UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Matemáticas y Computacionales

Automatic classification of medical images

Trabajo de integración curricular presentado como requisitopara la obtención
del título de Ingeniero en tecnologías de la Información

Autor:

Selena Jahaira Jiménez Lara

Tutor:

Manuel Eugenio Morocho Cayamcela, Ph.D.

Urcuquí - Marzo, 2021

Urcuquí, 16 de marzo de 2021

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES
CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN
ACTA DE DEFENSA No. UITEY-ITE-2021-00001-AD

A los 16 días del mes de marzo de 2021, a las 13:30 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.
Miembro No Tutor	Dr. CUENCA LUCERO, FREDY ENRIQUE , Ph.D.
Tutor	Dr. MOROCHO CAYAMCELA, MANUEL EUGENIO , Ph.D.

El(la) señor(ita) estudiante **JIMENEZ LARA, SELENA JAHAIRA**, con cédula de identidad No. **0950076539**, de la **ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES**, de la Carrera de **TECNOLOGÍAS DE LA INFORMACIÓN**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-43-No.496-2014**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **AUTOMATIC CLASSIFICATION OF MEDICAL IMAGES**, previa a la obtención del título de **INGENIERO/A EN TECNOLOGÍAS DE LA INFORMACIÓN.**

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

Tutor	Dr. MOROCHO CAYAMCELA, MANUEL EUGENIO , Ph.D.

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Presidente Tribunal De Defensa	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.	9,0
Tutor	Dr. MOROCHO CAYAMCELA, MANUEL EUGENIO , Ph.D.	9,0
Miembro Tribunal De Defensa	Dr. CUENCA LUCERO, FREDY ENRIQUE , Ph.D.	9,7

Lo que da un promedio de: **9.2 (Nueve punto Dos)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

Certifico que *en cumplimiento del Decreto Ejecutivo 1017 de 16 de marzo de 2020, la defensa de trabajo de titulación (o examen de grado modalidad teórico práctica) se realizó vía virtual, por lo que las firmas de los miembros del Tribunal de Defensa de Grado, constan en forma digital.*

JIMENEZ LARA, SELENA JAHAIRA
Estudiante

Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.
Presidente Tribunal de Defensa

Dr. MOROCHO CAYAMCELA, MANUEL EUGENIO , Ph.D.
Tutor

Dr. CUENCA LUCERO, FREDY ENRIQUE , Ph.D.
Miembro No Tutor

TORRES MONTALVÁN, TATIANA BEATRIZ
Secretario Ad-hoc

Autoría

Yo, **Selena Jahaira Jiménez Lara**, con cédula de identidad **0950076539**, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el autor del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Marzo 16 del 2021.

Selena Jiménez

Selena Jahaira Jiménez Lara
CI: 0950076539

Autorización de publicación

Yo, **Selena Jahaira Jiménez Lara**, con cédula de identidad **0950076539**, cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorizacion escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Urcuquí, Marzo del 2021.

Selena Jiménez

Selena Jahaira Jiménez Lara
CI: 0950076539

Acknowledgments

Agradezco a mi familia, a mis amigos, a mis compañeros de clases, y a mí misma por poder sobrellevar el trabajo en tiempos complicados, donde la esperanza parecía más lejana que nunca. A, mi profesora y amiga, Lorena Guachi, quien me enseñó y quien me ha apoyado en todas las ocasiones necesarias, tanto en clases como en lo personal, compartiendo su conocimiento. Agradezco a, mi tutor y amigo, Diego Peluffo, quien me ayudó y comprendió en todo momento y fue la pieza más importante y necesaria para el desarrollo de este trabajo investigativo, ya que me ha enseñado todo lo necesario para llevar a cabo la investigación con mucha paciencia, y por demostrarme que todos tenemos nuestro lado humano

Abstract

Lung cancer is a disease where abnormal cell multiplying and growing into a tumor. This type of cancer is diagnosed via analyzing Computed Tomography images (CT images) from patients. In this sense, for medical and diagnostic purposes, the quality of the images is a key factor in detecting any type of abnormality. Also, early diagnosis and treatment can save the life of the patient. Given improvements on both lung cancer tomography and regions-ofinterest identification techniques, it has been possible to produce clearer images to be subsequently analyzed. Additionally, through image processing techniques and machine learning, the regions of interest can be identified and classified between normal and pathological. This means that a computerassisted diagnosis is possible, which can support physicians in the process of accurately identifying cancer cells. At present, there exists a wide range of computer-aided diagnostic approaches for lung cancer. In this research, an exploratory study is carried out in order to evaluate basic image processing and machine learning techniques to identify and classify regions of interest in tomography images to diagnose lung cancer. Particularly, in this work, the use of basic techniques is preferred for the sake of lower computational complexity. For experiment purposes, several basic image processing techniques were implemented, such as: media filter, Gaussian blurring, Gabor filter, thresholding-based segmentation technique, the morphological erosion operator, and watershed segmentation. Subsequently, a feature extraction process was carried out, which produced a data set. Such a data set is used to train a binary classifier based on support vector machines. As an overall result, it was obtained that the performance of the automatic classification, in terms of confusion-matrix-based measurements, amounts to 91 % of precision, 98 % of sensitivity, and 85 % of specificity.

Keywords: **Digital tomography, early detection, Gabor function, image processing, thresholding-based segmentation**

Resumen

El cáncer de pulmón es una enfermedad en la que las células se multiplican anormalmente y se convierten en un tumor maligno. Este tipo de cáncer se diagnostica mediante el análisis de imágenes de tomografía computarizada de los pacientes. En este sentido, para efectos médicos y diagnósticos, la calidad de las imágenes es un factor importante para detectar cualquier tipo de anomalía. Con la mejora de las tomografías de cáncer de pulmón y las técnicas de identificación de las regiones de interés, se ha logrado producir imágenes más claras para ser subsecuentemente analizadas. Adicionalmente, a través de técnicas de procesado de imágenes y aprendizaje automático, se puede identificar las regiones de interés y clasificarlas entre normales y patológicas. Es decir que es posible llevar a cabo un diagnóstico asistido por computadora, el cual puede apoyar a los médicos en el proceso de identificación de células cancerosas con precisión. En la actualidad, existen diversos enfoques de diagnóstico asistido por computador para cáncer pulmonar. En esta investigación, se realiza un estudio exploratorio con el fin de evaluar técnicas básicas de procesado de imágenes y aprendizaje automática para identificar y clasificar regiones de interés en imágenes de tomografía para diagnosticar cáncer de pulmón. Particularmente, en este trabajo, se prefiere el uso de técnicas básicas con el fin de reducir la complejidad computacional. Para efectos de experimentación, se implementaron varias técnicas básicas de procesado de imágenes, tales como: filtro de media, desenfoque Gaussiano, filtro de Gabor, técnica de segmentación por umbral, el operador morfológico de erosión, y segmentación de cuencas. Posteriormente, se llevó a cabo un proceso de extracción de características, el cual produjo un conjunto de datos. Dicho conjunto de datos se usó para entrenar un clasificador binario basado en máquinas de vectores de soporte. Como resultado general, se obtuvo que el desempeño de la clasificación automática, en términos de mediciones basadas en matriz de confusión, asciende a 91% de precisión, 98% de sensibilidad y 85% de especificidad

Palabras Clave: **Cáncer pulmonar, detección temprana, función de Gabor, procesamiento de imágenes, segmentación por umbral, tomografía digital**

Contents

Agradecimientos **ii**

Resumen **iii**

Abstract **iv**

1 Introduction **1**
- 1.1 Problem Statement 2
- 1.2 Document Organization 2
- 1.3 Objectives 3
  - 1.3.1 General Objective 3
  - 1.3.2 Specific Objectives 3

2 Overview and background **4**
- 2.1 Lung Cancer 4
  - 2.1.1 Diagnosis 5
- 2.2 Data Collection 6
- 2.3 Image Processing 6
- 2.4 Feature Extraction 6
- 2.5 Machine learning 7
  - 2.5.1 Support vector machines 8
  - 2.5.2 Classification 9
- 2.6 Related works 9

3 Methodology **12**
- 3.1 Introduction 12
- 3.2 Database 13
- 3.3 Image Pre-Processing 14
  - 3.3.1 Gray Scaled 14
  - 3.3.2 Median Blur Filter 15
  - 3.3.3 Gaussian Noise Filter 16
  - 3.3.4 Gabor Filter 17
- 3.4 Image Segmentation 18

3.4.1 Threshold Segmentation 19
3.4.2 Morphological Operator: Erode 19
3.4.3 Watershed Segmentation and Markers 20
3.5 Feature Extraction 21
3.5.1 Area 21
3.5.2 Perimeter 21
3.5.3 Eccentricity 21
3.6 SVM Classification 22

4 Experimental Setup **23**
4.1 Performance metrics 23
4.2 Experimental Settings 24
4.3 Applied Tests on Database 25

5 Results **27**
5.1 Experimental setup 27
Hardware Description 27
5.2 Results of the Generic Data 27
5.3 Features Extraction 31
5.4 Experiments 32
5.4.1 Experiment 1 32
5.4.2 Experiment 2 34
5.4.3 Experiment 3 36

6 Final remarks **39**
6.1 Conclusions 39
6.1.1 Future Research 40

Bibliography **41**

List of Figures

2.1 CT image of diagnosed patient 5
2.2 CT image of not diagnosed patient 5

3.1 Flow graph of lung cancer detection development 12
3.2 CT images of a 40 year old female no diagnosed, subject with a family history of lung cancer was CT screened for lung cancer. 13
3.3 CT images of a female who smoked, diagnosed with NSCLC, BAC. 14

5.1 Side by side in (A) an image of a diagnosed patient, (B) an CT image of a NOT diagnosed patient. 28
5.2 CT images after the application of median blur filter. 28
5.3 CT images after the application of Gaussian noise filter. 29
5.4 CT images after the application of Gabor Filter. 29
5.5 CT images after the application of morphological operator: Erode. 30
5.6 CT images after watershed segmentation, ... 30
5.7 CT images after the application of the markers algorithm. . . . 31
5.8 Experiment 1 - Comparison of iterations through their accuracy. 33
5.9 Experiment 1 - Comparison of iterations through their sensitivity. 34
5.10 Experiment 1 - Comparison of iterations through their specificity. 34
5.11 Experiment 2 - Comparison of iterations through their accuracy. 35
5.12 Experiment 2 - Comparison of iterations through their sensitivity. 36
5.13 Experiment 2 - Comparison of iterations through their specificity. 36
5.14 Experiment 3 - Comparison of iterations through their accuracy. 37
5.15 Experiment 3 - Comparison of iterations through their sensitivity. 38
5.16 Experiment 3 - Comparison of iterations through their specificity. 38

List of Tables

4.1 Variable setup of the different parameters used for Python implementation. 25

5.1 Obtained results for the extracted features of a subset of 15 CT images. 32
5.2 Experiment 1 - Comparison of iterations through mean and standard deviation measures. 33
5.3 Experiment 2 - Comparison of iterations through mean and standard deviation measures. 35
5.4 Experiment 3 - Comparison of iterations through mean and standard deviation measures. 37

List of Abbreviations

CT image: Computed tomography image.
SVM: Support vector machine.
ROI: Region of interest.

Chapter 1

Introduction

In modern days, lung cancer has progressed from an uncommon disease to one of the most serious diseases around the world, and the most common reason for death caused by cancer -leading the cause of cancer death for about 19% above average, in both men and women [1]. Lung cancer is a malignant tumor, i.e. a group of malignant cells which has been growing uncontrolled within the tissues of the lung [1]. One common symptom of this disease is coughing, plus chest pains and weight loss, however not all patients demonstrate those symptoms not even in the final stages. Some causes of Lung Cancer are the increase in smoking, chronic and autoimmune diseases, genetic, as well as environmental factors. Through both active or passive smoking, most of the cases are caused by tobacco consumption -amounting approximately 90% of the lung-cancer diagnosed patients [2].

The treatment to diagnose patients is defined by how much the disease is distributed across the lungs and the whole immune system of the patient. Still, most of the cases are not curable due to the late diagnosis. In order to diagnose lung cancer, the patient performs a chest radiography, since it can reveal the possible lumps in the chest which are the main reason to diagnose this type of cancer, plus it is used to get more information about the spread of the disease since as mentioned before, lung cancer evolves through different stages. Commonly, lung cancer shows up as a pulmonary lump on a chest radiography or CT image, plus it may reveal more information regarding the type and widespread of the disease [3].

Additionally, medical images such as computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) provide essential information for the detection, these methods had evolved as the prime methods for lung cancer screening. Besides, other diagnostic methods such as blood study may have future potential in lung cancer detection but currently, there is no research to support them [4]. Therefore, not much medical technologies are currently available that can take over from CT images, however, if a new method is considered then it must be compared to the diagnoses made with CT images, plus the efficiency of this new method should be higher than the already efficiency of CT images. Hence, that is the reason for the use of CT images in this exploratory study.

1.1 Problem Statement

The reason for the insufficient survival rate is that lung cancer is hard to detect since it just comes up with symptoms in a final [5]. In Ecuador, it tends to be diagnosed at late stages since there is no adequate screening for the patient. Most of the people who survive this cancer are because their cancer has been treated in the early stages. Therefore, the research of medical images accelerates the process of detection. In other words, given that lung cancer can still be cured when detected at an early stage, it is important to design computerized methods that facilitate its early detection.

Therefore, this thesis presents an exploratory study of lung cancer CT images classification into diagnosed and not diagnosed, applying image processing techniques and support-vector-machine-based classifiers.

1.2 Document Organization

This manuscript is structured into 5 chapters, namely: preliminaries, an overview background, methodology, analysis of results with discussion, and final remarks.

- Chapter 1 presents an introduction, along with the problem statement behind this research project. Additionally, the objectives had been settled in this chapter, both the general objective and specific objectives.
- Chapter 2 provides an overview of the background about the whole theory behind image processing in medical images, in our specific case, in lung cancer medical images. Moreover, the chapter presents an explanation of machine learning and the use of super vector machine(SVM). Furthermore, there can be found examples of existing implementations in this field. Thus, there are explanations of how these implementations were carried out.
- In Chapter 3, it is explained the methodology at the detail, the whole process that was carried out, and how they were performed. Section

3.2 provides information about the data-set used on the explanatory study, and where it was obtained. In Section 3.3, it is explained the process behind the image pre-processing. Next, in Section 3.4 outlines the explanation of the segmentation process, next at Section 3.5, the extracted features are described, and finally at 3.6, the used SVM model is explained.

- In chapter 4, it is explained the experimental settings and parameters used in the model.
- In Chapter 5, the analysis of the results is presented, and the experimental setup is discussed. Also, the performance of the experiments are shown and explained.
- Finally, in chapter 6, as the concluding chapter of this thesis,the finals remarks of the project are drawn, and some subsequent research for future works are stated.

1.3 Objectives

1.3.1 General Objective

- To develop an exploratory study of image processing techniques applied to lung cancer detection using benchmark segmentation approaches and support-vector-machine-based classifiers.

1.3.2 Specific Objectives

- To establish the image segmentation approaches, and the support-vector-machine classifiers, regarding recommendations by scientific literature, in order to identify those reaching a good trade-off between simplicity and readily interpretation.
- To design an experimental setup for assessing and comparing the established image segmentation and classification approaches on a lung cancer detection dataset through a comparative study.
- To validate the ability of the designed experimental setup for lung cancer region classification in terms of specificity, precision, and sensitivity.

Chapter 2

Overview and background

In this chapter, basics and information about topics of interests are presented, in order to provide an enough background to understand the proposed exploratory study. The application of machine learning in the field of medicine is relatively new since the years go back until recently of 50 years. Additionally, the importance of the pre-processing stage and segmentation section in CT images is explained, in order to classify them according to the diagnosis of the patient.

Incorporating machine learning algorithms in the field of medicine basically guarantees improvements of performance in it. Because the fieldwork time would be automated based on the patient's need. That is why in recent times, several private medical companies seek to create their own diagnoses techniques through machine learning [6].

However, to ensure a good classification, other processes are necessary beforehand, such as the pre-processing of the image to be studied, since the CT images of patients are to be further classified.

2.1 Lung Cancer

Lung cancer is a malignant disease produced by the malformation of tumors in the lung area, which is basically cells with uncontrolled growth, so the extension and size of the tumor vary depending on the case and the life span of the cancer [7]. People diagnosed with this disease should undergo specific treatment depending on how advanced their cancer is. In figure 2.1, we can observe the CT image of a patient diagnosed with lung cancer, in one of its early stages, while in Figure 2.2, it is shown the image of a patient who despite having a history of the disease in its genealogical tree is free of any tumor presence.

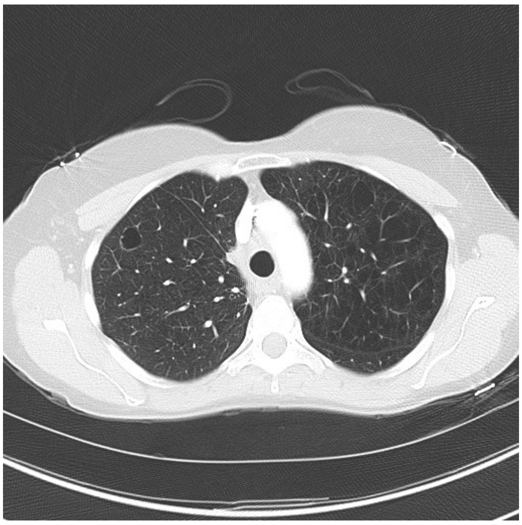

FIGURE 2.1: CT image of diagnosed patient

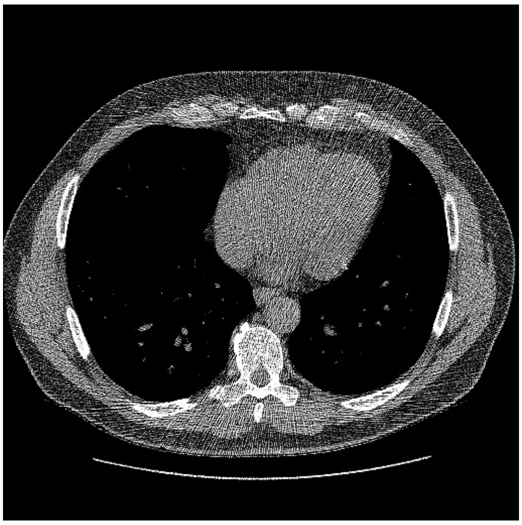

FIGURE 2.2: CT image of not diagnosed patient

It should be noted that in this exploratory study, we will study primary lung cancer, which is classified regarding the type of cells that begin to grow, namely:

- non-small-cell: the most common form, which represents more than 87% of cases. There are three types of non-small-cell: squamous cell carcinoma, adenocarcinoma, or large-cell carcinoma.
- small-cell: Less common form that spreads faster than non-small-cell lung cancer.

The sizes of normal lung nodules are between 0.5 cm to 2.5 cm, then if the nodule is greater than the normal size it can start to being considered malignant.

2.1.1 Diagnosis

To diagnose whether a person has lung cancer, a diagnosis from an expert doctor is needed. Then, the doctor studies the medical history, age, behaviors, and the obtaining of CT images. A pulmonary nodule can be categorized based on its size, location, shape, and texture.

Nowadays, lung cancer is detected by the use of CT images, which are medical images taken at the chest of the patient, that uses computer-processed combinations of many X-ray measurements. Several radiologists extract pulmonary information from several CT images taken. After that study, the diagnosis is revealed as the possibility of malignant nodules, so that the patient can follow a treatment plan [8]. Furthermore, the doctor has to do several tests so that the patient can be correctly diagnosed with lung cancer or any other kind of disease that could be.

2.2 Data Collection

Data collection is the very first step for doing this research specifically, regarding the automatic classification. The appropriate data to develop the exploratory study are CT images from real patients since such an information is totally reliable. The CT images have to be specifically from the chest or lung sector, besides in order to classify them it has to have a binary output so that the result is *diagnosed* or *notdiagnosed*. In other words, the given CT images have been validated by medical and scientific communities as well as properly labeled by experts regarding the diagnosis (normal or pathological). Also, the considered CT image database counts on information about the original size of the nodules (Region of Interest - ROI), which enables the assessment of automatic segmentation tools.

2.3 Image Processing

Image processing plays a vital role in diagnosing and analyzing diseases that require medical images as CT images. In the medical field CT images have background noise, irregular color changes, variations of the data due to factors such as human error, the status of the tomography machine, even the temperature.

Image processing techniques are applied to eliminate these problems that affect the extraction of needed features for the classification. Therefore, doctors can make a faster diagnosis for treatment still efficiently and accurately [9]. The images used in this research hold the same format so that image transformations were not required.

2.4 Feature Extraction

The feature extraction is a key stage to classify the CT images, since it defines the feature set (set of characteristic is) to subsequently train the automatic classification model [10]. For this purpose, a feature extraction procedure is applied to the segmented image. Since the machine learning method applied is supervised (as explained in Chapter 2), the data is also labeled with the information of lung-cancer-diagnosed or -not-diagnosed patient. Some important features extracted from the segmented images are based on previous studies [11], namely:

1. Area: Total number of white pixels in the extracted area.
2. Perimeter: Length of the perimeter of extracted ROI.
3. Eccentricity: It refers to the roundness, circularity or irregularity complex.

2.5 Machine learning

Machine learning is an area devoted to recognizing patterns from data, which can be used for several applications and in a wide range of fields. Indeed, it has been shown to be a powerful tool for medical images [12]. It is considered as a branch of artificial intelligence and is based on learning patterns and similarities from the examples. Nowadays, machine learning has been widely used in medical settings, such that it can be used for diagnosis, to categorize different diseases, and even offer a totally personalized medication plan based on the patient's needs.

Besides, in our research, we used a set of data which has relevant knowledge about itself, such as the ROI and being diagnosed or not diagnosed, so with this data, our algorithm system can learn from the training data such that it can be applied to the testing data and returns precise results.

However, the use of machine learning in terms of professionalism is kind of controversial since the medical field is sensible at diagnosing patients because their lives are at the line. On the other hand, the evident great performance of machine learning in the medical field is undeniable [10]. Then the use of these investigations will greatly help the study and early diagnosis of patients with diseases.

There is a broad range of machine learning techniques, which can be divided into two main groups [13]:

- **Supervised Learning:**

  Supervised learning encompasses predictive models that handle the problem of classification employing its input data and labels thereof. In other words, these methods need a set of classified data in order to train and get the results. Besides, these are supervised learning methods since it needs data labeled and classified so that the output answer is already given, then the training phase requires less computational complexity [13]. However, at the testing stage, the algorithm provides its own output answer. Some examples of supervised learning methods are Linear Discriminant Analysis (LDC), k-Nearest Neighbor (k-NN), Artificial Neural Networks (ANN), and the used method for this research, Support Vector Machine (SVM).

- **Unsupervised Learning:**

  Unsupervised learning is a predictive model too, but its principal characteristic is that the learning process is done from non-labeled data. Instead, it considers the similarities and differences among data, such that is no necessary prior knowledge, so the method extract relevant information and hierarchies of the given dataset [14]. Due to the consideration of similarities, unsupervised methods rely on distance measures, since the similarity between two samples can be understood as the pairwise distance between data points. Then, the lower the distance, the higher the probability of belonging to the same group [15]. Some well-known unsupervised learning methods are k-Means and Isodata clustering.

2.5.1 Support vector machines

Support vector machine (SVM) is a statistical and machine learning technique, which has as a primary goal to linearly identify patterns on data for prediction and classification. SVMs were initially conceived for binary classification, i.e. problems holding two classes. This approach was later extended to address continuous outcomes and classification with more than two classes. Besides, SVM can be applied to continuous, binary, and categorical outcomes analogous to Gaussian, logistic, and multinomial regression [16]. SVM can be used in three different approaches, as follows:

1. Binary: It is often referred to as classification in statistical learning.
2. Multinomial: It is used for multi-class classification.
3. Continuos: Simply called regresion.

Therefore, the method applied in this work is an SVM binary, since the labeled data, has a target of $true$ when diagnosed and $false$ at not diagnosed.

2.5.2 Classification

Classification implies to assign a class to a group of pixels after the CT image is segmented the ROI is marked such that the classifier then determines whether the ROI in the testing data represents diagnosed or not diagnosed [17]. The classification is used labeled data since SVM is developed. Then the labeled data is a set of samples, in this research are the extracted features from images, each with their correct answer or target output. Therefore, that is why the explanation of the extracted features in Section 2.4. Features are necessary to extract so that the relevant information for the classification task.

Then, from the extracted features, the whole dataset is divided into training data and testing data, such that the algorithm learns with the training data, so that it can further be applied the learnt model over the testing data [12]. In this research, the machine learning method developed is Support Vector Machine(SVM).

2.6 Related works

- **Analysis and comparison of image enhancement techniques for the prediction of lung cancer**
  Several researchers have proposed and implemented methods of lung cancer detection by using different approaches, including alternatives to image processing. Nowadays, there are plenty of methods on how to detect lung cancer through processing images. Corresponding to the work and specific implementation in precise image processing methods such as Gabor filters and watershed segmentation technique in [18]. A particular metric in the mentioned research is that the experiments were performed in both healthy people and infected people, in order to compare their results.

- **Lung cancer detection using digital image processing and artificial neural networks** Besides, new approaches were released to get better results, which have been developed by some researchers by taking advantage of artificial neural networks techniques [19]. The authors used the Lung CT-Diagnosis database from `The Cancer Imaging Archive` as their data-set. Their experiments were performed using 70 images, and in the processing image steps, first, histogram equalization was executed to get better contrast, after that, thresholding segmentation, since it is highly efficient in CT images, they use a value from 150 to 220, so its binary image can be appreciated. Finally, the feature extraction was completed and just after this process, the ANN was called which contains 6 input neurons and 2 output neurons, along with 12 hidden layer neurons. As a result, the researchers concluded that the system can perform even better if the ANN is trained using larger databases, enabling even its use on lung cancer detection.

- **Brain Tumor and lung cancer Detection Using Digital Image Processing Techniques** Besides in other cases, some researchers focus not on the whole lung cancer but at the same time applied these processing methods in other diseases such as [20]. The bases of this research are tumor detection based on segmentation and morphological operators. The segmentation method is used to separate the tumor area from the background and then morphological operators are applied to detect the tumor in Magnetic resonance imaging (MRI) and cancer cell in computerized tomography (CT) scan. The proposed method in this research started in the pre-processing image with the use of the Median filter since is less sensitive than the outliers and can deal with noise in a simple technique, after that, methods such as thresholding segmentation, being specific, was used maximum entropy method for threshold segmentation. Then, watershed segmentation was implemented but computing the local minimum of the image gradient as a marker. As mentioned, the techniques used to work with CT images are similar to the methods used in our research, however, there is not accurate results or percentages of the work accomplished.

- **Study of medical image processing techniques applied to lung cancer** Additionally, in the research [21], Moreno et al. review of image processing techniques applied to the study of lung cancer and focuses on two specific tasks involved in the normal processing workflow: lung nodule segmentation, and feature extraction for tumor classification and prognosis. This research claims that in order to develop an efficient method for the detection of lung cancer, some stages have to be performed. Thus, the processing steps in lung cancer images are the following:

    1. Lung segmentation: In this stage, the structures have to be differentiated, it means the background from the object of interest. This stage has complexity since the lung structure has similar intensity values.

2. Tumor detection: In this stage, the region of interest has to be selected.
3. Extraction: This step is to analyze and highlight the features of the tumor itself, such as size and shape.

A notable difference as an improvement with our work is that the segmentation stage in terms of computational complexity is reduced. Since the input image has an enhancement stage, before passing to the segmentation phase. Therefore, irrelevant information such as noise are eliminated. The enhancement stage provides an image that can be differentiated easier, in comparison to the research [21] in the specific Lung segmentation stage.

- **Tumor Detection and Classification of MRI Brain Image using Different Wavelet Transforms and Support Vector Machines**

  In the research of Gurbinua [22], Gurbinua et al. applied a different wavelet transforms together with a support vector machines, in order to detect and classify MRI brain tumors. Since the classification of MRI images are really important. As a difference in this research the methods were applied to detect disease in the brain, so that demonstrates that the Machine Learning can be performed at several fields in Medicine. Afterwards, the classification of each result, these are interpreted since their classes are in benign or low-grade, from grade 1 to grade 2, and malignant tumors or high-grade, from grade 3 to 4. The proposed methodology by the authors seeks to differentiate between normal brain and tumor brain

Chapter 3

Methodology

3.1 Introduction

To develop this thesis as an exploratory study of lung cancer detection, it begins collecting CT images from a medical database to then process and classify it. Firstly, the CT image has to go through to image processing techniques which, consists of two main steps: Image Pre-Processing and Image Segmentation. Secondly, features are extracted from the already segmented CT images, those features are the perimeter, eccentricity, and area of the segmented lump in the lung. Finally, the features are classified through a SVM-based classifier, which is trained to then build a model to predict the class of a new sample. Therefore, considering the previous explanation, the whole stages of the advancement of this exploratory study are briefly defined in Figure 3.1.

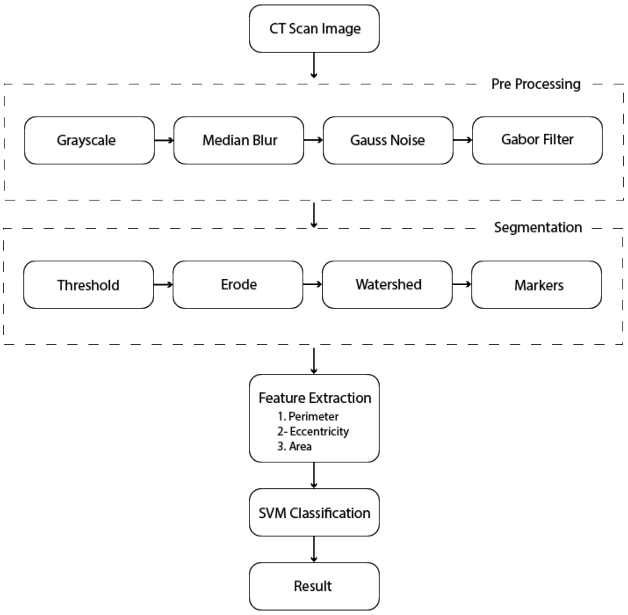

FIGURE 3.1: Flow graph of lung cancer detection development

3.2 Database

The CT images of patients with possible Lung Cancer were provided by *"Give a Scan: The world's first patient-powered open-access database for lung cancer research"*[23]. This database was created at 2013 and consists of a collection of donated metadata of 76 patients. Every patient has several screening studies to follow their diagnostic.

These images have been anonymized and are available for free on the Give A Scan® website.

The Give a Scan imagery set is organized in a conventional format: Patient/Study/Series/Dataset. Inside each data set of each study may include multiple scans of varying quality and anatomical orientation. Therefore the directory *Series* may be divided into multiple datasets. The format of these images is DCM, which is due to the Digital Imaging and Communications in Medicine (DICOM) format, since it has become the default standard in the clinical and research communities for CT data. However, in order to simplify the development, each CT image was transformed to a jpg format with size 512x512, RGB color. The CT images also have the information of the donor such as age, heritage, active smoker, extra diseases, etc. As it can be seen in Fig. 3.2 and Fig. 3.3.

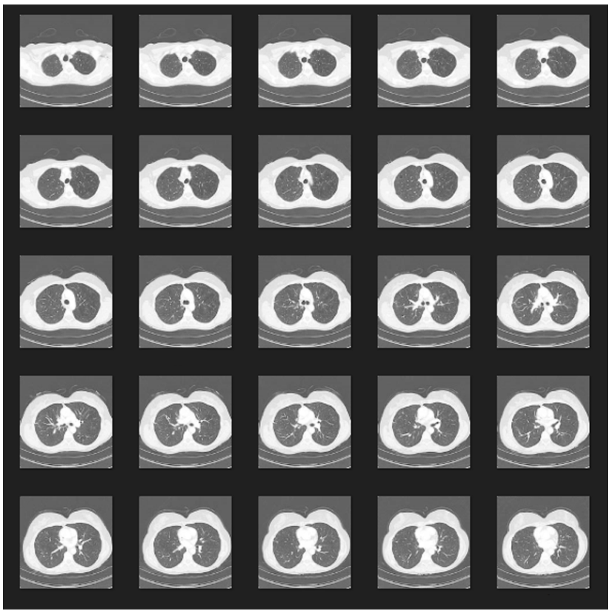

FIGURE 3.2: CT images of a 40 year old female no diagnosed, subject with a family history of lung cancer was CT screened for lung cancer.

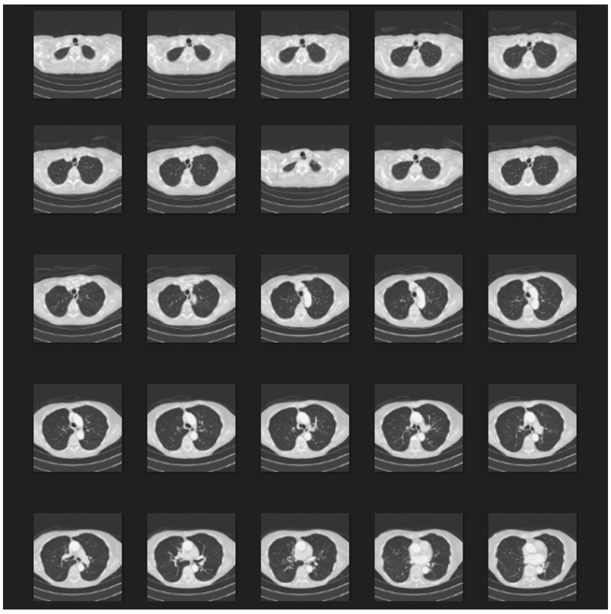

FIGURE 3.3: CT images of a female who smoked, diagnosed with NSCLC, BAC.

3.3 Image Pre-Processing

Raw images, recovered from the data-set, show some noise which could leads to misinterpretation of tumor detection. For this reason, pre-processing the original CT images is fundamental to get accurate results. Image Enhancement is used to improve the quality of the image and to get a better image than the provided one, such that the object of interest, in our case, lumps or nodes in the lungs, are clearly visible.

We use different techniques to enhance the image and to reduce the noising, corruption or interference. Enhancement technique provides better input for the subsequent noise reduction technique. For image enhancement, the image is computed in terms of gray-scale such that the complexity is reduced, then we use different types of filtration methods for the removal of noise such as media filter, Gauss Noise Filter and Gabor Filter.

3.3.1 Gray Scaled

The first step to prepare the image is to convert it to gray-scale images. In the method presented in this thesis, the gray-scale image works better with the techniques applied later, such as the watershed algorithm[24]. Besides, the time used in the pre-processing shall not be longer than the other steps to be developed. Gray scaled images are not like a simple black and white images, they provides a grayscale shade instead of providing only two shades black and white. On images, the grayscale is one in which the amount of each pixel is single sample represents the amount of light it contains or we can say that it carries only the intensity values. Varying from black at the lowest intensity to white at the highest intensity [25]. Grayscale images are distinct from bit by bit on black and white images. The illusion of grayscale shading in a halftone image is obtained by rendering the image as a combination of black dots on a white background or vice versa, and these grayscale images are a result of measuring the intensity of light at each pixel according to a particular weighted combination of frequencies or wavelengths.

3.3.2 Median Blur Filter

The first image pre-processing technique performed to the original gray-scale images is a smoothing technique, also known as blurring, called Median blur. The smoothing is performed in order to reduce salt and pepper noise present in the original medical image. To achieve this result, a filter called the Median filter is applied to the original image [26]. As the name implies, the Median filter replaces the value of a pixel by the median of the neighbor pixel's spectrum level. The median filter is widely used because of its capability to decrease the noise, above other methods.

Besides, the computational complexity of this algorithm is mainly in the calculation of the kernel. Because the filter must calculate pixel by pixel, going through all the neighborhoods [27]. Therefore, its efficiency should be taken into account when implementing it for large images, since the computational complexity is essential due to the time it takes to achieve it.

When the Median filter is linear, the output pixel $g(x, y)$ is computed as the weighted sum of the input pixel $f(i+k, j+l)$, so:

$$g(i, j) = \sum_{k,l} f(i+k, j+l)h(k, l), \tag{3.1}$$

where $h(k, l)$ is known as the kernel. The kernel acts as a window that slides all over the image providing the coefficients to the filter. The filter used in the median blur is the Median Filter. This filter replace each pixel of the input image with the median of the neighbor pixels located within a square around the filtered pixel.

The function used to perform the Median blur is $medianBlur()$. This function is part of the Opencv library for Python[28], and takes in two parameters:

1. Src: Where src is the source of the input image to be smoothed.
2. $ksize$: The size established for this work is $ksize = 5$.

3.3.3 Gaussian Noise Filter

The images obtained after applying the Median Blur Filter are then passed to the Gaussian Noise Filter. Gaussian Noise is also known for having a Gaussian normal distribution.

The Gaussian noise filter is commonly used to remove noise from digital images which is produced due to the acquisition of the image, plus, medical CT images tend to be blurry depending on the acquisition protocol [29].. Some factors could be poor illumination, electronic circuit noise even the room temperature. The noise is removed by smoothing each pixel of the CT image [30]. As the main point, Gaussian filter convulse each pixel with a Gaussian kernel, such that the output pixel is computed as the sum of all pixel convolution [31]s. Then, the transformation applied to each input pixel can be computed as:

$$G(x,y) = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{3.2}$$

where (x,y) are the distances from the origin in the horizontal and vertical axis respectively. And σ is the standard deviation of the Gaussian distribution.

The function used to perform the Gaussian blur is $GaussianBlur()$. This function is part of the Opencv library for Python[28], and takes in three parameters, that were computed through several iterations of study. Besides, in terms of medical processing images the kernel has a prearranged size [10]:

1. src: It is the input image to be smoothed.
2. $ksize$: It is a duple to determine the kernel width and kernel height. The size established for this work is $ksize = (5,5)$.
3. $sigmaX$: It is the kernel standard deviation in X. The size established for this work is $sigmaX = 4.0$.

3.3.4 Gabor Filter

Then, as the last step of image pre-processing, the implementation of the Gabor Filter was computed. The Gabor filter is a technique that consists of a linear filter used for texture analysis and discrimination. Gabor filter determines if there are specific frequency contents in any direction around the region of analysis. When a Gabor filter is applied to an image, the edges, and pixels where texture changes receive the highest response. Gabor Filter results from a Gaussian kernel function, modulated by a sinusoidal plane wave, since previous designs of Gabor Filter had involved the computation of Fourier transformations for textures of interest, and at the same time deciding the discriminating frequencies which fit better to the target image.

Additionally, the Gabor filter used the deductions made from the research of the human visual system and understanding each texture of interest according to the image[32]. However, Gabor filter in terms of resolution in space and spatial frequency is not the best method to apply in order to optimize it, so, it has to be used knowing the process necessary to fulfill the research objective [33]. In this thesis, Gabor filter is used successfully, since in medical images there are a texture or pattern derived from the tomography machine or environment in general. Most commonly, Gabor Filter performance is better in texture characterization. Since the texture is composed of harmonics discrete frequencies, and the use of Fourier express the frequency of the texture in an image [34]. Finally, this filter consists of a real part and an imaginary part described below:

- Real:

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right) cos\left(2\pi\frac{x'}{\lambda}+\psi\right), \tag{3.3}$$

- Imaginary:

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right) sin\left(2\pi\frac{x'}{\lambda}+\psi\right), \tag{3.4}$$

where $x' = xcos\theta + ysin\theta$, $y' = -xsin\theta + ycos\theta$, λ represents the wavelength of the sinusoidal factor, θ represents the orientation of the normal to the parallel stripes of a Gabor function, ψ is the phase offset, σ is the standard deviation of the Gaussian envelope and γ is the spatial aspect ratio.

Therefore, to create a Gabor filter the function *getGaborkernel()* is used. This function is part of the OpenCV library for Python[28]. It requires seven parameters:

1. Ksize: It is a duple of the size of the Gabor kernel.
2. sigma σ: The standard deviation.
3. theta θ: Orientation of the normal to the parallel stripes.
4. lambda λ: Wavelength of the sinusoidal factor.
5. gamma γ: The spatial aspect ratio.
6. psi ψ: It is the offset.
7. Ktype: It indicates the type and range of values that the Gabor kernel can hold.

Until this point, the CT image was already got rid of salt and pepper noise and speckle noise thanks to median and Gaussian filters. Afterward, in order to enhance the image quality, the Gabor filter was performed. After the consecutive application of the three before mentioned filters (Median, Gaussian, and Gabor Filters), the noise of CT images was reduced to a minimum. Then we continued with next phase of the proposed image analysis: Image Segmentation Thus, continue with the processing stage of the CT image, in the next section.

3.4 Image Segmentation

The objective of this is to segment the object of interest of the CT image, The objective is accomplished by using the Threshold method, Morphological Operator - erode, Watershed Segmentation and Markers, as it can be seen in Fig. 3.1. It has to be considered that to get an efficient segmented CT image, the output image should be uniform in color, with no region with holes or empty spaces, since this would mean that there is some loss of information. Additionally, the boundaries of each area/segment of the image have to be simple but detailed. Also, in the next stage, where the feature's information will be collected [35], thus, in this image segmentation stage, the achieving of distinctive features are demonstrated through the output image.

3.4.1 Threshold Segmentation

The first stage of the segmentation section is thresholding, which is intended to binarize the image in such a manner it can be readily analyzed in further steps. Specifically, the Otsu's thresholding method is used, which is a local thresholding approach. Its purpose is to find a global maximum among the gray-level variance values found between the region of interest (ROI) and the background. On doing so, the ROI is completely distinguished from the background. Besides, Otsu's thresholding is commonly used on no-noise images, so that the resulting segmented image relies on how properly the noise removal algorithm is working[36]. Moreover, an advantage of the Otsu's thresholding is that its calculation involves a no large computational complexity [37].

Therefore, to perform this technique, the $threshold()$ function from the OpenCV library for Python is used[28]. This function takes four parameters.

1. The first parameter is the input image. This image must be grayscale.
2. The second parameter is the value used to classify the pixel's value.
3. The third parameter is the value to be given if a pixel is greater/less than the threshold value.
4. The last parameter defines the type of thresholding.

3.4.2 Morphological Operator: Erode

After performing the Otsu's thresholding and obtaining the binary image from it, a morphological transformation is applied to the resulting image. In Morphological operators, there are two fundamental operation, which are erosion and dilation. The one used for this research is Erosion, which in summary is an operation that receives an image as an input so that it reduces the shapes of pixels contained within the image [38]. The objective of this morphological transformation is to erode the boundaries of the foreground object. This goal is achieved by sliding a kernel through the image where a pixel of the original image will be considered 1 only if all the pixels under the kernel is 1 as well. Then, the erosion of a set A by a structuring element (SE) B is defined as:

$$A \ominus B = \{z|(B)_z \subseteq A\}. \tag{3.5}$$

To perform this technique the function $morphologyEx()$ is used. This function belongs to the OpenCV library for Python[28], and takes in four parameters:

1. src: The input image.
2. op: the morphological operation to be performed, in our case is Erode.
3. $kernel$: It is the structuring element.
4. $iterations$: Times that erosion is applied.

3.4.3 Watershed Segmentation and Markers

Finally, the watershed segmentation technique is performed. This technique separates and identifies objects in the image. It extracts seed which points to the existence of an object or background in the image. The philosophy behind this technique comes from geography, it says that any grayscale image can be viewed as a topographic regional surface where high intensity denotes peaks while low intensity denotes valleys. Then, every isolated valley (local min) is filled with different colored water (labels)[39]. In other words, it separates adjacent objects in an image, so is one of the most difficult images processing operations, thus the watershed segmentation is applied to such an issue. Afterward, the watershed segmentation is performed, just remains one stage to complete the image processing techniques[40].

Additionally, OpenCV implemented a marker-based watershed algorithm where all valley points to be or not merged can be specified. Different labels for the object we know are given. Markers algorithm functionality is designed to make the Region of Interest of an image more highlighted. Nowadays, the Markers algorithm is used for augmented reality and robotics applications because they facilitate the localization and landmark detection in featureless environments [41]. Therefore, in this thesis project, it was labeled the region which we are sure of being the foreground or object with one color (or intensity), label the region which we are sure of being background or non-object with another color and finally the region which we are not sure of anything, label it with 0. That is the marker. Then apply a watershed algorithm. Then the marker will be updated with the labels we gave, and the boundaries of objects will have a value of -1. And now it is easy to check if there is a malignant tumor in the lung of the patient.

3.5 Feature Extraction

After the whole section of image processing is performed, the segmented lung nodule is used for feature extraction. After performing noise reduction, the next phase of our work consist on identifying regions of interests within CT images. In our work, a region of interest is a portion of a CT image where a tumor might be depicted. We aimed at measuring the area, perimeter and eccentricity of such regions.

3.5.1 Area

The area is a scalar value which returns an actual number of overall nodule pixel in the extracted Region of Interest. Moreover, the value is the total number of white pixels in the extracted area.

3.5.2 Perimeter

The Perimeter is also a scalar value that returns the actual number of pixels that are in the edges of our Region of Interest. The perimeter can be expressed as follows:

$$Perimeter = (P_{ij}, X_{edge}[P] = i, X_{edge}[P] = j), \tag{3.6}$$

where X_{edge} and Y_{edge} are vectors that represent the coordinate of the i-th and j-th pixel that creates the curve of the region of interest, respectively.

3.5.3 Eccentricity

Eccentricity refers to the roundness or circularity or irregularity complex of the Region of Interest. Eccentricity values are in a range of 0 to 1, such that when eccentricity is 1, it means that the region of interest is perfectly circular, and if it is less than 1, then it can be any other shape. Thus, eccentricity is explained in values of:

$$Eccentricity = \frac{L_{MA}}{L_{mA}}, \tag{3.7}$$

where L_{MA} and L_{mA} are the lenth of major and minor axes, respectively.

3.6 SVM Classification

In this thesis research, support vector machine (SVM) is used to train and predict whether an image is a case of negative or positive lung cancer, the whole process follows the features extracted in the previous section. SVM is a supervised learning model that is used in classification and regression based on characteristics of defined decision boundaries [42], [43].

Therefore, from the 120 images that were used for the SVM, and to differentiate the results. The experiments were classified into three groups as follows:

1. Experiment 1: 80 images used for training and 40 for testing.
2. Experiment 2: 90 images used for training and 30 for testing.
3. Experiment 3: 1000 images used for training and 20 for testing.

The selection of images of diagnosed or not diagnosed images was obtained by random. Besides data points were normalized between 0-1 and labeled positives 1 and negatives 0 for cancerous and noncancerous respectively. The label information was provided by the database[23]. Then, for the SVM, the extracted features were stored in a CSV, so that it can be easily read by the algorithm, the process can be explained in some steps as:

1. First, read the data.
2. Split the dataset into training and test samples.
3. Next, the values of predictors and targets have to be classified.
4. Then, the SVM can be initialized and fitted with the training data
5. So, the classes for the test set are predicted.
6. Attach the predictions to the test set to compare.
7. Compute metrics.

Additionally, a kernel was used such that the linear kernel SVM is also used to classify the image into normal or cancerous images, since in research as [44] the use of kernel is widely explained.

Chapter 4

Experimental Setup

In this chapter, the performance measures and used parameters which were applied for every test are going to be described both by definition and value. All the experiments were implemented, by using several Python libraries, namely `OpenCV`, `Scipy`, `Pandas`, `Numpy`, `CSV`, and `Sklearn`, which are widely used for experimental research and data analysis since their properties and tools are flexible to setup exploratory studies. The principal library used was `OpenCV` [45], which was the main library at the development of enhancement, pre processing, and segmentation the CT image, as it can be seen at Section 3.3 and Section 3.4. Also, libraries `Scipy`, `Pandas`, `Numpy`, `CSV`, and `Sklearn` [46] [47] [48] [49], are used in order to facilitate SVM process since, for example, `Numpy` reduces the management of matrixes; [46] provides a friendly interface to perfom graphics, and `Sklearn` which reduces the complexity of SVM implementation.

4.1 Performance metrics

The following measures were used to qualify the results of the classifications, such as Sensitivity(Se), Specificity(Sp), Accuracy(Ac), Standard deviation(Std), and Mean Error (Me). The corresponding expressions for the measures are:

$$Std = \sqrt{\frac{\sum_{i=1}^{n}(er_i - \bar{er})^2}{n-1}}, \tag{4.1}$$

$$Me = \frac{\sum_{i=1}^{n} er_i}{n}, \tag{4.2}$$

$$Se = \frac{Tp}{Tp + Fn}, \tag{4.3}$$

$$Sp = \frac{Tn}{Tn + Fp}, \tag{4.4}$$

$$Ac = \frac{Tn + Tp}{Tn + Tp + Fp + Fn}, \tag{4.5}$$

according to the following notation:

- er: An error vector.
- n: Length of er.
- TP: True Positives are the cases in which the values predicted and the actual values are both true.
- TN: True negatives are the cases in which the values predicted and the actual values are both negative.
- FP: False positives are the cases in which the actual values is false and the predicted ones differ.
- FN: False negatives are the cases in which the actual values is true and the predicted ones differ.

4.2 Experimental Settings

To determine the correct functioning of the explanatory study, the process was separated as it can be seen at Figure 3.1.

Once the CT image was processed by the image processing techniques, and the features were correctly extracted, these characteristics were the input of the SVM classification. Then, this classification process was carried out by several iterations so that the results are constructed by the computation of the metrics. Table 4.1 summarizes the used techniques and algorithms settings in the stage of image processing, those values were selected by performing several iterations beforehand. however, since these values are not focused in this exploratory study presented, the information of the iterations is not presented.

	Parameter	**Value**
	No. of Images	120
Median Blur Filter	Ksize	5
Gaussian Noise Filter	Ksize Standard deviation	(5,5) 4.0
Gabor Filer	Ksize Sigma Theta Lamda Gamma psi Ktype	61 4.0 from 0 to π by $\frac{\pi}{6}$ 10.0 0.5 0 cv.CV_32F
Thresholding	Threshold value maxVal Type	150 255 Otsy and Binary
Morphological	Type kernel Iterations	Erode (3,3) 1
SVM	kernel training array testing array	rbf length (80, 90, 100) length (40, 30, 10)

TABLE 4.1: Variable setup of the different parameters used for Python implementation.

4.3 Applied Tests on Database

In the proposed methodology the separation in stages is explained. First, every CT image from the database has to be pre-processed and segmented. Therefore the stage of feature extraction is easily developed. Besides, every CT image has the information if it presents cancer or not, so in order to create the supervised classification, that information is added as one of the extracted features.

In order to compare results, there are 3 experiments that change the size number of the training set and test set, the experiments are establish at Section 3.6. However, it is important to emphasize the size of the training vector with which SVM learns since its size would change the results, creating a difference.

For all these experiments, the classification was applied under 20, 30 and 40 iterations. Therefore the obtained results are those obtained with running by the dataset only.

Chapter 5

Results

The results of the process explained in Chapter 3 are presented through figures and tables in order to have a better visualization of the data. The performance metrics explained in Section 4.1 are the main part of the results.

5.1 Experimental setup

Hardware Description

The thesis was executed using the following hardware characteristics:

System Specs	
Name	**Value**
CPU	Intel(R) Core(TM)i7 2.80Ghz
GPU	3GB
Memory	16GB, Type: DDR3, Speed: 1600 MT/s
OS	Windows 10 x64
Kernel	4.18.0-17-generic
Python version	Python 3.7.7

5.2 Results of the Generic Data

In order to understand the image processing, it will be walked through step by step along with a proper explanation.

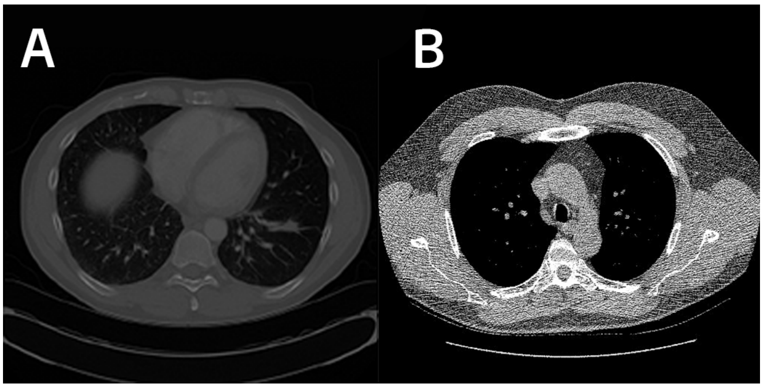

FIGURE 5.1: Side by side in (A) an image of a diagnosed patient, (B) an CT image of a NOT diagnosed patient.

First, some CT images are selected as benchmark in each step, so the understand of the results will be easily portrayed. In Figure 5.1 we have remarkable samples of the data-set with 120 samples, of both a diagnosed and a not diagnosed patient.
Therefore, the use of techniques such as median blur and Gaussian blur for smoothing the image made the first step of processing. In the original images there are parts of terminal bronchi which can confuse the algorithm. In the following Figure 5.2 can be observed the result of image of the Median Filter application.

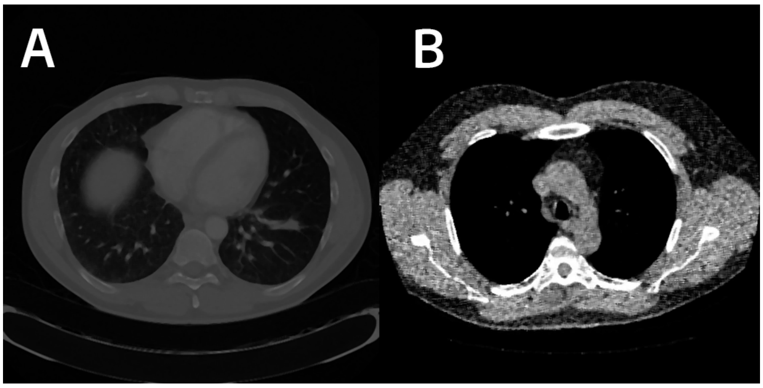

FIGURE 5.2: CT images after the application of median blur filter.

After that, the Gaussian Noise filter was applied in the CT images, such that the presence of noise texture can be completely removed from our image, as it can be seen at Figure 5.3.

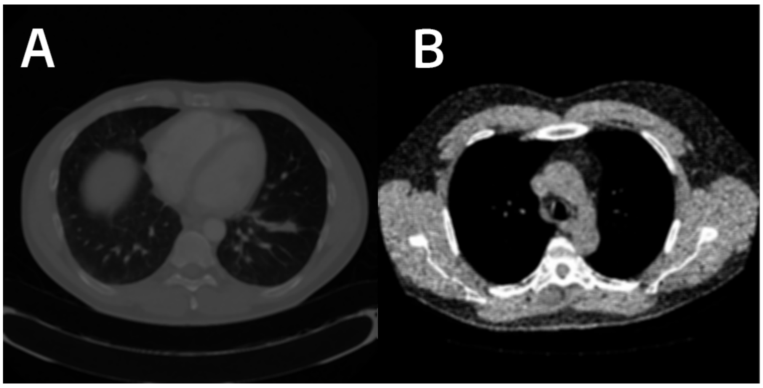

FIGURE 5.3: CT images after the application of Gaussian noise filter.

Considering filtering an image with Gabor functions is related to processes in the visual cortex. Specifically, they are a good model for the receptive fields of simple cells of the cortex if they are supposed to have linear behavior. In Figure 5.4, the tumor-regarded region is defined. This process improves the image for the thresholding step.

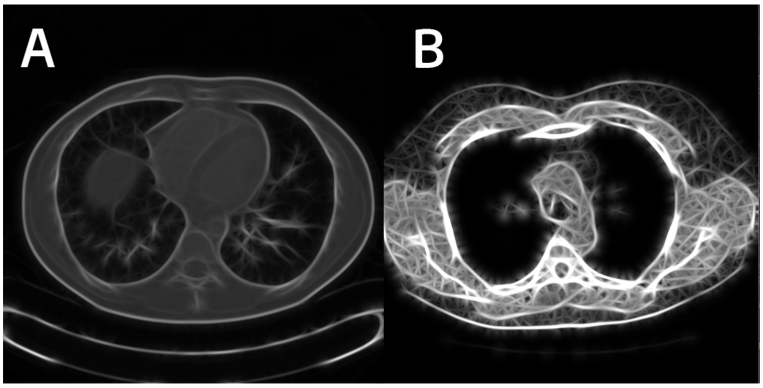

FIGURE 5.4: CT images after the application of Gabor Filter.

Now, the image segmentation process takes places, as follows: First, the thresholding step is applied. After applying the thresholding process to the previous image, the morphological operator: Erode is a crucial step in the segmentation of the image, as can be observed in Figure 5.5,that the presence of the tumor is clearly differentiated in image A, we also see that in Figure B there are barely no nodules inside the lung, which will considerably reduce the processing time involved in feature extraction procedures.

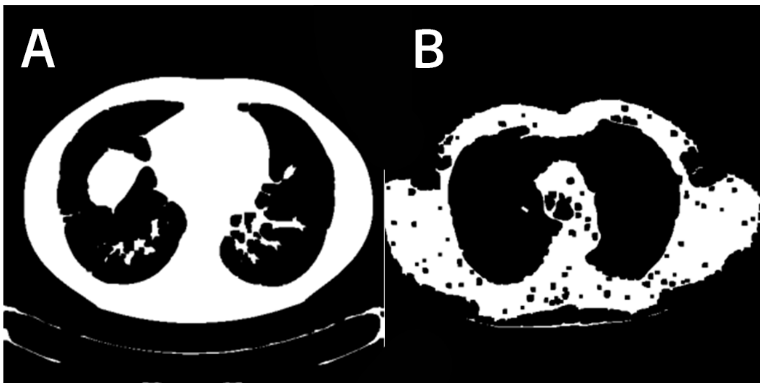

FIGURE 5.5: CT images after the application of morphological operator: Erode.

Afterward, the lasts steps consists in watershed method algorithm and markers which will define with lines the different section segmented in the images. So, in Figure 5.6, the regions of the CT image are segmented by colors, such that the tumor in Figure A, and some errors in Figure B.

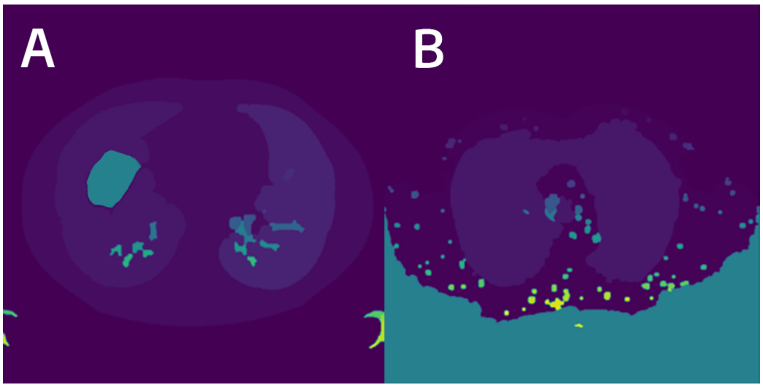

FIGURE 5.6: CT images after watershed segmentation,

Finally, to end the image processing techniques phase, there is markers which facilitate the process of feature extraction since it is easier to follow the edges created by the algorithm, as it can be seen at Figure 5.7.

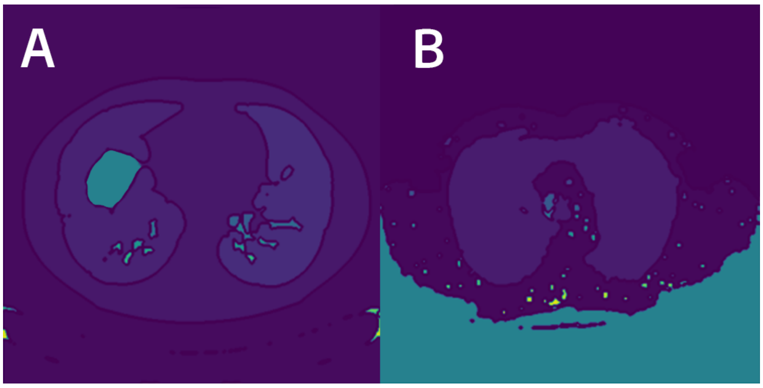

FIGURE 5.7: CT images after the application of the markers algorithm.

Every CT image used were provided by the Database [23]. The processed CT images are from different patients. The results show that when there is evidence of tumor within lungs, the resulting image perfectly segments it.

5.3 Features Extraction

Table 5.1 shows the extracted features of some of the images from the dataset. In order to consider this table as our input to the SVM, the target is added, such that 1 is for CT images of diagnosed patients so $1 = True$, and 0 for CT images of not diagnosed patients so $0 = False$.

Reg.	Extracted Features			
	Perimeter	Eccentricity	Area	Target
1	228	0.8732299393	2086	1
10	278	0.7730792138	1709	1
20	198	0.6773480291	1812	1
30	86	0.8269207862	778	0
40	177	0.71262247839	971	0
50	152	0.6968299712	1535	1
60	188	0.6573850131	1689	1
70	101	0.1553242594	859	0
80	106	0.3747252747	1170	0
90	335	0.8546739984	3197	1
100	164	0.5998460354	1755	1
110	97	0.5001539646	766	0
120	189	0.2423698384	733	0

TABLE 5.1: Obtained results for the extracted features of a subset of 15 CT images.

5.4 Experiments

There are a total of three experiments, every result is performed in the next subsections.

5.4.1 Experiment 1

This experiment is carried out by using a training array with 80 samples that are classified as either positive or negative, and an array of 40 samples to test the SVM according to the metrics in 4.1. Besides, the results are compared in terms of iterations, as 20, 30, and 40 iterations.

Table 5.2 presents the mean and standard deviation from the error obtained by each number of iterations. In this experiment, the lower error was at the execution of 20 iterations. Besides, the standard deviation is low for testing with 30 iterations. However, since the difference is not large enough, the data can be considered accurate.

In Figure 5.8, it can be seen the accuracy of the relation in a box plot which presents that the accuracy has a maximum at 40 iterations. In Figure 5.9 the sensitivity is portrayed, as well as at accuracy the outstanding results are performed at 40 iterations, and its minimum value is of 85%, and, in Figure

5.10, the specificity is displayed, which is the most stable and its average varies from 82% to 85.2%, approximately.

Number of iterations	Mean	Standard deviation
20	0.083750	0.039745
30	0.104999	0.032712
40	0.086250	0.038729

TABLE 5.2: Experiment 1 - Comparison of iterations through mean and standard deviation measures.

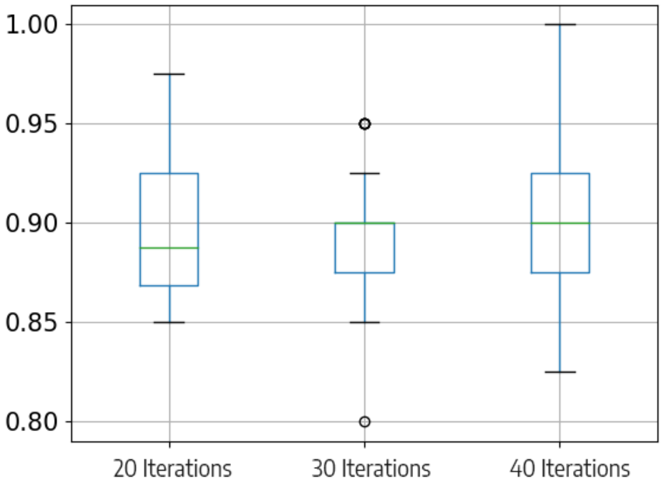

FIGURE 5.8: Experiment 1 - Comparison of iterations through their accuracy.

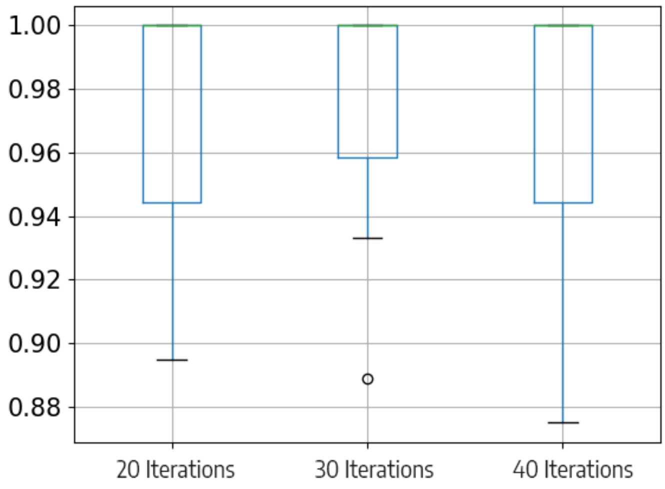

FIGURE 5.9: Experiment 1 - Comparison of iterations through their sensitivity.

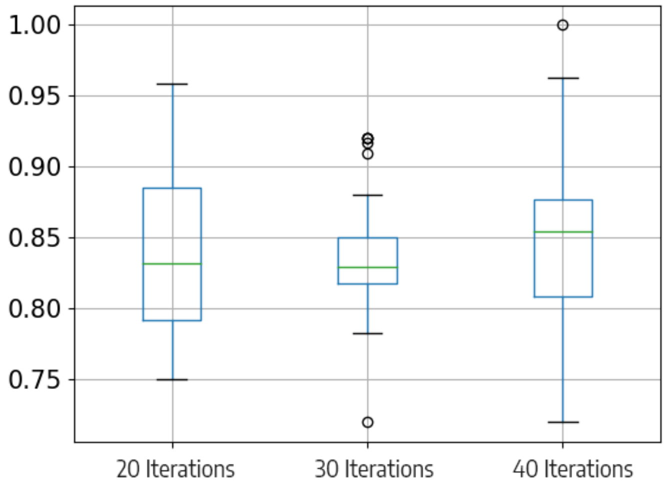

FIGURE 5.10: Experiment 1 - Comparison of iterations through their specificity.

5.4.2 Experiment 2

This experiment is carried out using a training array with 90 samples that are classified as either positive or negative, and an array of 30 samples to test the SVM according to the metrics presented in Section 4.1.

Table presents the mean and standard deviation from the error obtained by each number of iterations. In this experiment, the lower error was at execution of 40 iterations. However the difference between the Mean Error of

30 iterations and 40 iterations is really close. Besides, the standard deviation is significantly low for the testing with 30 iterations, being about 0.02. However, the difference between the 20 iterations and 40 iterations are small too.

In Figure 5.11, it can be seen the accuracy of the relation. According to the box plot, the accuracy exhibits to be stable at 20 and 40 iterations, reaching a median value of 90% of accuracy. In Figure 5.12 the sensitivity is shown. As noted, the results results are relatively high. Its minimum value is 86%, and there is presence of dot values. In Figure 5.13 the specificity, which is more stable and its average goes from 82% to 89% approximately.

Number of iterations	Mean	Standard deviation
20	0.099999	0.050552
30	0.085555	0.037449
40	0.083333	0.051071

TABLE 5.3: Experiment 2 - Comparison of iterations through mean and standard deviation measures.

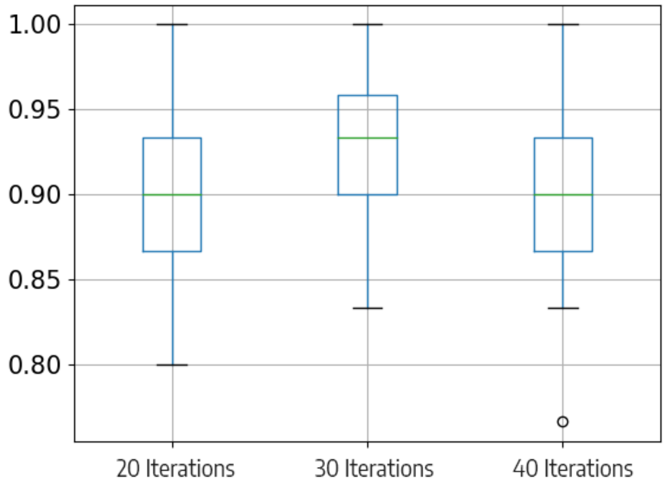

FIGURE 5.11: Experiment 2 - Comparison of iterations through their accuracy.

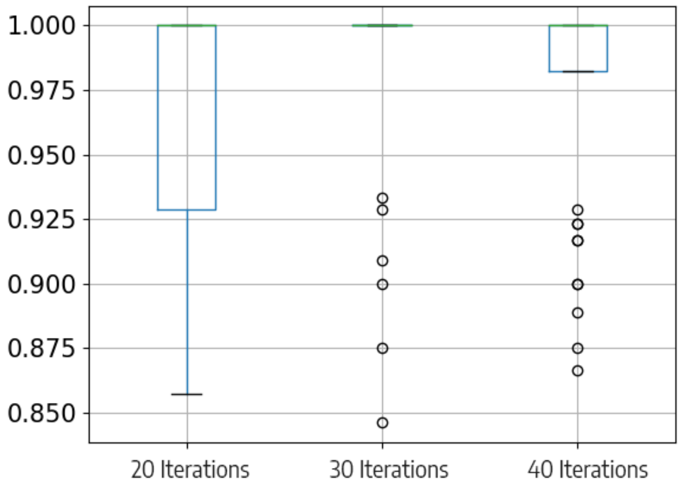

FIGURE 5.12: Experiment 2 - Comparison of iterations through their sensitivity.

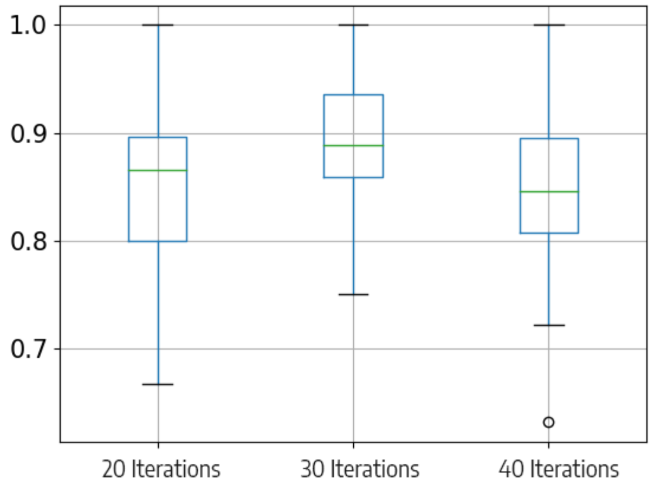

FIGURE 5.13: Experiment 2 - Comparison of iterations through their specificity.

5.4.3 Experiment 3

This experiment is carried out employing a training array with 100 samples that are classified as either positive or negative, and an array of 20 samples to test the SVM according to the metrics described in Section 4.1.

Table presents the mean and standard deviation from the error obtained by each number of iterations. In this experiment, the lowest error was reached at the execution of 20 iterations amounting 0.0775. The difference between the mean error of 30 iterations and 40 iterations is about 0.03. Besides, the standard deviation is lower for the testing with 30 iterations. However, the difference between the three different number of iterations is close from each other. In Figure 5.14, it can be seen the accuracy. The box plot presents that the accuracy is stable for every number. Accuracy is maintained at 90%, but it has a minimum value at 20 iterations of 75%. In Figure 5.15, the sensitivity is displayed, showing that the results are barely perfect, and still there is the presence of dot values. In Figure 5.16, the specificity is shown, which is the most irregular metric -in this case, it ranges from 82% at 40 iterations and at 90% at 30 iterations.

Number of iterations	Mean	Standard deviation
20	0.077500	0.06689
30	0.10500	0.05273
40	0.10500	0.06015

TABLE 5.4: Experiment 3 - Comparison of iterations through mean and standard deviation measures.

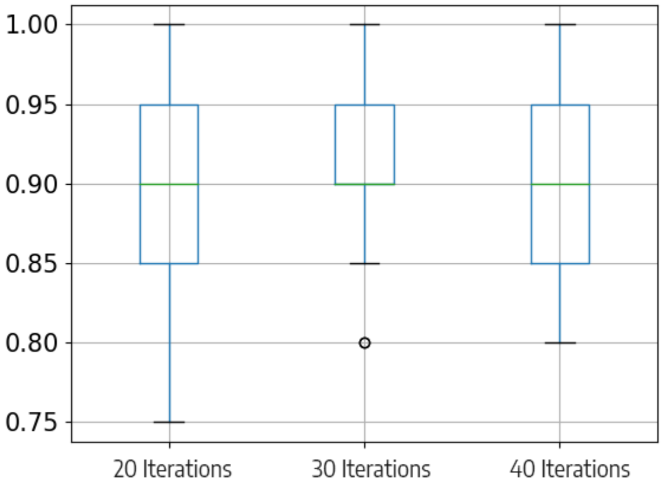

FIGURE 5.14: Experiment 3 - Comparison of iterations through their accuracy.

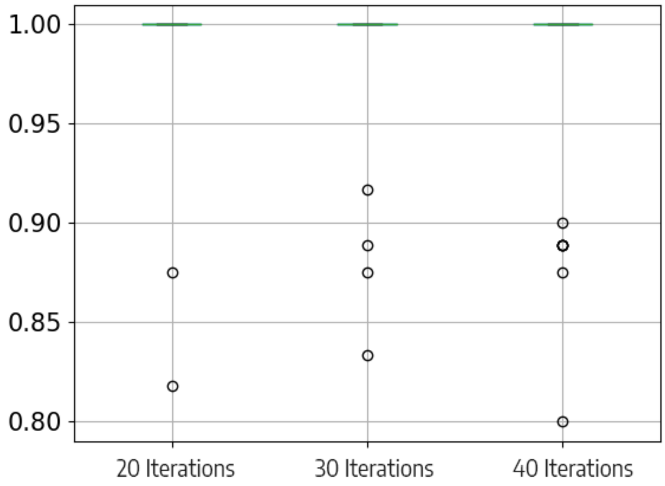

FIGURE 5.15: Experiment 3 - Comparison of iterations through their sensitivity.

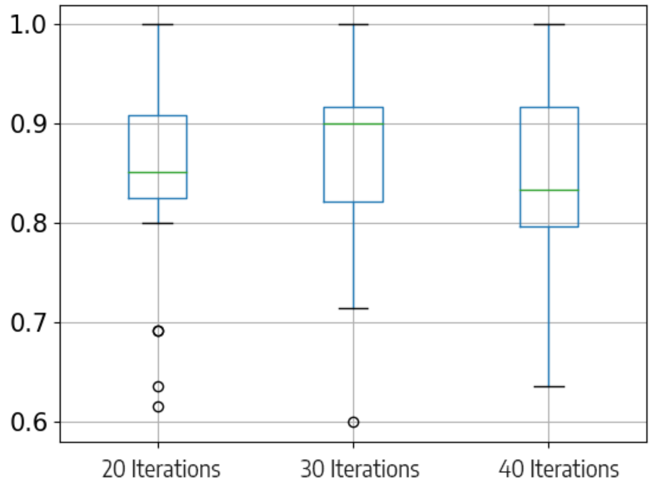

FIGURE 5.16: Experiment 3 - Comparison of iterations through their specificity.

Chapter 6

Final remarks

6.1 Conclusions

In general terms, the objective of this research is to develop an exploratory study of image processing techniques applied to lung cancer detection using benchmark segmentation approaches and support-vector-machine-based classifiers. To do so, supervised learning techniques were studied as well as image processing techniques were developed. The main conclusions reached in this research are the following ones:

- CT images were collected from a database so that they can be tested. Then, every experiment has a high percentage of precision, and, in fact, the experiments work better at 40 iterations to train, in this thesis. So, it can be concluded that the machine needs to train more to learn and synthesize all the information provided by the dataset.

- In conclusion, image processing is a powerful tool for medical settings. The hardware that can be easily acquired today is powerful enough to run algorithms like the one purposed in this work. This leads to the creation of many solutions for many problems that societies have to deal with today. Like, in this case, lung cancer tumors.

- The metrics accuracy, specificity, and sensitivity were selected after a study in what metrics are reliable and suitable for exploratory studies. All of these metrics were implemented in Python given the tools provided by this programming language. Additionally, every metric was employed in every experiment.

- The detection of the presence of Lung Cancer at early stages may save lives [2]. Since the amount of time required to classify the image is considerably minor to the time required to classify them manually.

6.1.1 Future Research

As future work, since there is a lot of medical diseases that have a similar process as lung cancer, this method can be implemented to detection for other diseases that affect high percentages. Furthermore, the research can be implemented with different supervised learning methods such as the k nearest neighbour (k-NN), including more stages and iterations to the process in order to get developed research. The k-NN is considered since it is based in distance that determines the nearest k training instances to target instance, whereas the implemented SVM uses support vectors. Besides, the results and graphics can be upgraded by more specific experiments so that is further studied and improved by subsequent researchers.

Bibliography

[1] P. A. T. E. Board, “Non-small cell lung cancer treatment (pdq®): Patient version”, *PDQ Cancer Information Summaries [Internet]*, 2002.

[2] L. A. Torre, R. L. Siegel, and A. Jemal, “Lung cancer statistics”, in *Lung cancer and personalized medicine*, Springer, 2016, pp. 1–19.

[3] M. A. Grippi, J. A. Elias, J. A. Fishman, A. I. Pack, R. M. Senior, and R. Kotloff, *Fishman's Pulmonary Diseases and Disorders, 2-Volume Set*. McGraw Hill Professional, 2015.

[4] P. M. de Groot, C. C. Wu, B. W. Carter, and R. F. Munden, “The epidemiology of lung cancer”, *Translational lung cancer research*, vol. 7, no. 3, p. 220, 2018.

[5] T. Inage, T. Nakajima, I. Yoshino, and K. Yasufuku, “Early lung cancer detection”, *Clinics in chest medicine*, vol. 39, no. 1, pp. 45–55, 2018.

[6] D. S. Char, N. H. Shah, and D. Magnus, “Implementing machine learning in health care—addressing ethical challenges”, *The New England journal of medicine*, vol. 378, no. 11, p. 981, 2018.

[7] R. M. Hoffman and R. Sanchez, “Lung cancer screening”, *Medical Clinics*, vol. 101, no. 4, pp. 769–785, 2017.

[8] G. Zhang, S. Jiang, Z. Yang, L. Gong, X. Ma, Z. Zhou, C. Bao, and Q. Liu, “Automatic nodule detection for lung cancer in ct images: A review”, *Computers in biology and medicine*, vol. 103, pp. 287–300, 2018.

[9] M. I. Razzak, S. Naz, and A. Zaib, “Deep learning for medical image processing: Overview, challenges and the future”, in *Classification in BioApps*, Springer, 2018, pp. 323–350.

[10] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine”, *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[11] V. A. Gajdhane and L. Deshpande, “Detection of lung cancer stages on ct scan images by using various image processing techniques”, *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 16, no. 5, pp. 28–35, 2014.

[12] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging", *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[13] F. C. Pereira and S. S. Borysov, "Machine learning fundamentals", in *Mobility Patterns, Big Data and Transport Analytics*, Elsevier, 2019, pp. 9–29.

[14] M. Pecht, "Prognostics and health management of electronics", *Encyclopedia of structural health monitoring*, 2009.

[15] A. I. Károly, R. Fullér, and P. Galambos, "Unsupervised clustering for deep learning: A tutorial survey", *Acta Polytechnica Hungarica*, vol. 15, no. 8, pp. 29–53, 2018.

[16] N. Guenther and M. Schonlau, "Support vector machines", *The Stata Journal*, vol. 16, no. 4, pp. 917–937, 2016.

[17] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[18] S Avinash, K Manjunath, and S Senthilkumar, "Analysis and comparison of image enhancement techniques for the prediction of lung cancer", in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, IEEE, 2017, pp. 1535–1539.

[19] S Kalaivani, P. Chatterjee, S. Juyal, and R. Gupta, "Lung cancer detection using digital image processing and artificial neural networks", in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, vol. 2, 2017, pp. 100–103.

[20] B. Tirumalasetty, S. Ambati, and M. A. Shaik, "Brain tumor and lung cancer detection using digital image processing techniques", 2019.

[21] S. Moreno, M. Bonfante, E. Zurek, and H. San Juan, "Study of medical image processing techniques applied to lung cancer", in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, 2019, pp. 1–6.

[22] M. Gurbină, M. Lascu, and D. Lascu, "Tumor detection and classification of mri brain image using different wavelet transforms and support vector machines", in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, 2019, pp. 505–508.

[23] L. C. Alliance, *Give a scan®: The world's first patient-powered open access database for lung cancer research*, urlhttp://www.giveascan.org/, 2013.

[24] J. C. M. Román, H. L. Ayala, and J. L. V. Noguera, “Image color contrast enhancement using multiscale morphology”, *Journal of Computational Interdisciplinary Sciences*, vol. 8, no. 3, 2017.

[25] S. R. Nayak, J. Mishra, and P. M. Jena, “Fractal dimension of grayscale images”, in *Progress in Computing, Analytics and Networking*, Springer, 2018, pp. 225–234.

[26] L. Cadena, A. Zotin, F. Cadena, A. Korneeva, and A. Legalov, “Noise reduction techniques for processing of medical images”, in *Proceedings of the World Congress on Engineering*, vol. 1, 2017, pp. 5–9.

[27] G. George, R. M. Oommen, S. Shelly, S. S. Philipose, and A. M. Varghese, “A survey on various median filtering techniques for removal of impulse noise from digital image”, in *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, IEEE, 2018, pp. 235–238.

[28] Itseez, *Open source computer vision library*, `https://github.com/itseez/opencv`, 2015.

[29] M. Mafi, H. Martin, M. Cabrerizo, J. Andrian, A. Barreto, and M. Adjouadi, “A comprehensive survey on impulse and gaussian denoising filters for digital images”, *Signal Processing*, vol. 157, pp. 236–260, 2019.

[30] N. Kumar and M Nachamai, “Noise removal and filtering techniques used in medical images”, *Orient. J. Comput. Sci. Technol*, vol. 10, no. 1, pp. 103–113, 2017.

[31] J. Nader, Z. A. Alqadi, and B. Zahran, “Analysis of color image filtering methods”, *International Journal of Computer Applications*, vol. 174, no. 8, pp. 12–17, 2017.

[32] D. Dunn and W. E. Higgins, “Optimal gabor filters for texture segmentation”, *IEEE Transactions on image processing*, vol. 4, no. 7, pp. 947–964, 1995.

[33] B. E. Shi, “Gabor-type filtering in space and time with cellular neural networks”, *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 45, no. 2, pp. 121–132, 1998.

[34] C. Palm and T. M. Lehmann, “Classification of color textures by gabor filtering”, *Machine Graphics and Vision*, vol. 11, no. 2/3, pp. 195–220, 2002.

[35] R. M. Haralick and L. G. Shapiro, “Image segmentation techniques”, *Computer vision, graphics, and image processing*, vol. 29, no. 1, pp. 100–132, 1985.

[36] T. Y. Goh, S. N. Basah, H. Yazid, M. J. A. Safar, and F. S. A. Saad, “Performance analysis of image thresholding: Otsu technique”, *Measurement*, vol. 114, pp. 298–307, 2018.

[37] M. H. Merzban and M. Elbayoumi, “Efficient solution of otsu multilevel image thresholding: A comparative study”, *Expert Systems with Applications*, vol. 116, pp. 299–309, 2019.

[38] F. G. De Natale and G. Boato, “Detecting morphological filtering of binary images”, *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1207–1217, 2017.

[39] S Avinash, K Manjunath, and S. S. Kumar, “An improved image processing analysis for the detection of lung cancer using gabor filters and watershed segmentation technique”, in *2016 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, vol. 3, 2016, pp. 1–6.

[40] B. Abdillah, A. Bustamam, and D. Sarwinda, “Image processing based detection of lung cancer on ct scan images”, in *Journal of Physics: Conference Series*, IOP Publishing, vol. 893, 2017, p. 012 063.

[41] J. DeGol, T. Bretl, and D. Hoiem, “Chromatag: A colored marker and fast detection algorithm”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1472–1481.

[42] S. Mustafa, A. B. Dauda, and M. Dauda, “Image processing and svm classification for melanoma detection”, in *2017 international conference on computing networking and informatics (ICCNI)*, IEEE, 2017, pp. 1–5.

[43] M. A. I. Mahmoud and H. Ren, “Forest fire detection and identification using image processing and svm”, *Journal of Information Processing Systems*, vol. 15, no. 1, pp. 159–168, 2019.

[44] D. P. Kaucha, P. Prasad, A. Alsadoon, A Elchouemi, and S. Sreedharan, “Early detection of lung cancer using svm classifier in biomedical image processing”, in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, IEEE, 2017, pp. 3143–3148.

[45] *Opencv-python*. `https://pypi.org/project/opencv-python`, 2020.

[46] *Pandas-python*. `https://pandas.pydata.org/`, 2020.

[47] *Scipy-python*. `https://www.scipy.org/`, 2020.

[48] *Numpy-python*. `https://numpy.org/`, 2020.

[49] *Sklearn-python*. `https://scikit-learn.org/stable/`, 2020.