Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

MALIGNANT TUMOR REGION COMPARISON THROUGH MEDICAL DIAGNOSIS AND STATISTICAL ANALYSIS

Thesis presented to the

Physics College

as partial requeriment to obtain a

Degree of Bachelor of Physics

by

Sofía Pacheco Mex

advised by

Dr. Cristian Heber Zepeda Fernández/Dr. Juan Moisés Arredondo Velázquez

Puebla Pue.
November 6, 2023

Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

MALIGNANT TUMOR REGION COMPARISON THROUGH MEDICAL DIAGNOSIS AND STATISTICAL ANALYSIS

Thesis presented to

Physics College

as a partial requirement to obtain a

Degree of Bachelor of Physics

by

Sofía Pacheco Mex

advised by

Dr. Cristian Heber Zepeda Fernández/Dr. Juan Moisés Arredondo Velázquez

Puebla Pue.
November 6, 2023

Title: MALIGNANT TUMOR REGION COMPARISON THROUGH MEDICAL DIAGNOSIS AND STATISTICAL ANALYSIS
Student: Sofía Pacheco Mex

COMMITTEE

Dr. Eduardo Moreno Barbosa
President

Dr. Benito de Celis Alonso
Secretary

Dr. Lucio Fidel Rebolledo Herrera
Spokesperson

Dr. Cristian Heber Zepeda Fernández/Dr. Juan Moisés Arredondo Velázquez
Advisor

Dedication

To my parents, Gloria and Israel, for motivating me and helping me achieve my goals.
To Ale, Glo and Isra, for being my main support and never letting me give up.

Acknowledgement

I can not express all the gratitude that my parents, Israel and Gloria, deserve since they have always guided my steps, they have supported me in every decision and have given me the foundation to be the person I am. I am extremely grateful to Ale, Glo and Isra for accompanying me on every sleepless night, for every word of encouragement, for trusting me and for wiping away every tear I shed along the way.

Thanks to all the teachers who were part of my academic development. Special thanks to my advisors, Dr. Cristian Heber Zepeda Fernández and Dr. Juan Moisés Arredondo Velázquez, without them this work would not have been possible. Many thanks to the academic group of Medical Physics for all the support, every advice and every teaching they gave me.

Words cannot express my gratitude to Carlos who is a very important part of my life, thanks to him for all the experiences and teachings at his side, for being a source of inspiration, for all the unconditional support and for all the love he has given me.

I am also grateful to Giovana for being a great friend, with whom I shared my entire bachelor degree journey and was a fundamental support for the realization of this thesis. Lastly, I'd like to mention all my closest friends, Salvador, Monserrat, Victor, Iván, Iñaki and Edmundo for all the experiences we lived together and for the support they gave me over 4 and a half years.

Abstract

Breast cancer is a disease that consist of the uncontrolled growth of cells in the breast and it is one of the leading causes of death worldwide. That is why early detection of this disease is important. Mammography is an example of soft tissue radiography, in this image technique are employed low doses of radiation and it is the principal radiographic technique to diagnose breast cancer, however, sometimes this technique present limitations since the breast is composed of different tissue types that limit the possibility of detecting cancerous abnormalities with similar image characteristics as healthy tissue and therefore same grey intensity signal. Image processing by means of computerized methods helps to obtain important information from a mammographic image, like the size, location, form and other characteristics of a tumor or another abnormalities.

This thesis proposes a study that consisted of the analysis of mammographic images with malignant abnormalities. This was accomplished through image processing and data analysis were made with **ROOT** and codes in language `C++`. Mammograms were parameterized and normalized and then the tumor region was located in each image through the analysis of intensity values of pixels. Finally, to evaluate the quality of the results obtained, a comparison was made with the medical diagnostic region, which was found in the data base of mammograms. The results indicated that it was possible to identify the tumor region effectively, since the relative errors between data (medical diagnostic and data analysis) were at most 10%. These results imply this method may be a complement to the medical diagnosis.

Contents

Chapter 1

Introduction

Considering that cancer is one of the leading causes of death, early detection of this disease remains a challenge today. If this disease is detected in early stages it increases the chances that the patient will survive because the effectiveness of treatments and therapies are reduced and also are more invasive when cancer is in advanced stages. In consideration of mammography is a less invasive imaging technique due to the low doses of radiation that are employed, it deserves to continue exploring diagnostic techniques that involve its use. Locate the region of the tumor and determine the size, form and other characteristics of the abnormality is desirable to be able to correctly diagnose the illness and to decide what is the best way to treat it. Due to the above, this study presents a data analysis of the intensity values of the pixels of mammograms with malignant tumors to identify the tumor region to then calculate relative errors between the information found in the database and the results obtained, in order to compare the results to the confirmed diagnoses and determine if this method is an effective way to locate the abnormality area.

1.1 Problem Statement

Breast cancer is the most common type of cancer in the world, mainly affecting women. It is a disease that consists of the accelerated and uncontrolled proliferation of cells in the breast.

Cancer is the leading cause of death in the world [1], being breast cancer one of the most common types of cancer detected. According to the **World Health Organization (WHO)** [2], 85% of the cancer cases are detected in the ducts while the remaining 15% are detected in the lobes [3].

In 2020, 2.3 million women were diagnosed with breast cancer of which around 685,000 died as a result of this disease. Most cases and deaths of breast cancer were recorded in low-income and middle-income countries, being Africa where it is recorded that 50% of breast cancer deaths occur in women under 50 years of age, while in Latin America and the Caribbean the percentage of women diagnosed with this disease before 50 years is 32% and 19% respectively. In this same year, in high-income countries the breast cancer mortality rate had a 40% decrease thanks to the introduction of early detection programs and standardized treatment protocols [3].

In Mexico, during 2021 according to the **Instituto Nacional de Estadística y Geografía (INEGI)**[4] 7,973 deaths were documented due to breast cancer of which 99.4% were women and 0.6% were men; 17% of the deaths due to this disease were due to malignant tumors. The highest rate of deaths from breast cancer was recorded in women aged 60 years and more while the lowest rate was documented in women between 20 and 29 years.

Although there are many techniques to detect and treat breast cancer, many times the existing techniques have limitations, so medical physics focuses on improving them or looking for other methods to make them more efficient or accurate, whether for detection of the disease in earlier stages and get a better diagnosis or for enhance the therapies.

Delimiting a region of interest **ROI** which in this case may be a tumor or another abnormality that may turn into cancer is an important duty for many reasons like the evaluation, treatment and therapies which are mentioned below:

- **Prognosis:** The ROI size is related to the severity of the disease. Frequently a bigger tumor indicates an advanced cancer and may have a less favorable prognostic.
- **Treatment:** Depending on the tumor or abnormality size, treatment options may be determined. Commonly a small size may be candidate to a surgery while tumors with bigger size may required radiotherapy or chemotherapy.
- **Monitoring and evaluation:** Known the initial tumor size is primordial to evaluate the response to the treatment. Shrinking the tumor over time is an indicator that treatment is working.
- **Propagation prediction:** The abnormality size can also help predict the probability that the cancer has spread to other organs.
- **Clinical research:** In clinical research, the tumor size is an important variable to understand the efficacy of new treatments and experimental therapies.

1.2 State of the art

Characterize the size, shape and patterns, among other things, of a tumor or another abnormality present in the breast that may cause cancer, is an important duty since it is fundamental for the diagnosis and treatment of this disease. In recent years, there have been many related research works on this.

Guevara et al. in the paper *"Detection of Breast Cancer using Convolutional Neural Networks with Learning Transfer Mechanisms"* [5], used deep learning models to discover the features in 7803 images with benign and malignant abnormalities to identify and classify breast cancer, they employed 4 Convolutional Neural Networks (CNN) to do it. Likewise, Arevalo et al. described a learning framework to diagnose breast cancer in a mammography that integrates deep-learning techniques. It was built a biopsy proven benchmarking from 736 film mammography with the aim of improving image details, this can be found in the article called *"Representation learning for mammography mass lesion classification with convolutional neural networks"* [6].

"An Improved Fully Automated Breast Cancer Detection and Classification System" [7] is about a proposed computerized method to detect and categorize tumor masses in the breast, Shawly, T. and Alsheikhy, A. A. used two deep-learning models and a classifier to develop the method which was evaluated in 5 data-sets and turns out that the method has the ability to classify the cancer as benign or malignant and also can categorize the healthy tissue. On the other hand, Kuo et al. in *"Complete, Fully Automatic Detection and Classification of Bening and Malignant Breast Tumors Based on CT Images Using Artificial Intelligent and Image Processing"* [8] used 174 breast tumors and by means of image processing it was developed an automatic detection and diagnostic system that classify the breast tumors of the CT scan images, different

methods were used to detect, locate and circle the tumor to then classify it as benign or malignant.

Additionally, Strelcenia, E. and Prakoonwit, S. in the paper *"Improving Cancer Detection Classification Perfomance Using GANs in Breast Cancer Data"* [9] proposed a K-CGAN method trained in different settings to generate synthetic data to compared it with a Breast Cancer data set. This study applied five methods of classification and feature selection to "non-image" sample consisting of 357 malignant cases and 212 benign cases for evaluation, this study had the objective of develop a highly efficient classification framework. Furthermore, in the paper *"Breast ultrasound image classification and physiological assessment based on GoogLeNet"* [10], Chen et al. used 880 breast ultrasound images where 103 was normal images, 467 had malignant tumors and 210 with benign tumors and by means of a CNN model of GoogLeNet the images were processed and with this method it was possible to detect and classified the breast cancer.

Similar studies exist to the one in this thesis, an example is a bachelor thesis under the name *"Separación de Regiones de Interés para el Diagnóstico y Tejido de fondo en Mamografías mediante Análisis de Datos"* [11] where López proposed a data analysis method to isolate the regions of interest from the rest of the breast by means of an analysis of the intensities of the pixels that make up the image. It is essential refer to a study named *"Tumor and microcalcification characterization using Entropy, Fractal Dimension and intensity values statistical analysis in a mammography"* [12] by Zepeda, C. et al. since in this study was used a data analysis method to segment and distinguish malignant tumors from the rest of the breast by means of fractal dimension, entropy and pixels intensity, the results where that the highest intensity pixel value was located in the centre of the **ROI** and suggested that the pixels of the **ROI** had intensity values up to 0.7.

1.3 Hypothesis

Starting from data analysis of the intensity values of the mammograms with malignant abnormalities, it is possible to locate the area of the abnormality.

1.4 Objectives

1.4.1 General

To comparing a Region of Interest **(ROI)** between medical diagnosis and a data analysis method applied to mammograms with malignant tumors.

1.4.2 Particulars

- To parameterize and normalize mammograms.
- To create computational codes in `C++` to perform the analysis.
- To analyze the mammograms with malignant abnormalities.
- To get the region of the malignant abnormalities present in the mammograms.
- To evaluate the quality of the results obtained by means of the comparison of the area obtained and the medical data.

Chapter 2

Theoretical Framework

This chapter contains information that is necessary for the understanding of this study. It consist of a general description of the breast, the tissue types, the breast cancer and some abnormalities like masses and calcifications. Furthermore, discusses mammography, digital image processing and finally it is presented the mathematical concepts that were used for data analysis and the reliability method.

2.1 Breast Cancer

The female breast is made up of 10 or 20 sections called **lobes**, in turn, these lobes are divided into smaller sections called **lobules**. The glands responsible for milk production are contained in the lobules and the milk goes from the lobules to the nipple through tubes called **ducts**. The space between lobules and ducts it is composed of fat and fibrous tissue. [13] This can be seen in the figure 2.1.

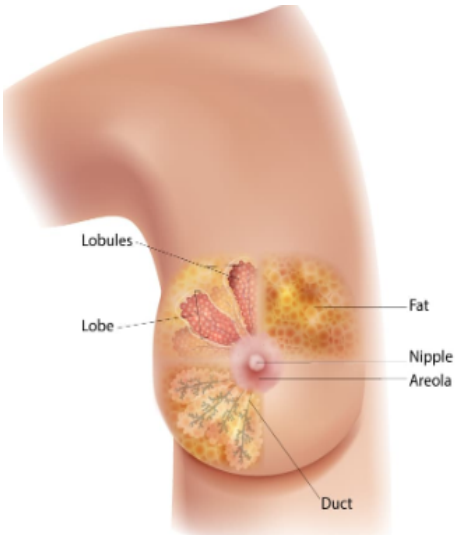

Figure 2.1: Anatomy of the female breast. Taken from: *What is Breast Cancer?. Centers for Disease Control and Prevention, 2022 [14]*

Breast cancer is a disease that consists of the accelerated and uncontrolled proliferation of cells in the breast [13]. Most breast cancers begin in the ducts or lobules but can begin in other

parts. Also, the cancer can spread outside the breast and when this happen it is said to have metastasized. There are different kinds of breast cancer, and depends on the type of cell that turns into it, the most common kinds are the **invasive ductal carcinoma** which is the type that begins in the ducts and the **invasive lobular carcinoma** which begins in the lobules [14, 15].

There are risk factors for this disease, and some of them can be:

- **Age:** the risk of breast cancer increases with increasing age.
- **Personal history:** people who have previously had breast cancer are more likely to develop it again.
- **Genetic predisposition:** family history or mutations in certain genes.
- **Hyperplasia:** increased number of cells in an organ or tissue that can become cancerous.
- **Children:** not having children or even having children at a late age.
- **Menstruation:** starting menstruation at an early age.
- **Density of the breast.**
- **Obesity.**

Breast Cancer Stages

Staging is the process by which it is determined if the cancer has spread in the body and if so, how far. This helps to defined how serious cancer is and what is the best way to treat it. Staging is divided in **stage 0** that is the earliest stage breast cancer which is called *carcinoma in situ*, then ranges from **stage I** to **stage IV**. A higher number means the cancer has spread out more of breast.

TNM system [16] (which means Tumor, Node and Metastases) is the most common staging system used for breast cancer, it is an international standard for classifying the malignancy of an abnormality. For breast cancer the TNM system has two stages:

- **Pathological stage:** known also as **surgical stage**, is determined by examining tissue removed during a surgery.
- **Clinical stage:** if is not possible to do a surgery, then the cancer will be given a clinical stage which is used to help plan treatment. This stage is based on the results of other test like biopsy, imaging test or physical exam.

In both stages mentioned there are 7 key characteristics that are used: the size of the tumor, the spread to nearby lymph nodes, metastasis to distant sites, estrogen receptor status, progesterone receptor status, HER2 status (is a protein produced by cancer) and grade of the cancer [17].

2.1.1 Tissue types

The density of the breast is a measure of how much fibrous tissue and glandular tissue are in the breast in comparison of fatty tissue. This can be determined by a mammogram and it is important to know it because women with dense breast tissue have an increased risk of breast cancer [18].

There are three tissue types:

- **Fibrous tissue**: it is the tissue that holds the breasts in place.
- **Glandular tissue**: it correspond to the part of the breast that contains the lobes and ducts.

- **Fatty tissue**: this tissue fills the space between fibrous tissue, lobes and ducts. It is also the one that gives the breasts their size and shape.

The breasts are considered dense when they are mainly composed by fibrous and glandular tissue and do not have too much fatty tissue [19].

Breast density may be assigned with any of the following categories:

- **Low density**: occurs when breasts are composed almost entirely of fatty tissue or have little areas of dense tissue scattered throughout the breasts.

- **High density**: occurs when breasts are composed almost entirely of dense tissue.

The figure 2.2 shows mammograms of different breast densities. Starting from left to right, the first is a breast composed of fat, the second one shows little areas of scattered fribroglandular tissue, these two are in the category of low density, while the last two are in the high density category as they show a breast composed almost entirely of fibroglandular tissue.

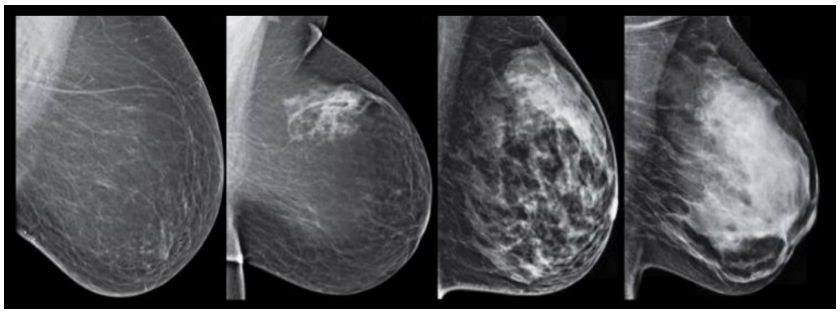

Figure 2.2: Breast density. Taken from: *What Is Density in the Breast?. Connecticut Breast Imaging, 2022*

2.1.2 Abnormalities

Tumors

Tumors are lumps, in other words, they are abnormal tissue masses due to the uncontrolled growth of cells. Not every tumor is sign of cancer, that is why they can be classified as benign (non-cancerous) or malignant (cancerous). The difference between the two types of tumor are that benign tumor do not spread outside the part of the body in which they grew, while the malignant one does and metastasis is said to have occurred [20, 21].

Masses

A mass is a breast lesion that can be seen in a mammogram as an area of abnormal breast tissue. It can have a different shape and edges compared to the rest of the breast tissue. A mass can be benign or malignant and sometimes may be accompanied by calcifications. A cyst is a type of mass and it is a small sac that may contain fluid, air or solid material. Another type of mass may be the solid masses that can be more concerning than cysts, but frequently solid breast masses are non-cancerous [22]. The shape of a mass can be round, oval, lobular, or irregular, on the other hand, the margins can be described as circumscribed, microlobulated, obscured, ill-defined or spiculated [23].

Calcifications

Calcifications are small calcium deposits (Understanding Breast Calcifications, n.d.) [24] on the soft tissue background of the breast. According to American Cancer Society [22], in a mammogram, calcifications look like white dots and they may or may not develop into cancer. There are two types of calcifications:

- **Macrocalcifications:** are big calcium deposits most likely due to old injuries, breast inflammation or arteries aging. These type of calcification are commonly related to non-cancerous conditions.
- **Microcalcifications:** They are tiny spots of calcium in the breast and are more delicate than macrocalcifications although it does not always mean that cancer is present. It can determine that the calcification is due to cancer or not by their shape, layout and whether they are near a mass. Sometimes a biopsy is necessary to check for cancer.

2.1.3 Mammography

Mammography is an example of soft tissue radiography, it is the method of diagnostic imaging that is obtained by the interaction of the x-rays with the breast, in this image technique low doses of radiation are employed and it is the principal imaging technique to diagnose breast cancer. However, this method of diagnosis is not infallible, because sometimes its sensitivity is limited by the fact that the tissues it studies present similar radiological features [18].

The mammogram machine is called a *"mammograph"* and it is composed by 2 clear plastic plates that compress the breast to spread out the tissue apart, it takes two X-rays pictures (one from the top and the other one from the side) and then it is the same for the other breast [25]. There are three types of target materials: molybdenum, rhodium and tungsten or even a combination of two of the materials mentioned before [26]. There are certain standards to do mammograms related to radiation dose, personnel, equipment and image quality.

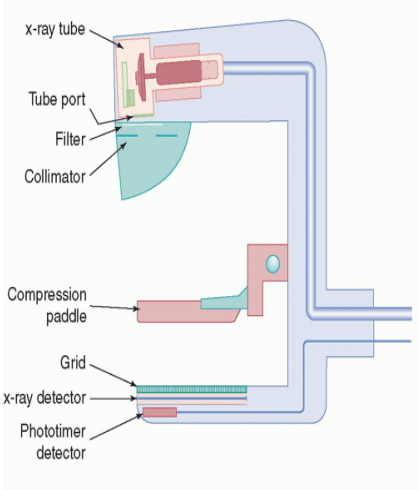

Figure 2.3: Mammograph components. Taken from: *Themes (2021).*

The generation of X-rays is an important concept to obtain a mammogram. The process by which kinetic energy is converted into electromagnetic energy due to interaction with the target's nuclear field it is called **Bremsstrahlung X-rays**. When the electron interacts with the

target, the electron slows down and changes its trajectory causing the loss of kinetic energy to reappear in the form of X-rays. Characterisctics X-rays are generated when a projectile electron interacts with electrons in the inner shell of the target in the X-ray tube, causing the ejection of the electron from the target atom leaving a vacancy in its energy level, so the electrons in the upper shells cascade down to take the place of the ejected electron. Characteristics X-rays have specific discrete energies determined by atomic structure. Whereas Bremsstrahlung X-rays have a continuous spectrum of energy ranging from zero to the maximum possible energy in the X-ray tube [27].

Two physical effects are involved in obtaining a mammogram, by the interaction of X-rays with breast tissues. The **Photoelectric Effect** occurs when a photon in the X-ray beam interacts with an electron in the innermost shell of an atom, in this case the X-ray is completely absorbed and this generates important diagnostic information as it produces the clear or brightest areas on the mammogram. On the other hand, the **Compton Effect** occurs when a photon in the X-ray beam interacts with an electron in the outermost shell of the atom, causing the atom to ionize and the beam to scatter and lose energy; scattered X-rays produce a decrease in the contrast of the image and can produce noise [27]. Also, X-rays, which do not interact with the body at all, are also important because they are responsible for generating the dark areas on the mammogram [27].

Clinically, mammography can be classified into two types:

- **Screening Mammography:** it is used in women with no symptoms (asymptomatic). Consist in two projections to detect an unexpectedly cancer. X-ray pictures of each breast are taken usually from 2 different angles.
- **Diagnostic Mammography:** it is used in patients with severe symptoms or elevated risk factors. It may be necessary to make two or three projections of each image. It can be also used to check for breast cancer after a lump, a sign or a symptom of disease has been found. Since this procedure takes multiple images of the breast (more than the other type), the total dose of radiation is higher than in a screening mammogram due to the fact that time to expose is longer in this exam.

Mammograms can show different types of abnormalities, where the main types of breast lesions found are calcifications, masses, asymmetries and distortions. A mammogram usually is not enough to decide if the abnormality is or is not a cancer sign, to reach this diagnostic it is necessary to do more tests [25].

Magnifications techniques are frequently used in mammograms which can produce images up to twice the normal size. Usually this techniques are unnecessary for most of the patients because with a normal one it is possible to detect abnormalities. An amplification allows better observations of some really small lesions or microcalcifications since this technique helps in defining the borders but to achieve this the radiation dose to the breast increases [26].

2.2 Image Processing

A digital image may be defined as a two-dimensional function where x and y are spatial coordinates and the amplitude at any pair of coordinates is called the "intensity" or "gray level" of the image at that point (Gonzalez and Woods, 2002, p.1) [28]. Digital Image Processing is the set of techniques and algorithms to manipulate and analyze digital images to obtain important information of the image, this techniques included tasks like improve the quality, color correction, edges detection, noise removal, objects segmentation and patterns recognizing. Generally, it is difficult to say where image processing stops since it is highly related to image analysis and computer

vision, but some authors defined image processing as "a discipline in which both the input and output of a process are images" (Gonzales and Woods, 2002, p.2), however, it is possible to define three types:

- Low-level: inputs and outputs are images.
- Mid-level: involves segmentation, description of those objects and classification, inputs are images but outputs are characteristics (edges, contours, etc.) from the images.
- High-level: involves "making sense" of a group of recognized objects, normally associated with computer vision.

2.3 Mathematical Concepts

Standard Deviation

The standard deviation S is a measure of dispersion that is understood as the estimate of the average uncertainty of the measurements, this means, it is a measure of how disperse the data is in relation to the mean. It can be calculatated by the following equation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}} \tag{2.1}$$

If the mean and median are similar and the standard deviation is maximum 30% of the value of the mean, it can be supposed that the data distribution is **normal or Gaussian**.

Relative Error

Relative Error is used to quantify the accuracy of a calculation or measurement in relation to the actual or theoretical value. It is defined as the ratio of absolute error to actual value. The closer to zero the relative error, the more precise the result is considered and is usually expressed as a percentage or in a decimal form. The equation to calculate the error is the following.

$$\delta = |\frac{\nu_A - \nu_E}{\nu_E}| \times 100 \tag{2.2}$$

Where ν_A is the measured value and ν_E is the actual value.

Histogram

A histogram is a type of chart in which are presented the frequency distribution of numerical data. It is similar to a bar chart but the difference is that in a histogram the bars are consecutive and class intervals in a histogram are all of a equal length due to the class frequencies are plotted proportionally to the bars heights. With this chart it is easy to compare the differences between bars frequencies with the same base, it is also easy to identify the data variability and the center of each class intervals.

The histograms can be classified by the way in which they are distributed:

- **Symmetrical**: This type of distribution it seems like a bell, this means that the left half is a reflected image of the right half.

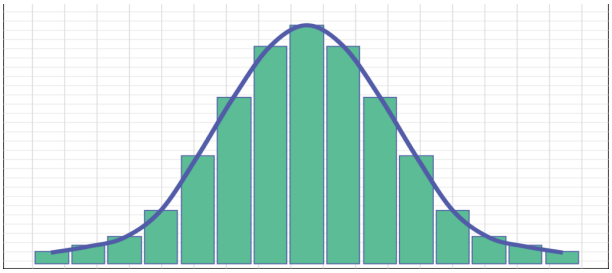

Figure 2.4: Symmetrical histogram. Taken from: *"Las Entradas y Salidas del Histograma", Sabbah, Z. (2023).*

- **Skewed**: A **Right-Skewed Histogram** or Positively Skewed Histogram occurs when the data distribution indicates that high values occur infrequently. "The tail" is said to be longer on the right side. Also, on a right-skewed histogram, the mean, median, and mode are all different and the peak of the distribution are on the left side. On the other hand, a **Left-Skewed Histogram** or Negatively Skewed Histogram it is when the distribution has a "tail" on the left side. In this case, the mean is lower than the median and the peak of the distribution are on the right side.

Figure 2.5: Right-Skewed Histogram and a Left-Skewed Histogram. Taken from: *Torturando Los Datos: Capítulo I. Blog. Merkle.*

- **Multimodal**: It is when the distributions has more than one peak. When it has two peaks it is called *bimodal*, if it has three peaks is *trimodal*.

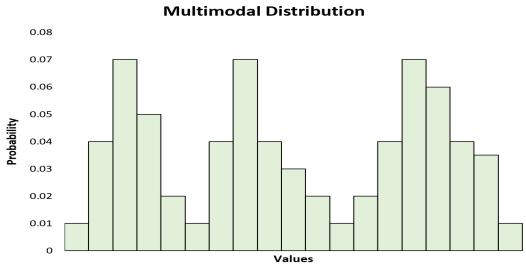

Figure 2.6: Multimodal distribution histogram. Taken from: *Zach. (2021). What is a Multimodal Distribution? Statology.*

Chapter 3

Methodology

The methodology followed to obtain the tumor area by means of the data analysis is shown in this chapter; includes the criteria under which downloaded mammographic images (MI) were selected and the data analysis which consisted of the parametrization and normalization of the images and the realization of histograms of image data.

3.1 Data base of mammograms

A sample of 45 Mammographic Images (MI) was taken from *"The mini-MIAS database of mammograms"* [29] on PEIPA's page. There are 322 mammograms in the data base, and each image have a reference number and information about them. The MI are gray-scale images in .pgm format, and they have a size of 1024×1024 pixels.

The details of every image can be consulted in the data list found on the website, this table consist of 7 columns as seen below:

- **1st column**
  Reference number of each MI
- **2nd column**
  Character of background tissue
- **3rd column**
  Class of abnormality present
- **4th column**
  Severity of abnormality
- **5th column**
  x image-coordinate of centre of abnormality
- **6th column**
  y image-coordinate of centre of abnormality
- **7th column**
  Approximate radius (in pixels) of a circle enclosing the abnormality

There are some important facts about that list: the list is arranged in pairs of films, where each pair represents both mammary glands of a single patient, the origin of the coordinate system of the images is fixed in the bottom-left corner and in some cases calcifications are

not concentrated at a single site so for these cases centre locations and radii have been omitted [29].

The figure 3.1 below is an example of mammogram that can be found in the PEIPA's database and then the details of the image are included as an example that illustrates the seven columns of information in it.

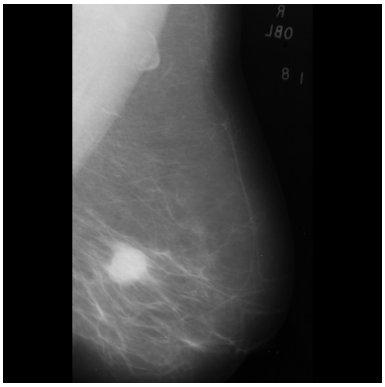

Figure 3.1: mammographic image under the name "mdb028". Taken from: *The mini-MIAS Database of Mammograms.*

mdb028 F CIRC M 338 314 56

The seven columns above are the information about the mammographic image 3.1, it starts with the name of the file which in this case is "mdb028", continuing with the background tissue type that is Fatty (F), then there is the abnormality type CIRC which means a well-defined/circumscribed mass, the fourth column indicates the severity of this abnormality that is malignant, the number 338 in the fifth column is the x-coordinate of the centre of the abnormality while the sixth column is the y-coordinate of the centre of the abnormality which in this case is 314 and finally, the last column is the approximated radius in pixels of the abnormality which is 56.

3.2 Sample taken

For this project, 45 MI with malignant abnormalities were selected regardless of tissue and abnormality type, however, calcifications and those that did not have the centre location and radii were discarded. The images were downloaded from the PEIPA's imagery [29] where are 322 mammograms images in .pgm format which had to be changed to .png format to made modifications on each mammography. These modifications consisted of removing artifacts from the MI, i.e, text boxes, numbers or part of the chest, since those artifacts do not represented important information for this study. Artifacts areas were filled with black color in *kryta*, which is a program that allows to make changes in any image. The images were not cut because this option makes them smaller and for this work the MI size always had to be 1024×1024 pixels.

Figure 3.2 shows a MI downloaded from the imagery, the reference number of this film is "mdb072" and this is the original one, meaning that the image had not suffer any changes yet, also a part of the chest can be appreciated.

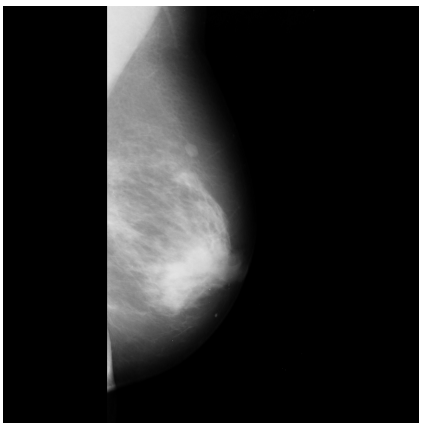

Figure 3.2: Mamographic image obtained from PEIPA's page.

3.3 MI parametrization and normalization

In order to be able to do the parametrization of the images, a code in ROOT through `C++` language was developed, using a class reference of ROOT called **TASImage** [30], which is the interface that helps image processing allowing them to be read and written in different formats. It also allows manipulation of the images (scaling, tiling, merging, etc.). ROOT is a framework used for data processing developed by **the European Organization for Nuclear Research (CERN)**. It serves to save data, provides mathematical and statistical tools, it can displayed the results as histograms, scatter, plots or any type of chart [31].

This program takes as input the images obtained from the database in .png format. From this program, it could be obtained a new image, which it is called **parameterized image (PI)**, as it is shown in figure 3.3. It is called a parameterized image, because each pixel has three coordinates: x, y, and the intensity value z. For ease of analysis, these three coordinates were normalized and are unit-less, this information was recorded in a data file. In this transformation, the 1024×1024 pixels still maintained.

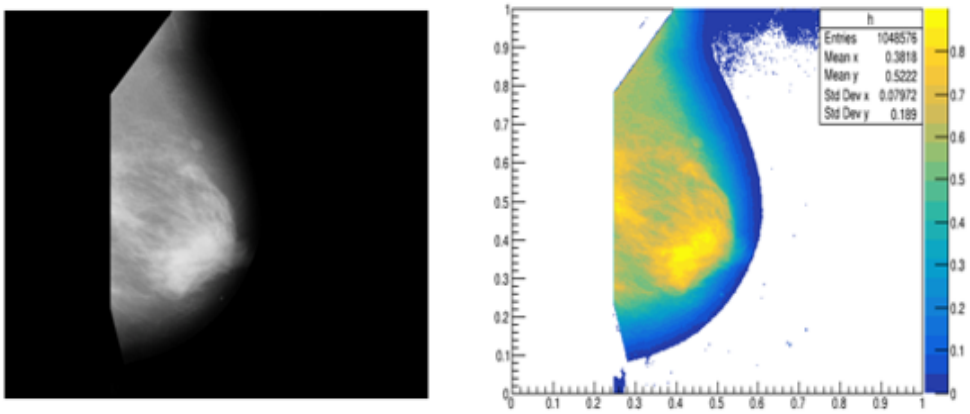

Figure 3.3: Left: Original image from the data base and right: Parameterized Image (with arbitrary units a.u.).

Afterwards the respective files of each MI were obtained and saved, it was verified that the x and y coordinate returned by ROOT on the brightest pixel coincided with a point inside the mammary gland.

Since this study looks for pixels with intensity values greater than 0.7, since it is the value that Zepeda et. al suggested [12], a code was created to calculate an area that included only pixels with the intensity mentioned before. It was sought that this region was proportional to the radius of the abnormality of each MI provided on the PEIPA page; considering that the image scale was readjusted, the diagnostic radius were divided by 1024 in order to obtain a region proportional to the original. Also with this code it was possible to obtain a histogram for each PI in which they are shown two lines, the first one (blue line) corresponded to the medical diagnosis and the second one (red line) was the analysis data from ROOT, both of them correspond to the respective abnormality area and this can be appreciated in figure 3.4.

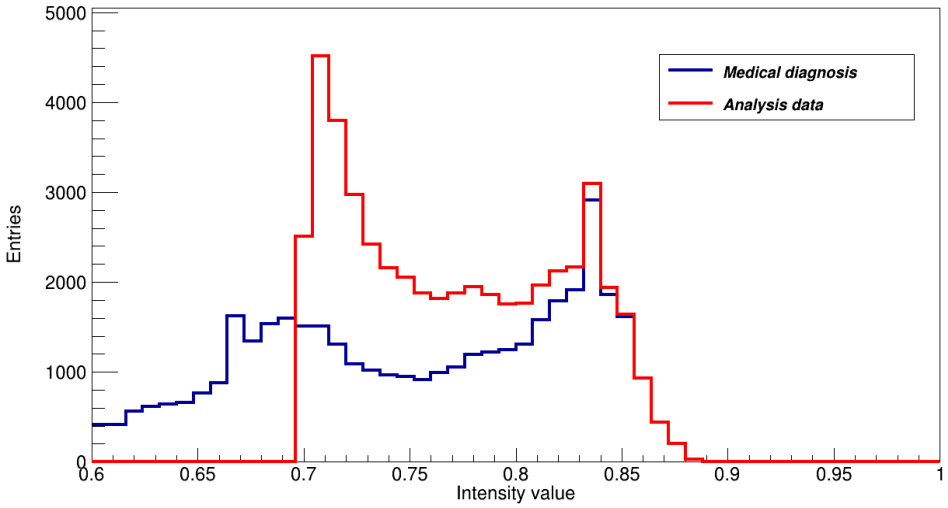

Figure 3.4: 2D areas histogram. The histogram shows the total number of pixels that make up each regions. This is an example of an MI, 45 histograms similar to this figure were obtained.

Chapter 4

Results and Discussions

In this chapter a statistical method that consists of the calculation of relative errors to compare how similar the results obtained are with the data found in the database is presented. This with the aim of knowing how good the results obtained are. It is also presented the limitations of the method applied.

4.1 Reliability Method

At the same time that the histograms of the area of each PI were obtained, a file was created also for each PI, but this time that file had six columns, which the first two corresponded to the total number of pixels of the ROI of both images (MI and PI), the third and fourth columns were respectively the standard deviation of medical data and analysis data and the last two columns were the intensity value of brightest pixel (medical and analysis). After obtaining the files mentioned before, a code was developed to analyze all that data by comparing the results obtained with the medical diagnosis in the data base of mammograms to get the reliability of this method. This time the purpose was obtain the relative errors. The code compared the medical and the analysis data to be able by means of the equation 2.2 to get the error between them and with this got three errors and their respective histograms which are shown below.

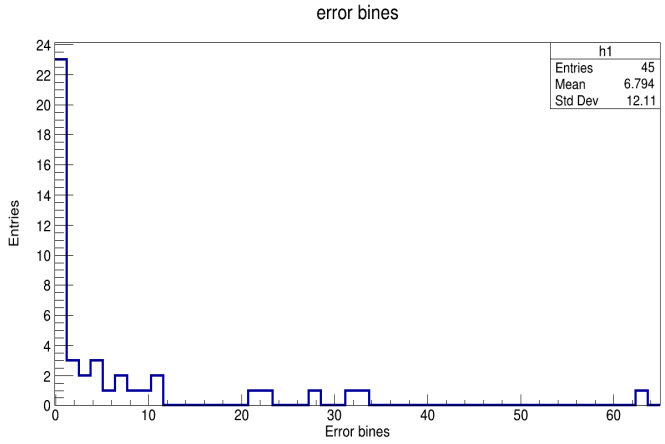

Figure 4.1: Relative error of pixels

The figure 4.1 is the histogram where the relative error of the pixels can be appreciated, this was obtained by comparing the total number of pixels of the abnormality region in the MI, and the total number of pixels in the area found by data analysis of the PI. It can be seen in the histogram that 77.7% of the images, i.e., 35 mammograms have a relative error of less than 10.

The histogram of the relative error of the standard deviation can be seen in the figure 4.2, to obtained this error chart the standard deviation of the MI and the PI were compared using equation 2.1. For this case, 41 mammograms which represents the 91% of the sample, have a relative error of less than 2.

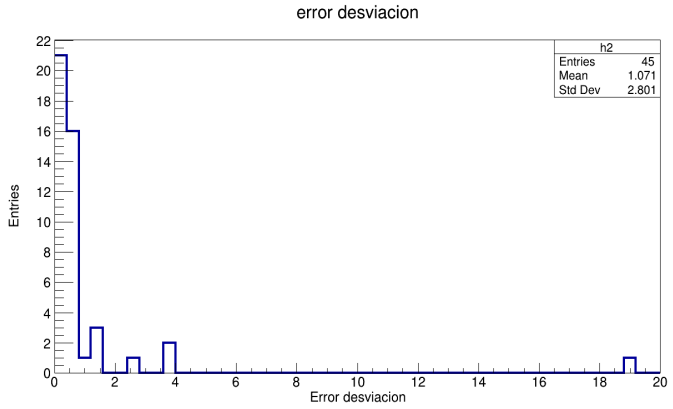

Figure 4.2: Relative error of the standard deviation

Finally the figure 4.3 is the chart of the relative error of the intensity value, it was made by comparing the pixels of higher intensity whose values are shown in the table 4.2, in this chart it is observed that 93.3% of the sample (42 images) have a relative error of less than 0.2.

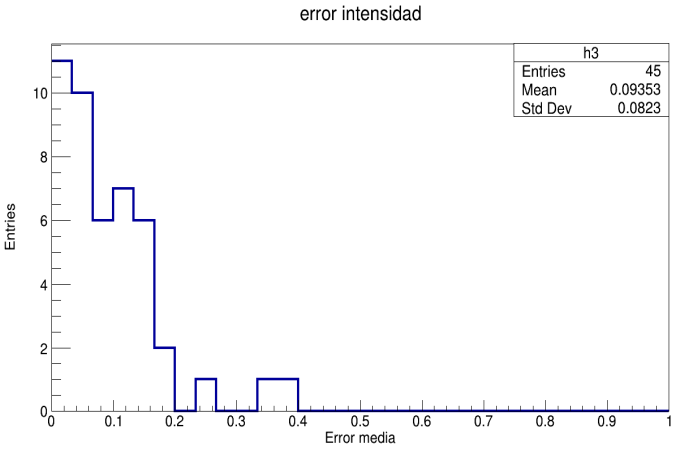

Figure 4.3: Relative error of the intensity

The relative error results indicate that the data analysis method employed is compatible with the medical data found in the database. Also, with the results in the figure 4.3 it can be said that the brightest pixel correspond to the abnormality centre. To be more certain about the results of the charts, further analysis should be done to determine exactly how much variation there is between medical and analysis data.

4.2 Table of results

Name	Medical Pixels of ROI	Analysis Pixels of ROI
mdb023	3365	96440
mdb028	12545	8095
mdb058	2917	2670
mdb072	3081	35133
mdb075	2117	1301
mdb090	9605	80478
mdb095	3365	28626
mdb102	5777	36455
mdb105	38417	174886
mdb110	10405	57168
mdb111	45797	47864
mdb115	35803	62841
mdb117	28225	8592
mdb120	24965	34118
mdb124	4357	105388
mdb125	14401	85056
mdb130	3137	104675
mdb134	9605	4387
mdb141	3365	6485
mdb148	121105	57698
mdb155	36101	9625
mdb158	30977	9632
mdb170	26897	53079
mdb171	15377	170867
mdb178	19601	34022
mdb179	17957	99681
mdb181	11665	9830
mdb184	51985	39349
mdb186	8837	16502
mdb202	5403	61587
mdb206	1157	5155
mdb209	30277	27509
mdb211	703	44933
mdb213	8101	1129
mdb231	1157	2575
mdb238	6401	217762
mdb239	5777	129477
mdb241	9217	23604
mdb249	3081	241940
mdb253	5477	2331
mdb256	14401	40856
mdb265	12545	19845
mdb267	20737	3305
mdb270	18497	7785
mdb271	60517	302

Table 4.1: table of the ROI in pixels

Name	Medical S.D	Analysis S. D.	Medical Brightest Pixel	Analysis Brightest Pixel
mdb023	0.0443503	0.0332183	0.804	0.708
mdb028	0.00184178	0.0365587	0.604	0.812
mdb058	0.0365876	0.0232975	0.7	0.708
mdb072	0.0371385	0.0490462	0.668	0.708
mdb075	0.121788	0.0737316	0.932	0.708
mdb090	0.0291343	0.0469103	0.836	0.836
mdb095	0.0581538	0.0326303	0.732	0.724
mdb102	0.0495472	0.0407087	0.836	0.764
mdb105	0.0183063	0.0649382	0.932	0.876
mdb110	0.0403358	0.0407903	0.836	0.708
mdb111	0.0730184	0.0503375	0.836	0.708
mdb115	0.0415081	0.0545784	0.604	0.836
mdb117	0.0596026	0.297683	0.628	0.708
mdb120	0.0496887	0.0286317	0.74	0.74
mdb124	0.0519237	0.0467402	0.876	0.748
mdb125	0.0478509	0.0405734	0.804	0.708
mdb130	0.0310133	0.0366252	0.82	0.708
mdb134	0.0598028	0.0257042	0.756	0.756
mdb141	0.0385487	0.0102886	0.668	0.708
mdb148	0.0679025	0.0398279	0.668	0.708
mdb155	0.0406953	0.0272946	0.668	0.708
mdb158	0.0329392	0.0166355	0.668	0.708
mdb170	0.0565925	0.0347072	0.74	0.74
mdb171	0.0128555	0.0609702	0.924	0.9
mdb178	0.0729915	0.0403622	0.821	0.708
mdb179	0.024732	0.0605705	0.924	0.924
mdb181	0.0313419	0.0223272	0.644	0.708
mdb184	0.0921129	0.0677743	0.668	0.708
mdb186	0.104151	0.0348243	0.604	0.708
mdb202	0.0503264	0.0326983	0.7	0.708
mdb206	0.0365925	0.0141175	0.612	0.708
mdb209	0.0613698	0.0340877	0.788	0.756
mdb211	0.0194635	0.0432292	0.86	0.708
mdb213	0.0238232	0.0119117	0.636	0.708
mdb231	0.0397485	0.0135071	0.668	0.708
mdb238	0.0484978	0.055985	0.876	0.836
mdb239	0.0192612	0.0305675	0.788	0.716
mdb241	0.0505891	0.0364844	0.796	0.708
mdb249	0.0164152	0.0399953	0.836	0.836
mdb253	0.0149919	0.00969561	0.636	0.708
mdb256	0.0539033	0.0335427	0.804	0.732
mdb265	0.0330559	0.0183186	0.644	0.708
mdb267	0.0416002	0.0338219	0.652	0.708
mdb270	0.0615071	0.03322722	0.764	0.764
mdb271	0.0266423	0.00342454	0.628	0.708

Table 4.2: Table of metrics

Table 4.1 shows the total number of pixels of the ROI of each mammogram used, while the table 4.2 is the table of the metrics used for the reliability method when comparing the diagnostic region and the one given by the data analysis.

ROI visual comparison

Figure 4.4 illustrates a visual example of the comparison between the medical diagnosis and the data analysis of the abnormality region. In each image is marked the centre of the abnormality.

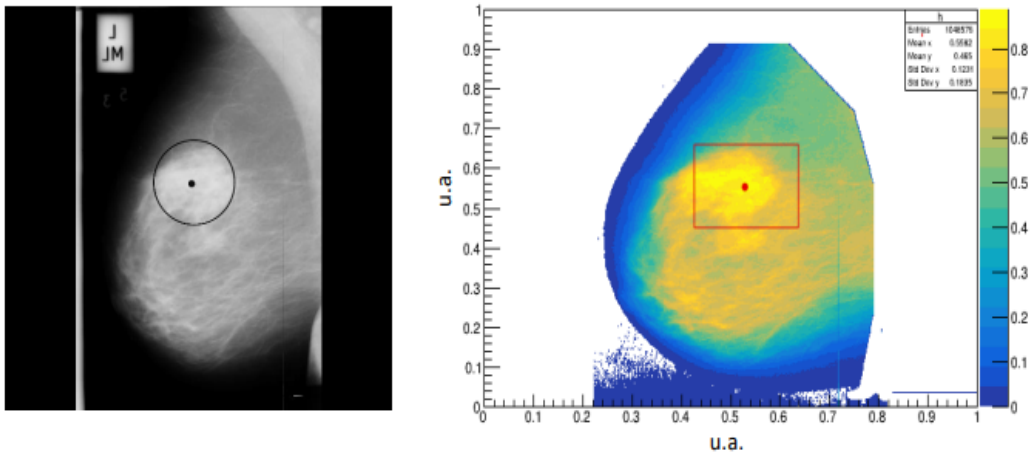

Figure 4.4: The image on the left is the diagnosis showing the ROI. The image on the right is the ROI calculated with the data analysis method with a.u.

It may be noted that in the first table the total number of pixels in some mammograms vary considerably between the two amounts. This is because ROOT detected every pixel with intensity greater than 0.7 and it exists the possibility that some of those pixels are part of the abnormality or just dense tissue patterns. It can be difficult to identify them based in the intensity considering the limitations set out below. To determine if it is part of the tumor or not, more detailed tests may be done to differentiate tumor tissue from dense healthy tissue with respect to intensity or use some other criteria than intensity to achieve it.

4.3 Method limitations

Although this method of data analysis was able to achieve the main objective, there are certain important limitations that must be mentioned:

- **Form:** This method can be seen as an approximation since the region was calculated as if the **ROI** has a defined form, in this case, it was calculated as a square and actually abnormalities present in the breast have undefined shapes and they do not have well-defined edges.
- **Some abnormalities:** Since calcifications are deposits of calcium, which means there are a set of elements, the problem is that this method can determinate the area of only one element and it will depend on where the pixel with the highest intensity is located.
- **Sample size:** The sample taken for this thesis is made up of 45 mammograms and although the results are good, it cannot be ruled out to do the same analysis with a bigger sample.

- **Mammography:** Sometimes the grey-scale of a mammogram is a limitation because in some cases the tissues present similar radiological features [18] that can cause that the intensity values of an abnormality area may be similar to the intensity values of patterns belonging to dense breast.

Chapter 5

Conclusions

Breast cancer is a disease of global importance and still has much that can be studied about it. For example, to improve existing techniques or develop new diagnostic and treatment techniques, specially for better detection in early stages. There are many research works that help against this disease, either to categorize and diagnose it or to classified it. As were mentioned in The State of the Art in Chapter 1. Other methods related to Artificial Intelligence and Artificial Neural Networks currently are used for those purposes. Several articles exists on the application of these methods with different objectives and procedures that help to known more about cancer.

The results obtained in this work support the achievement of the general objective which is comparing the abnormality area between the medical diagnosis and the data analysis using a sample of 45 mammograms. At the same time, the results of this work reinforcing the results of the study of Zepeda et. al [12], since they suggested that the ROI has intensity values grater than 0.7 and that the brightest pixel correspond to the center of the abnormality.

It was sought to make a statistical analysis as a method to give reliability to the results, which consisted in calculating the relative errors to be able to notice how similar they were respect to the given medical diagnosis. It was found that respect the total number of pixels in the abnormality region, 77.7% of the images had a error less than 10 (see figure 4.1). Respect the standard deviation in the figure 4.2, 91% of the sample had a error less than 2. In the equation 2.1 the closer the ratio is to zero, it means that there is not much error between the data, in this case, indicates that the results of this method is compatible with the data of the medical diagnosis. Finally for the relative error of the mean that can be observed in the figure 4.3 it is concluded that the 93.3% had a error less than 0.2. However, more analysis should be done to determine how much the data analyzed varies.

While analyzing the results it was found that some pixels that had tumor intensity were not considered in the medical diagnosis but they were detected by ROOT. However it is difficult to determine the reasons why those pixels were not considered in the region of the tumor, it could be that those pixels were discarded by something specific or like was mentioned in chapter 4 it may be because the tissues may have similar radiological characteristics causing it to be difficult to determined which ones belong to the tumor.

Unless the procedure used to identify the area of each mammogram is a little simple, it can be said that the study fulfilled the hypothesis and the objectives set. Considering that the parameterized image PI preserves the exact location of each pixel in the MI and it also can be noticed that this method of data analysis matches with the medical diagnosis which means this method can be reliable to determine the tumor area. The use of this method would be

recommended as a complement to medical diagnosis as an approximated area of the tumor.

Although the results are good and reliable since the errors are small and they were the expected ones. In order for the results of this analysis to improve, future work could involve examining a larger sample of mammograms.

Bibliography

[1] World Health Organization: WHO. (2022). Cáncer. www.who.int. https://www.who.int/es/news-room/fact-sheets/detail/cancer

[2] Organización Mundial de la Salud. (2021). Cáncer de mama. https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

[3] Cáncer de mama. (n.d.). OPS/OMS | Organización Panamericana De La Salud. https://www.paho.org/es/temas/cancer-mama

[4] Instituto Nacional de Estadística y Geograf. (2022, October 17). Estadísticas a Propósito del Día Internacional de la Lucha Contra el Cáncer de Mama. Press release. https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2022/EAP_CANMAMA22.pdf

[5] Guevara-Ponce, V., Roque-Paredes, O., Zerga-Morales, C., Flores-Huerta, A., Aymerich-Lau, M., & Iparraguirre-Villanueva, O. (2023). Detection of Breast Cancer using Convolutional Neural Networks with Learning Transfer Mechanisms. International Journal of Advanced Computer Science and Applications, 14(6). https://doi.org/10.14569/ijacsa.2023.0140661

[6] Arévalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & López, M. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. Computer Methods and Programs in Biomedicine, 127, 248–257. https://doi.org/10.1016/j.cmpb.2015.12.014

[7] Shawly, T., & Alsheikhy, A. A. (2023). An improved fully automated breast cancer detection and classification system. Computers, Materials & Continua, 76(1), 731–751. https://doi.org/10.32604/cmc.2023.039433

[8] Kuo, C. J., Chen, H., Barman, J., Ko, K., & Hsu, H. (2023). Complete, fully automatic detection and classification of benign and malignant breast tumors based on CT images using artificial intelligent and image processing. Journal of Clinical Medicine, 12(4), 1582. https://doi.org/10.3390/jcm12041582

[9] E. Strelcenia and S. Prakoonwit, "Improving Cancer Detection Classification Performance Using GANs in Breast Cancer Data," in IEEE Access, vol. 11, pp. 71594-71615, 2023, doi: 10.1109/ACCESS.2023.3291336.

[10] Chen, S., Wu, Y., Pan, C., Lian, L., & Q, S. (2023). Breast ultrasound image classification and physiological assessment based on GoogLeNet. Journal of Radiation Research and Applied Sciences, 16(3), 100628. https://doi.org/10.1016/j.jrras.2023.100628

[11] López, A. (2022). Separación de regiones de interés para el diagnóstico y tejido de fondo en mammografías mediante análisis de datos. [Thesis to obtain the bachelor's degree, Benemérita Universidad Autónoma de Puebla] Repositorio Institucional de Acceso Abierto BUAP. https://repositorioinstitucional.buap.mx/handle/20.500.12371/16677

[12] Zepeda, C.H, Vázquez, M.G, Moreno, E., de Celis, B., Herrera, K. and Rodríguez, M. (2021, 26 enero). Tumor and microcalcification characterization using entropy, fractal dimension and intensity values statistical analysis in mammography. arXiv.org. https://arxiv.org/abs/2101.11090

[13] Sociedad Española de Oncología Médica (2023). Cancer de mama. https://seom.org/info-sobre-el-cancer/cancer-de-mama?start=4

[14] What is breast cancer? (2023, July 27). Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

[15] American Cancer Society. (n.d.). What is breast cancer? https://www.cancer.org/cancer/types/breast-cancer/about/what-is-breast-cancer.html

[16] Breast Cancer Stages and TNM Classification | Penn Medicine. (n.d.). Penn Medicine - Abramson Cancer Center. https://www.pennmedicine.org/cancer/types-of-cancer/breast-cancer/breast-cancer-staging

[17] Stages of Breast Cancer | Understand Breast Cancer Staging. (n.d.). American Cancer Society. https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html

[18] American Cancer Society. (n.d.). Densidad de los senos e informe de su mamograma. https://www.cancer.org/es/cancer/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/mamogramas/la-densidad-de-los-senos-y-el-informe-de-su-mamograma.html

[19] What does it mean to have dense breasts? (2023, July 27). Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/dense-breasts.htm

[20] What is cancer? (2021, 11 octubre). National Cancer Institute. https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[21] What is cancer? | Cancer Basics | American Cancer Society. (s.f.). American Cancer Society. https://www.cancer.org/cancer/understanding-cancer/what-is-cancer.html

[22] What does the doctor look for on a mammogram? (n.d.). American Cancer Society. https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/mammograms/what-does-the-doctor-look-for-on-a-mammogram.html

[23] Bassett, L. W. (2003). The abnormal mammogram. Holland-Frei Cancer Medicine - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK12642/

[24] Understanding breast calcifications. (n.d.). https://www.breastcancer.org/screening-testing/mammograms/what-mammograms-show/calcifications

[25] Breast Cancer mammogram | How does a mammogram work? (s.f.). American Cancer Society. https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/mammograms/mammogram-basics.html

[26] Shaw, E. (2007). Atlas of Mammography. Lippincott Williams & Wilkins. http://www.osumcradiology.org/resources/Supplimental-Books/Atlas-of-Mammography-(3rd-Edition).pdf

[27] Carlyle, S. (n.d.) Manual de Radiología para Técnicos. Elsevier

[28] Gonzalez, R. and Woods, R. (2002). Digital Image Processing. Prentice Hall

[29] The mini-MIAS database of mammograms. (n.d.). http://peipa.essex.ac.uk/info/mias.html

[30] ROOT: TASImage Class Reference. (n.d.). https://root.cern.ch/doc/master/classTASImage.html

[31] Team, R. (n.d.). About ROOT. ROOT. https://root.cern/about/

[32] Chapra, S. & Canale, R. Numerical Methods for Engineers. 7th Edition. McGraw-Hill.

[33] Taylor, J. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. 2nd Edition. University Science Books. https://www.niser.ac.in/sps/sites/default/files/basic_page/John%20R.%20Taylor%20-%20An%20Introduction%20to%20Error%20Analysis_%20The%20Study%20of%20Uncertainties%20in%20Physical%20Measurements-University%20Science%20Books%20(1997).pdf