seen previously. Only performing this step do not fulfill the needs, so we need to use another function to find the whole area of the blobs. This function uses the points from the one that returns the object's contours and creates a bigger contour which contains all those points. This is particularly important because when we draw the contours themselves some of them do not connect, so to close these loopholes this function is used in order to have just one big contour around the small object. As we may see in the figure 4.7 all the contours are well delimited, although they lack sensibility.

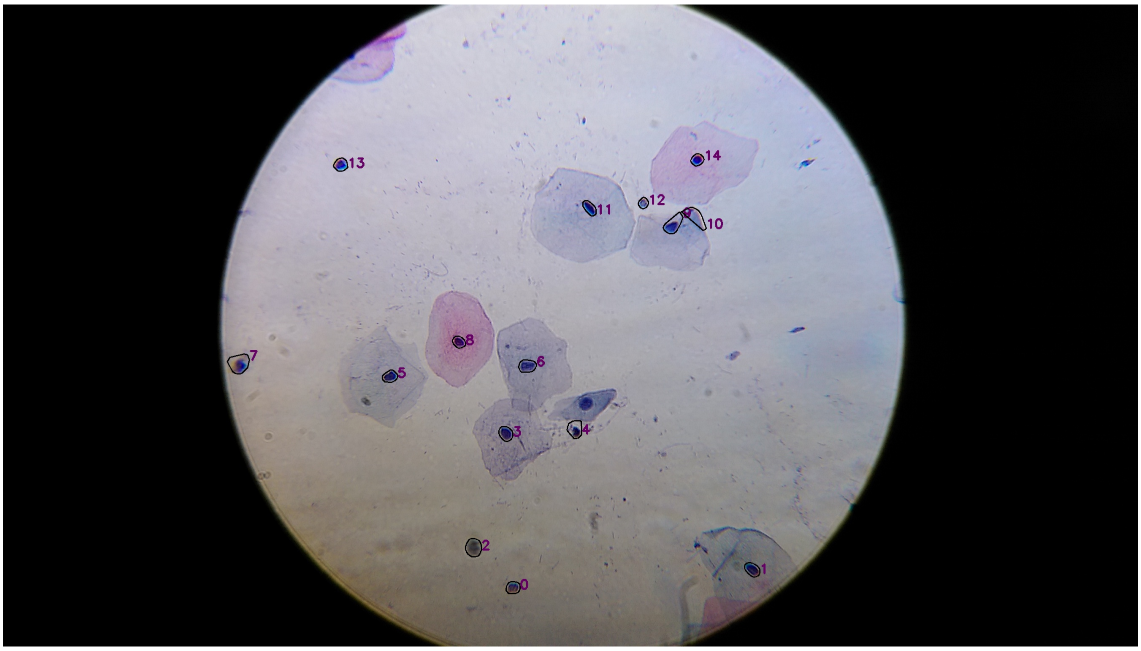


Figure 4.7: Resulting contours of both function combined

Regarding the segmentation, which is done in this phase, a ROI of each contour is created. The ROI is a rectangle which is slightly bigger than the contour and contains the small object itself. As stated earlier, the sub-step will use this image in order to find specific characteristics about each object. This sub-step is done by a function which at the same tries to detect a small object in the image and find its characteristics and work filtering function, for the function which this one is included, if no actual object

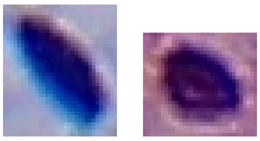


Figure 4.8: Example of two resulting ROIs.

is found on the ROI. As just said, the filtering involves passing through the finding object function, which will ignore contours that do not contain any relevant object, and if the ROI has any blobs on it, to make sure every image it is relevant for further investigation. Regarding the sub-step and since the function that detects the area around the objects do not give a very specific contour, the need of creating a function which detected the small object contour in a more detailed way appeared. Similar to other detecting processes in this project, this function also uses the revealing detection function created (i.e. the one that uses mean shift filtering and the resulting green channel). The difference in this one is that before finding the small object detailed contour we enlarge the image to a size two times bigger. Even though the resulting image do not have a high definition, the resulting contour is much more detailed than if we applied a typical OpenCV's function to find the contours on a smaller image. At the first glance, this was a great idea since we could get a good detailed contour but it still had to be drawn in the original image (i.e. the smaller one). The solution for this problem was to fill the drawn contour.

This way no detail was missed when the image was resized back to normal and with a simple color inversion we got the mask for the object in the ROI. This mask and ROI are very important since with them we can feed them to the feature extractor library and get all the characteristics we can from that object. Then, if in this sub-step any object is found a ROI and mask is saved from each one, which finishes the image processing module. Two examples of these masks can be seen in the figure 4.9, regarding the same two ROIs above.

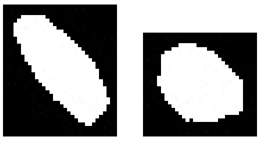


Figure 4.9: Example of two ROI contours.

4.2.3 Final Remarks

Despite the fact that this project seems to be a machine learning problem, while developing the image processing module it became clear that this was also within the problem. The obstacle in this project is that the images of the cervical cancer samples contain a lot of elements that do not give any valuable information, so we must develop a tool that can successfully detect the valuable elements in a trustworthy way. With the achieved results it is easily noted that most of the nucleus in an image are detected. Obviously, and since we are not working with the latest kind of technology, there are some of them that are not detected because of the quality of the image or just due to the fact that the cells overlap and are not well delimited. While developing this module, we tried our best to overcome this problem despite not being with an 100% success rate. Concluding, we have detected almost every nuclei in the image sample but also some of the inflammatory cells. This problem will be tackled through the machine learning module since there was no successful way to do it through image processing. Also in the next chapter it will be talked how can we detect that a sample contains evidence of cervical cancer or not.

4.3 Classification

Despite the fact that this project relies on a good image processing module, not every challenge can be tackled with it. Although we cannot disregard the fact that it is still a core component of it. Since the beginning of the development this tool was branched into two different modules which communicate with each other, the image processing one, which was addressed in the previous sub chapter, and the machine learning module. Obviously, besides the counting ability, the first one cannot infer anything about the images that it is processing. Therefore, in a complementary manner, the machine learning module is added so we can make conclusions about the images. This verdict depends on the information obtained from each and every nucleus and their characteristics. In the end we will have a probability of the presence of cervical cancer. Even though the characteristics extraction are outside the scope of the classification module, the option of talking about it in here was made considering that the library used was not developed during this project. This library called Texture-Color-Geometry Feature Extraction was developed by Fraunhofer AICOS and it was used due to the impossibility of creating such a robust and trustful library during the project development duration.

Regarding the actual classification, we split this module into two different classifications, one followed by the other, and both of them crucial to this project. Recapitulating the previous subchapter, it was not possible to differentiate the smaller objects from each other (i.e. the nuclei and the inflammatory cells). To facilitate, we could simply consider every identified object as a nucleus. The problem is that the morphology of the inflammatory cells is very irregular and if we just classify them as a nucleus we would conclude that the sample had cervical cancer every time. This was not an option since the goal is to develop a trustful tool that could help the cervical cancer screening. So a new classifier was implemented in order to decide if either it was an inflammatory cell or an actual nucleus. This is a very important aspect in this project due to the fact that although most of the inflammatory cells are easily identified as such, there are many of them that look like a nucleus and on top of that, sometimes, they are even above/below the cell cytoplasm.

Since the main goal is to identify the nuclei which are abnormal this will be the last step of the whole process. The purpose of this classifier is to, combined with the output of the last one, count and separate the nuclei in these two classes (i.e. abnormal and normal) which will be identified differently, in the original sample image, for later approval of the surgical pathologist. Once again, this tool is not trying to be a substitute for the clinician's opinion but a way to identify the nuclei which have a dangerous aspect (i.e. morphology irregularities). All this will be displayed in the android application, that will be later explained.

4.3.1 Additional Dataset for Validation

The validation of a classifier is crucial during the machine learning process, since it will provide an estimation of how good they are. That being said, this section's goal is to explain how these classifiers were validated. This process was made by testing each classifier with 20 images from each smartphone that was used during the development of this project. Half of these images were gathered using the positive sample and the other half was composed by images from the negative one. That being said, in the end of each sub-section, the results from this validation will be shown. These smartphones specifications can be found on the table 4.2.

Table 4.2: Used smartphones specifications

Smartphone	Camera Megapixels	Resolution	HDR
Samsung Galaxy S5	16	5312 x 2988 pixel	Yes
HTC One M8	Dual 4	2688 x 1520 pixel	Yes
LG Nexus 5	8	3264 x 2448 pixel	No

4.3.2 Nuclei/Non-Nuclei Classifier

As pointed out on the last chapter, this first classifier will work as a filter. This decision based on the fact that we could not filter everything during the image processing module since some of the objects identified were very similar to the nuclei we wanted. So, we found out about the characteristics of the identified objects, with the help of the TCGFE library [RCEc16], and created a classifier in order to identify which of them were actually a nucleus. The dataset used for this classifier was already described in the dataset creation sub-chapter, which has 3919 examples.

Also for this module many experiments were made in order to find which classification method would fit the best for this case and how could we improve it. Most of the classifiers used in these experiments, and for the implementation, are the ones included in the machine learning module of the OpenCV library. One of the experiments that did not used OpenCV's machine learning module was the decision tree. Even though the decision tree gave fairly good results when cross-validation was applied, as we can see on the table 4.3, when we tried to classify new examples with this model it did not give the results we expected. The results had less than 50% of accuracy for some of the images and others had almost 100%.

Table 4.3: Decision Tree Classification Results

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	1162	249	82.35%
pred. Nucleus	646	1862	74.24%
Sensitivity	64.27%	88.20%	

This method was obviously discarded due to its lack of consistency in the given results, despite the fact that is easy to implement and its own cross validation gave good results. Since we need a consistent and robust classifier the search went on and OpenCV classifiers were used. Regarding these, the ones that fitted the best for further investigation were the Adaboost and the Support Vector Machine (SVM). K-Nearest Neighbor was also tried and it had a simple implementation, but since it did not give any promising results, it was discarded, as is noticeable in the table 4.4.

Table 4.4: SVM and K-Nearest Neighbor results, respectively

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	1404	270	83,87%
pred. Nucleus	404	1841	82%
Sensitivity	77,65%	87,21%	

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	855	971	46,82%
pred. Nucleus	953	1140	54,47%
Sensitivity	47,29%	54%	

Table 4.5: RBF and Linear SVM, respectively, with default values regarding image 1

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	6	1	85%
pred. Nucleus	1	7	87%
Sensitivity	85%	87%	

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	6	1	85%
pred. Nucleus	1	7	87%
Sensitivity	85%	87%	

The search for the best SVM and Adaboost parameters were based on the same premise, we tried every SVM kernel and Adaboost type with the default values until we reached acceptable results. Only then we would change the custom values in order to improve it and find out which classifier was the best. To find the best values an exhaustive search was made, instead of trying one by one. A challenge faced during this experiment was that at the first glance almost every classifier was the best one. The reason behind this was that, for time consuming reasons, we would only test each classifier with one image and only when a classifier stood out we would test it with different images, to make sure that it was actually a good one. An example of this situation can be seen in the table 4.5.

In these first ones the classifiers would return almost perfect values, which can be a good sign of its accuracy rate. To check this out we would do further experiments on other images to confirm this robustness and consistency. On the latest, in the table 4.6, we can confirm that, although these classifiers had good results for just one image, the same accuracy was not preserved since some of them have a 0% rate. In order to a classifier being considered as a good one they need to have some degree of consistency, that is maintaining good results on a couple of images. Ultimately, it was found that in most of the cases Adaboost was the best classifier. There four types of Adaboost, being Gentle and Real Adaboost the preferable choice since they are the more versatile ones. Despite this, for this first classifier it was used the Discrete Adaboost since they gave, overall, the

Table 4.6: RBF and Linear SVM, respectively, with default values regarding image 11

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	1	1	50%
pred. Nucleus	4	5	55%
Sensitivity	20%	83%	

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	0	0	0%
pred. Nucleus	5	6	54%
Sensitivity	0%	100%	

best results. Even though the first image has worst results compared to other classifiers already seen, the consistency in this classifier is higher, as we can confirm in the table 4.7. The main goal of a classifier is to have good results independently of which image it is evaluating. Although sometimes we get worst results, the objective is to keep a satisfactory for each and every image.

Table 4.7: Chosen classifier results regarding image 1

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	6	4	60%
pred. Nucleus	0	4	100%
Sensitivity	100%	50%	

Table 4.8: Chosen classifier results regarding image 11

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	3	0	100%
pred. Nucleus	2	6	75%
Sensitivity	60%	100%	

In the end when the classifier was finally set, an XML with the training data and files with its maximum and minimum values were created. Considering that we need to normalize the data and then train a classifier with 3919 examples, this phase would take too much time if we had to perform this every time. The problem escalates even more in a smartphone, since they do not have the same capabilities as a computer. In order to solve this problem a trained file and its maximum and minimum values are created, so the training phase is skipped. This way we just have to initialize the classifier by reading the trained file and the classification it is done in less than one second, instead of more than 20 when the classifier was fully trained.

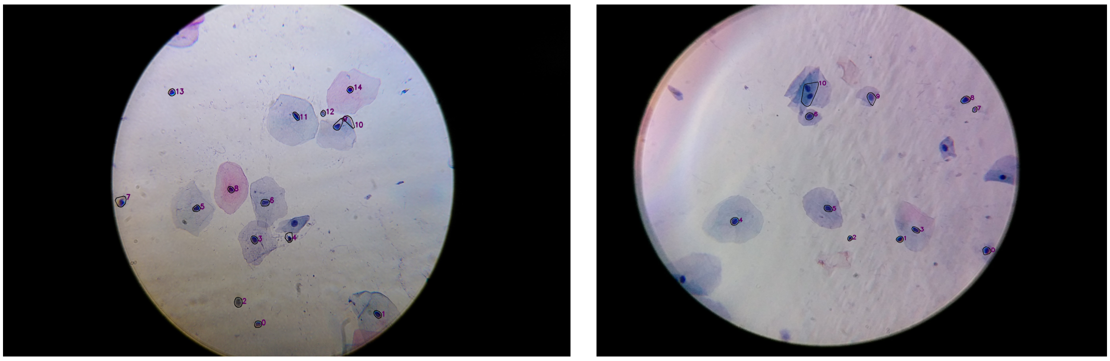


Figure 4.10: Image 1 and image 11, which were referred during this sub-chapter

4.3.2.1 Validation

The results from this validation process can be seen in appendix A in the section A.1. These results only regard the implemented classifier. As it is noticeable in the table 4.9, most of the sensitivity rates exceed the 70%. Also, in most cases the Non-Nuclei accuracy is lower than the opposite one. Although it means that the second classifier will still evaluate some objects which are not nuclei, the Nuclei/Non-Nuclei classifier represents a good filter for this kind of objects. As it is noticeable in the result's tables, most of the non relevant objects are filtered, even taking this issue in consideration. Nevertheless, in this kind of processes it is impossible to have a 100% success rate.

Even though, it is safe to say that a robust and trustworthy classifier was implemented. The high Nuclei sensitivity rate indicates that most of the nuclei in the images are successfully passed to the Abnormal/Normal classifier. Therefore, implying that there is a very low risk of missing an evidence of cervical cancer.

Table 4.9: Overview of the Results

	Non-Nuclei Sensitivity	Nuclei Sensitivity
HTC One M8 HSIL	80,8%	71,4%
HTC One M8 Negative	91,2%	92,1%
LG Nexus HSIL	82,3%	87,8%
LG Nexus Negative	76,9%	86,3%
Samsung S5 HSIL	68,1%	92,2%
Samsung S5 Negative	78,3%	89,5%

4.3.3 Abnormal/Normal Classifier

Although the implementation of this second classifier had some problems with its training set, which consumed some time, the investigation to find out which classifier fitted the best for this kind of situation was easier. Since we had already done some investigation about the classifiers behavior previously (i.e. for the first classifier), the implementation of this one took that into account and settled the starting point already in the Adaboost classifier. Even so, just to corroborate that what we have found previously was the best method some experiments were made anyway.

Table 4.10: Decision Tree Classification Results

	true Normal	true Abnormal	Class Precision
pred. Normal	576	48	92,31%
pred. Abnormal	689	809	54,01%
Sensitivity	45,53%	94,40%	

This time the decision tree gave some promising results, as we can see on the table 4.10, also with cross-validation. Even though the classification for normal nuclei had a very low accuracy, the decision tree was almost sure when a nucleus was indeed abnormal. The problem was that the normal nuclei did not have any attribute that could be recognized for (i.e. with a high accuracy rate) but the abnormal ones had, since they were very different from the common ones. Many tries were made in order to improve these results, hanging on the good accuracy of the abnormal classification, but these experiments did not improve substantially the decision process, consequently discarding this classification method. Regarding the SVM, the same situation happened comparing to the first classifier. Although SVM could give good results, again, they were not very consistent as we tried with different images. Even with cross-validation, since OpenCV's SVM has a specific function for that, the results did not improve or in the case it did it was not substantial.

Again, the classifier which was the most robust and consistent, compared to the others, was the Adaboost. The only problem encountered in the implementation of this classifier was about the optic circle of the images. Although this could be a problem of the previous implemented classifier, this one is quite more delicate and so more sensible to exterior factors. The role of the optic circle is that the nearest are the objects identified to it, the blurrier they are. Since for the first classifier we just need to know if they are a nucleus or not, in this one we need to know, the best we can, the morphology of it. Obviously, if the object is very blurry we can not know for sure its morphology and so this kind of situation can totally ruin the machine learning process. To neutralize this kind of situations, many examples, which were near the optic circle, were removed from the dataset, including from the first one. Regarding the first classifier, no improvements were noticed but in this second phase, plus the changes that were made in the dataset regarding the labeling, the results were much better.

Concerning the implemented solution and its classification, the results achieved were very promising. The results had a sensitivity of above 50%, as we can confirm on the table 4.11. Although the percentage seems a bit low, we need to take into account that there was only 6 normal nuclei, thus missing only three of them makes the sensitivity drop a lot. Another particularity of this classifier is that it tends to classify more nuclei as abnormal than normal. This does not mean that this is a bad classifier but both the dataset and the actual classifier are built in order to identify even the lesser morphological abnormalities. Considering that the goal is to give the clinicians an idea of situation regarding the cervical cancer, all the unusual nuclei most be identified and then it is the surgical pathologist's decision if they mean cervical cancer or not. An example of this classifier is shown in the image 4.11 and the result's table 4.11.

Table 4.11: Classification of the image 4.11

	true Normal	true Abnormal	Class Precision
pred. Normal	3	1	75%
pred. Abnormal	3	6	66,6%
Sensivity	50%	85,7%	

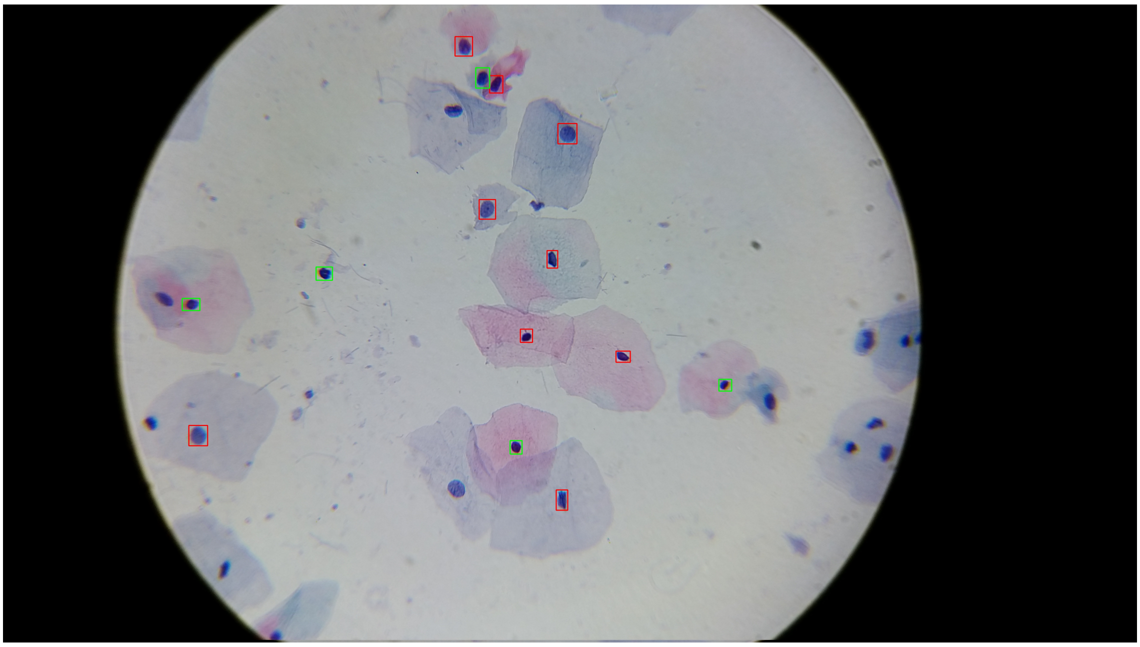


Figure 4.11: Output image from the second classifier, which gave the results on the table 4.11

4.3.3.1 Validation

Compared to the first classifier's validation, this one had some problems in between. Since some of the objects classified as nuclei in the Nuclei/Non-Nuclei classifier do not correspond to a true nucleus, the sensitivity and precision in this validation process were lower compared to the first one. Considering this, two set of tables were created in order to exhibit this problem. These tables are shown in the appendix A in the section A.2. The table 4.12 presents the difference between the Normal classification sensitivity and the table 4.13 the difference between the Abnormal classification. Firstly, in the table 4.13, we can notice that the negative sample always give a sensitivity of 0% for the Abnormal classification. That is normal, considering that the negative sample do not contain any evidence of cervical cancer. Another aspect that we can observe is the fact that the miss classification of the Nuclei/Non-Nuclei classifier spoil the results of this one. By ignoring the non-nuclei objects, for example, in the Samsung S5 HSIL test, we improved the Abnormal Sensitivity by 25,2% and the Normal Sensitivity by 9%. Thus, we can conclude that although this classifier had some setbacks regarding its dataset, if the first classifier achieved better results, it was also possible to achieve a bigger accuracy rate regarding the Abnormal/Normal classification.

Table 4.12: Overview of the Normal Classification

	Normal Sensitivity Not Ignoring	Nuclei Sensitivity Ignoring
HTC One M8 HSIL	61,3%	79,2%
HTC One M8 Negative	96,2%	98,4%
LG Nexus HSIL	82,9%	85,8%
LG Nexus Negative	96,2%	98,2%
Samsung S5 HSIL	72,2%	81,2%
Samsung S5 Negative	75%	77,9%

Table 4.13: Overview of the Abnormal Classification

	Abnormal Sensitivity Not Ignoring	Abnormal Sensitivity Ignoring
HTC One M8 HSIL	78,5%	84,6%
HTC One M8 Negative	0%	0%
LG Nexus HSIL	58,4%	75%
LG Nexus Negative	0%	0%
Samsung S5 HSIL	65,9%	91,1%
Samsung S5 Negative	0%	0%

4.3.4 Final Remarks

Since it was not possible to fully ignore all the unnecessary objects during the image processing module, besides the cervical cancer classification, machine learning was also used to help the tool to filter undesired information. It is safe to say that the first classifier, which did the filtering job, was the harder to implement, since all the investigation was done during its implementation. Also, this classifier would comprise the results of the second if it did not perform well. If we had too many non-nuclei objects, the second classifier would have terrible rates. The goal of both of them was to balance the false positive as well as the false negatives, so we did not try to classify a lot of non-nuclei objects but at same time did not ignore abnormal nuclei.

Chapter 5

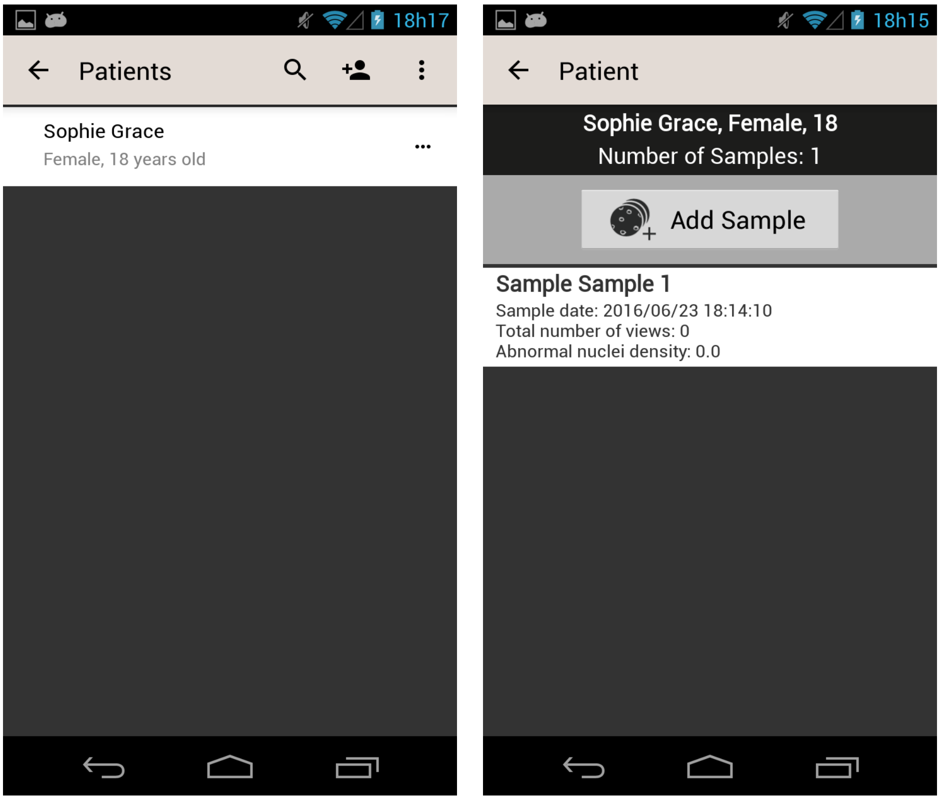
Mobile Integration

In order to fulfill all the objectives proposed in the beginning of this project, the integration of the developed tool into an Android application was done. The difference between the already marketed solutions and this one is, exactly, the portable nature of it. This way the tool could assist the clinicians anywhere, making cervical cancer diagnosis possible, even in the most remote locations.

As referred, the porting to Android was not made into a brand new application, but instead to an already developed one. The MalariaScope was created by Fraunhofer AICOS and it was used for a previous project regarding the malaria disease detection. Even though the cervical cancer and the malaria are totally different diseases, the core features are the same making the application easy to adapt. Also, this application had already some improvements in order to work on low-tier smartphones. This characteristic was very important taking into account the large quantity of image processing that this tool has. Even with an already adapted application for this kind of situations, the smartphones still have lower computational capabilities compared to a computer. The average time of the whole process, on a computer, is roughly one minute, something that in a smartphone would be impossible. The problem is that the image processing consumes most of the computation resources, due to the complex algorithms it uses. The time for the machine learning process to happen can be disregarded, since already trained files were included in the application. This was very important because this way there is no training time. Considering that all the attributes and values for the classifiers to take into account are specified in that file, there is no need for reading large sized datasets. Nevertheless, the benefits overcome these disadvantages, because the alternative would be not existing any kind of diagnosis. Currently, this application only works on offline mode. This is a core capability of this application, considering that most of the developing countries do not have the same Internet coverage as other countries. This way all of the diagnosis is done at the spot and all the information is stored internally in the smartphone. Some projects regarding the communication with a server are already in development at Fraunhofer but were not deployed yet. All the images that this application processes is captured by the device's camera, with special microscope adapters, or by transferring them from a computer.

Regarding the actual implementation of the solution into the Android programming language, some considerations were taken into account. Even though it was possible to use the already developed C++ code on the Android application, which spares several time, some modification needed to be done. By using the Android NDK this transaction is possible. Android NDK, stands for Android Native Development kit, which enables the use of native languages, like C++, for Android applications. This is useful, given the fact that these languages have far more capabilities that the Android one. Finally, in the figures 5.1, 5.2 and 5.3 some use-case examples of this application are shown.

Paying attention to the first set of figures 5.1 we have two screenshots showing the list of patients, on the left, and the list of each patients' samples, on the right. This application can hold record of several patients and samples associated to them. We can also add multiple views of one samples, as we can see on left in the set of figures 5.2. After adding some samples, they appear as blocks on a grid, as we can see on the left. Each view can be screened right away or just be saved and screened later. As the application finishes screening each view, the blocks turn into a different color to indicate evidence of cervical cancer, as we may notice on the left of the set of figures 5.3. Red represents high risk and Green low or non-existent. The last image, on the right, shows how the screening results are presented, showing the total number of nuclei in the view and consequently the number of normal and abnormal nuclei.



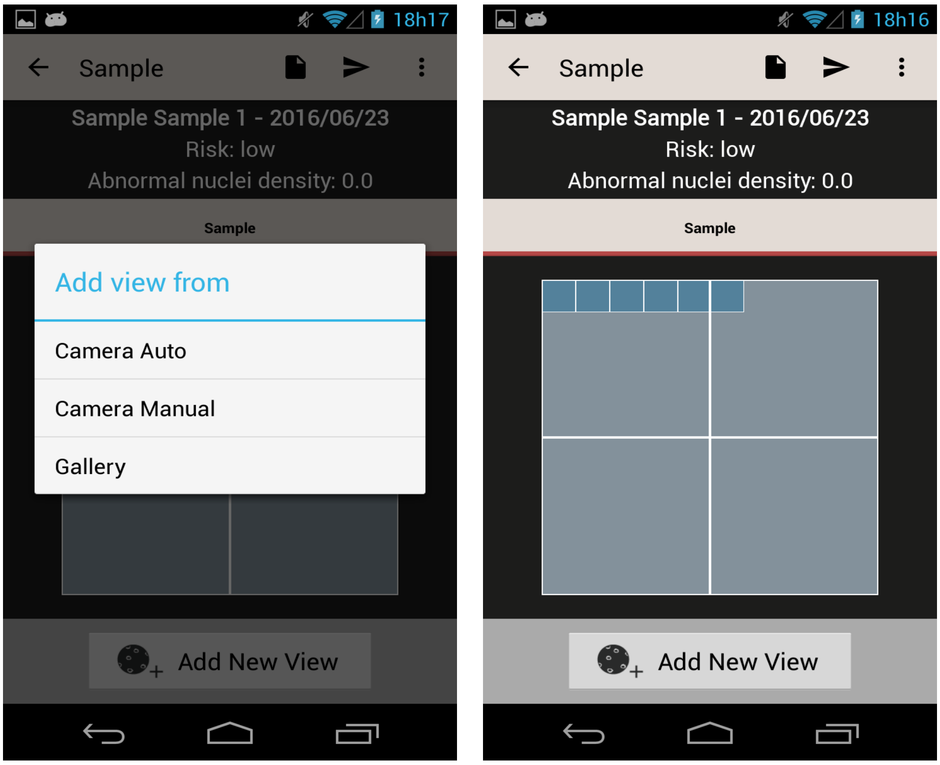
Sophie Grace
Female, 18 years old

Sophie Grace, Female, 18
Number of Samples: 1

Sample Sample 1
Sample date: 2016/06/23 18:14:10
Total number of views: 0
Abnormal nuclei density: 0.0

Figure 5.1: On the left, the list of the patients and on the right the list of samples of the same patient

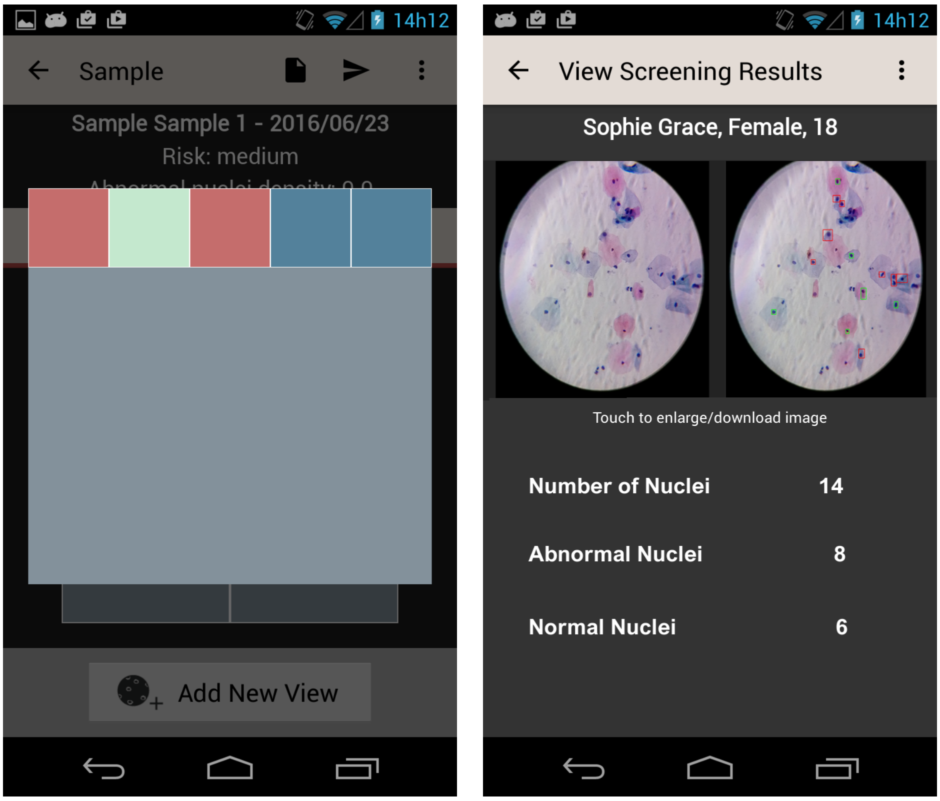
Figure 5.2: On the left, when we click right the different possibilities of adding images. On the right the views display.



Sample Sample 1 - 2016/06/23
Risk: low
Abnormal nuclei density: 0.0

Sample Sample 1 - 2016/06/23
Risk: low
Abnormal nuclei density: 0.0

Figure 5.3: On the left, the color changes when samples are screened. On the right, the screening results of a view.



Sample Sample 1 - 2016/06/23
Risk: medium
Abnormal nuclei density: 0.0

Sophie Grace, Female, 18

Touch to enlarge/download image

Number of Nuclei	14
Abnormal Nuclei	8
Normal Nuclei	6

Chapter 6

Conclusion and Future Work

Most countries in the World still lack medical equipment or technical expertise. Consequently, the people that live on these countries still do not have any surveillance for certain diseases. This enables the opportunity for cancers, like the cervical cancer, to take time to develop into later stages. This fact is one of the main reasons that the cervical cancer still kills in developing countries. As we have seen in the chapter 2, the solutions that already exist which can perform the screening of the cervical cancer, do not overcome this problem. Considering these facts, a new and low-cost tool was created in order to help to suppress them.

The detection of atypical cervical cells is a time consuming task. Firstly, the surgical pathologist of the screener must analyze images from the samples gathered during the gynecological observation. On each picture, the intention is to detect any kind of abnormality for each and every nuclei there is. The increasingly progresses in the computer vision field, made possible for this task to be made automatically. This defines the first stage of this cervical cancer screening tool. Also, the objects that are relevant for the screening are enhanced, in order for their features to be successfully extracted during the next stage. Considering that, during the first stage, segmenting the objects into nuclei and non-nuclei was a hard task, the first step of the machine learning module classifies them between these two groups. Finally, when we are left with the nuclei, we can classify them considering their morphology. This last classification will separate the normal from the abnormal nuclei, giving an idea about the presence of atypical cervical cells, some of which could be suspicious or even positive for cervical cancer.

During the development of this project we aimed to implement algorithms and techniques which could give good and trustworthy results. Likewise, these intentions endured when dealing with the creation of the dataset. In order to create this dataset, smartphones from three different brands were used to acquire the images. This way we could have some degree of variability in the quality of the images and consequently a more general training set. The dataset was created with several images, where 66% of them were taken from a cervical cancer positive sample and 34% from a negative sample. The same dataset was used for both the classifiers. The only difference between them was that the second dataset is a filtered version of the first. Labeling these datasets was also very important. With the help of a specialist, every entry was labeled as Nuclei/Non-

Nuclei and Abnormal/Normal, so we could use both of them for training the classifiers. Finally, another set of images was taken using the same smartphones, with the intention of testing both of the implemented classifiers.

To begin with, it was imperative to pre-process the images. To achieve this, we started by enhancing the valuable objects in the images, so we could extract their information in the next step. Initially, the goal was to identify the whole cell's structure, however it was discarded due to new information that was given by a surgical pathologist. Instead we only identify the cells' nuclei, since it was where the valuable information resided. The algorithm to detect the nuclei could also achieve far better results than the previous one. The reason behind this is because the nuclei are much brighter and intense than the cell's cytoplasm. Thus, during the pre-processing the number of identified nuclei exceed the number of identified cells.

As some of the problems were not tackled during the image processing module, two classifiers were implemented instead of one to identify cervical cancer. Even though the Nuclei/Non-Nuclei classifier was not in the initial plan of this project it was proved that it could retrieve good results, correctly splitting, into these two groups, most of the objects in the sample. This classifier was specially important not only for creating the second training set but also for filtering the undesired objects for the Abnormal/Normal classifier. The chosen classifier for this situation was the Adaboost, since it gave the best results. This classifier filters most of the non-nuclei objects. Out of 218, only 19 of them are considered nuclei, which results in a sensitivity of 91,2%, as is evidenced by the table A.2.

As expected, the Abnormal/Normal classifier ended up defining the success of the project. Like the previous classifier and considering that we were practically using the same dataset, also the Adaboost classification gave better results. Even though the accuracy rate is high, as we can see in the table A.17, there are more false positives than false negatives. This can be considered an advantage since we will highlight, possible, developing abnormal nuclei. Then, as this will be a screening tool, the screener will evaluate them and decide either they show or not evidence of cervical cancer.

Ultimately the tool was deployed as an Android application, for utilization purposes. This choice was based on the fact that the Android operating system works on smartphones which are cheaper when compared to most computers or any other complex equipment. This issue has always been taken into account, considering that this tool was developed for countries which do not have the capabilities of spending a lot of money on new equipment.

Despite the fact that the implementation was successful, some improvements could be done to this tool. Although the majority of the nuclei are identified during the image processing module, if the images were standardized before being pre-processed we could achieve better results. The standardization would consist on equalizing the image features and also making them to have the same size. This would improve the results because we could focus on a specific set of characteristics, and create the algorithm accordingly, instead of doing a more generalist approach. This could help during the segmentation phase, due to the fact that the objects would be identified with a higher accuracy. Splitting contours of nuclei that are identified together, could be another improvement done to the segmentation phase. By doing so, more nuclei could be identified and less opportunities of finding cervical cancer would be missed. Regarding the Abnormal/Normal classifier, some improvements to its training set could also be done. During the development of this project we only had access to two groups of samples, one positive and one negative. By having access to multiple of these, we could create an even more diverse dataset, which could help avoiding overfitting and achieve better results. Finally, some future work regarding the Android application could also be done. Considering that a smartphone does not have the same specifications compared to a computer, the application performance will be slightly slower. Thus, the tool could be optimized in order to increase the processing speed. Lastly, an autofocus algorithm could be implemented. This algorithm could improve the quality of images while taking them with a smartphone, consequently improving the quality of the whole process.

References

[Ada] Adaboost algorithms. https://en.wikipedia.org/wiki/AdaBoost. Accessed: 25-05-2016.

[AMS] Android market share. https://d28wbuch0jlv7v.cloudfront.net/images/infografik/normal/chartoftheday_4431_smartphone_operating_system_market_share_n.jpg. Accessed: 01-06-2016.

[BM14] Ewert Bengtsson and Patrik Malm. Screening for cervical cancer using automated analysis of PAP-smears. *Computational and Mathematical Methods in Medicine*, 2014, 2014.

[Can] Treatment options for cervical cancer, by stage. http://www.cancer.org/cancer/cervicalcancer/detailedguide/cervical-cancer-treating-by-stage. Accessed: 11-02-2016.

[CBB$^+$03] Massimo Confortini, Lucia Bonardi, Paolo Bulgaresi, Maria Paola Cariaggi, Silvia Cecchini, Stefano Ciatto, Ida Cipparrone, Laura Galanti, Cristina Maddau, Marzia Matucci, Tiziana Rubeca, Grazia Maria Troni, Patricia Turco, Marco Zappa, and Francesca Carozzi. A feasibility study of the use of the AutoPap screening system as a primary screening and location-guided rescreening device. *Cancer*, 99(3):129–134, 2003.

[CCS] Cervical cancer automated screening solutions. http://www.hindawi.com/journals/cmmm/2014/842037/. Accessed: 12-02-2016.

[Cro] Cross-validation. https://en.wikipedia.org/wiki/Cross-validation_(statistics). Accessed: 25-05-2016.

[CTUA14] Thanatip Chankong, Nipon Theera-Umpon, and Sansanee Auephanwiriyakul. Automatic cervical cell segmentation and classification in Pap smears. *Computer Methods and Programs in Biomedicine*, 113(2):539–556, 2014.

[DHS09] Edward R Dougherty, Jianping Hua, and Chao Sima. Performance of feature selection methods. *Current genomics*, 10(6):365–74, 2009.

[hig] High level image processing. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MARBLE/high/high.htm. Accessed: 08-02-2016.

[Hpv] Human papillomavirus (hpv): Causes, symptoms and treatments. http://www.medicalnewstoday.com/articles/246670.php?page=2. Accessed: 11-02-2016.

[HRDP06] World Health Organization. Reproductive Health, Research, World Health Organization. Chronic Diseases, and Health Promotion. *Comprehensive Cervical Cancer Control: A Guide to Essential Practice*. Integrating Heath Care for Sexual and Reproductive Health and Chronic Diseases. World Health Organization, 2006.

[KLS^{+}94] L G Koss, E Lin, K Schreiber, P Elgert, and L Mango. Evaluation of the PAPNET cytologic screening system for quality control of cervical smears. *American journal of clinical pathology*, 101(2):220–9, 1994.

[KMN^{+}02] Tapas Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, Ruth Silverman, and A.Y. Wu. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892, 2002.

[KS08] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–3, 2008.

[LCy] Thinprep pap test. http://www.thinprep.com/hcp/thinprep_difference/beyond_conventional_pap.html. Accessed: 12-02-2016.

[Log] Logit boost algorithm. https://en.wikipedia.org/wiki/LogitBoost. Accessed: 25-05-2016.

[low] Low level image processing. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MARBLE/low/low.htm. Accessed: 08-02-2016.

[Mac] Machine learning algorithms. http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/. Accessed: 01-06-2016.

[Opea] Opencv website. http://opencv.org/about.html. Accessed: 12-02-2016.

[Opeb] Opencv's adaboost. http://docs.opencv.org/2.4/modules/ml/doc/boosting.html. Accessed: 25-05-2016.

[PAH^{+}09] Tune H. Pers, Anders Albrechtsen, Claus Holst, Thorkild I A S??rensen, and Thomas A. Gerds. The validation and assessment of machine learning: A game of prediction from high-dimensional data. *PLoS ONE*, 4(8), 2009.

[Pap] Papanicolaou test wikipedia page. https://en.wikipedia.org/wiki/Pap_test. Accessed: 12-02-2016.

[PLP09] Jong Hyun Park, Guee Sang Lee, and Soon Young Park. Color image segmentation using adaptive mean shift and statistical model-based methods. *Computers & Mathematics with Applications*, 57(6):970–980, 2009.

[QNR^{+}09] M R Quddus, T Neves, M E Reilly, M M Steinhoff, and C J Sung. Does the ThinPrep Imaging System increase the detection of high-risk HPV-positive ASC-US and AGUS? The Women and Infants Hospital experience with over 200,000 cervical cytology cases. *Cytojournal*, 6:15, 2009.

[RCEc16] Luis Rosado, Jose Manuel Costa, Dirk Elias, and Jaime cardoso. Automated detection of malaria parasites on thick blood smears via mobile devices. In *Procedia Computer Science*, page TBD, 2016.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[SDN05] Yu Sun, Stefan Duthaler, and Bradley J. Nelson. Autofocusing algorithm selection in computer microscopy. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 419–425, 2005.

[SGKC$^+$13] Hélène Sancho-Garnier, Youssef Chami Khazraji, Moktar Hamdi Cherif, Abbes Mahnane, Mohamed Hsairi, Amr El Shalakamy, Nejat Osgul, Murat Tuncer, Aisha O. Jumaan, and Muhieddine Seoud. Overview of cervical cancer screening practices in the extended middle east and north africa countries. *Vaccine*, 31, Supplement 6:G51 – G57, 2013. Comprehensive Control of {HPV} Infections and Related Diseases in the Extended Middle East and North Africa Region.

REFERENCES

Appendix A

2 Appendix 1

A.1 Results from the Validation of the Nuclei/Non-Nuclei Classifier

Table A.1: Results of Nuclei/Non-Nuclei classification of the HSIL images taken by HTC One M8

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	59	24	71,0%
pred. Nucleus	14	60	81%
Sensitivity	80,8%	71,4%	

Table A.2: Results of Nuclei/Non-Nuclei classification of the negative images taken by HTC One M8

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	199	20	90,8%
pred. Nucleus	19	129	87,2%
Sensitivity	91,2%	92,1%	

Table A.3: Results of Nuclei/Non-Nuclei classification of the negative images taken by LG Nexus

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	238	16	93,7%
pred. Nucleus	51	116	69,4%
Sensitivity	82,3%	87,8%	

Table A.4: Results of Nuclei/Non-Nuclei classification of the HSIL images taken by LG Nexus

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	60	23	72,2%
pred. Nucleus	18	145	88,9%
Sensitivity	76,9%	86,3%	

Table A.5: Results of Nuclei/Non-Nuclei classification of the HSIL images taken by Samsung S5

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	45	7	86,5%
pred. Nucleus	21	83	79,8%
Sensitivity	68,1%	92,2%	

Table A.6: Results of Nuclei/Non-Nuclei classification of the negative images taken by Samsung S5

	true Not Nucleus	true Nucleus	Class Precision
pred. Not Nucleus	58	9	86,5%
pred. Nucleus	16	77	82,7%
Sensitivity	78,3%	89,5%	

A.2 Results from the Validation of the Abnormal/Normal Classifier

A.2.1 Results Not Ignoring Non-Nuclei Objects

2

Table A.7: Results of Abnormal/Normal Classifier classification of the HSIL images taken by HTC One M8

	true Normal	true Abnormal	Class Precision
pred. Normal	38	3	92,6%
pred. Abnormal	24	11	31,4%
Sensitivity	61,3%	78,5%	

Table A.8: Results of Abnormal/Normal Classifier classification of the negative images taken by HTC One M8

	true Normal	true Abnormal	Class Precision
pred. Normal	127	17	88,1%
pred. Abnormal	5	0	0%
Sensitivity	96,2%	0%	

Table A.9: Results of Abnormal/Normal Classifier classification of the HSIL images taken by LG Nexus

	true Normal	true Abnormal	Class Precision
pred. Normal	73	32	69,5%
pred. Abnormal	15	45	75%
Sensitivity	82,9%	58,4%	

Table A.10: Results of Abnormal/Normal Classifier classification of the negative images taken by LG Nexus

	true Normal	true Abnormal	Class Precision
pred. Normal	114	40	74%
pred. Abnormal	13	0	0%
Sensitivity	96,2%	0%	

Table A.11: Results of Abnormal/Normal Classifier classification of the HSIL images taken by Samsung S5

	true Normal	true Abnormal	Class Precision
pred. Normal	39	16	70,9%
pred. Abnormal	15	31	67,3%
Sensitivity	72,2%	65,9%	

Table A.12: Results of Abnormal/Normal Classifier classification of the negative images taken by Samsung S5

	true Normal	true Abnormal	Class Precision
pred. Normal	60	13	82,1%
pred. Abnormal	20	0	0%
Sensitivity	75%	0%	

A.2.2 Results Ignoring Non-Nuclei Objects

Table A.13: Results of Abnormal/Normal Classifier classification of the images taken by HTC One M8

	true Normal	true Abnormal	Class Precision
pred. Normal	38	2	95%
pred. Abnormal	10	11	52,3%
Sensitivity	79,2%	84,6%	

Table A.14: Results of Abnormal/Normal Classifier classification of thef negative images taken by HTC One M8

	true Normal	true Abnormal	Class Precision
pred. Normal	127	0	100%
pred. Abnormal	2	0	0%
Sensitivity	98,4%	0%	

Table A.15: Results of Abnormal/Normal Classifier classification of the HSIL images taken by LG Nexus

	true Normal	true Abnormal	Class Precision
pred. Normal	73	15	82,9%
pred. Abnormal	12	45	78,9%
Sensitivity	85,8%	75%	

Table A.16: Results of Abnormal/Normal Classifier classification of the negative images taken by LG Nexus

	true Normal	true Abnormal	Class Precision
pred. Normal	114	1	99,1%
pred. Abnormal	2	0	0%
Sensitivity	98,2%	0%	

Table A.17: Results of Abnormal/Normal Classifier classification of the HSIL images taken by Samsung S5

	true Normal	true Abnormal	Class Precision
pred. Normal	39	3	92,8%
pred. Abnormal	9	31	77,5%
Sensitivity	81,2%	91,1%	

Table A.18: Results of Abnormal/Normal Classifier classification of the negative images taken by Samsung S5

	true Normal	true Abnormal	Class Precision
pred. Normal	60	0	0%
pred. Abnormal	17	0	0%
Sensitivity	77,9%	0%	