UNIVERSITAT DE BARCELONA

Treball final de grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Vision Transformer for Classifying Benign and Malignant Breast Tumors in Mammography

Autor: Noah Márquez Vara

Director: Dr. Oliver Díaz Montesdeoca

Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 10 de juny de 2024

Abstract

The need for precise diagnostic tools to distinguish between benign and malignant breast cancers is underscored by the fact that breast cancer remains a major global health concern. This is the reason why the use of Vision Transformers (ViTs) to categorize breast cancers in mammography is studied. Using the OPTIMAM Medical Database, which includes mammography scans from the UK National Health Service Breast Screening Program, we assess how well ViTs perform on this particular assignment.

The methodology includes proper preprocessing, augmentation, and splitting of data, including significant model training for fine-tuning hyperparameters concerning the prevention of data leakage. Insightful metrics such as AUC-ROC are used in evaluating correctly the model's performance.

The results make ViTs a state-of-the-art alternative to CNNs due to their capacity to capture global context through self-attention mechanisms. This is especially useful in complex tasks like medical imaging interpretation. The potential of cutting-edge AI methods to improve diagnostic precision and enhance patient outcomes during breast cancer diagnosis is highlighted by this study.

Resum

La necessitat d'eines de diagnòstic precises per distingir entre càncers de mama benignes i malignes destaca pel fet que el càncer de mama continua sent una preocupació important globalment. És per això que s'estudia l'ús de *Vision Transformers (ViTs)* per categoritzar els càncers de mama en mamografies. Utilitzant la base de dades mèdica d'OPTIMAM, que inclou mamografies del Programa de Detecció de Càncer de Mama del Servei Nacional de Salut del Regne Unit, avaluem que tan bé funcionen els ViTs en aquesta tasca específica.

La metodologia inclou un adequat preprocessament, augment i divisió de les dades, incloent-hi un entrenament significatiu del model per ajustar els hiperparàmetres en relació amb la prevenció de la contaminació de dades. Mètriques significatives i intuïtives com l'AUC-ROC s'utilitzen per avaluar correctament el rendiment del model.

Els resultats converteixen els ViTs en una alternativa de l'estat de l'art a les CNN a causa de la seva capacitat per capturar el context global a través de mecanismes d'autoatenció. Això és especialment útil en tasques complexes, com ara la interpretació d'imatges mèdiques. El potencial dels mètodes avançats d'IA per millorar la precisió diagnòstica i millorar els resultats dels pacients durant el diagnòstic de càncer de mama es destaca en aquest estudi.

Resumen

La necesidad de herramientas de diagnóstico precisas para distinguir entre cánceres de mama benignos y malignos destaca por el hecho de que el cáncer de mama sigue siendo una preocupación importante a nivel global. Es por ello que se estudia el uso de *Vision Transformers (ViTs)* para categorizar los cánceres de mama en mamografías. Utilizando la base de datos médica de OPTIMAM, que incluye mamografías del Programa de Detección de Cáncer de Mama del Servicio Nacional de Salud del Reino Unido, evaluamos qué tan bien funcionan los ViTs en esta tarea específica.

La metodología incluye un adecuado preprocesamiento, aumento y división de los datos, incluyendo un entrenamiento significativo del modelo para ajustar los hiperparámetros con relación a la prevención de la contaminación de datos. Métricas relevantes como el AUC-ROC se utilizan para evaluar correctamente el rendimiento del modelo.

Los resultados convierten a los ViTs en una alternativa del estado del arte a las CNN debido a su capacidad para capturar el contexto global a través de mecanismos de autoatención. Esto es especialmente útil en tareas complejas como la interpretación de imágenes médicas. El potencial de los métodos avanzados de IA para mejorar la precisión diagnóstica y mejorar los resultados de los pacientes durante el diagnóstico de cáncer de mama se destaca en este estudio.

Acknowledgements

This thesis would not have been possible without the guidance of Oliver Díaz, my supervisor. I have learned about a wide range of topics related to medical imaging and medical physics from his vast knowledge and excitement for research.

I would especially like to thank my friend Alejandro for all of his help and encouragement along this adventure.

The biggest thanks go to my family. Mama, Angel y Ona, sin vuestro cariño y apoyo no hubiera llegado hasta aquí.

Above all, I thank Marina for her love and support. T'estimo.

Contents

1 Introduction **1**
- 1.1 Contextualization of the problem . . . 1
- 1.2 Objectives . . . 2
- 1.3 Project Timeline . . . 3
  - 1.3.1 Tasks to be Developed . . . 3
  - 1.3.2 Gantt Chart . . . 4

2 Background and Clinical Context **5**
- 2.1 Breast Cancer . . . 5
- 2.2 Screening Program . . . 6
- 2.3 Medical Imaging . . . 7
- 2.4 Breast Cancer Detection . . . 11
  - 2.4.1 Challenges of Dense Breast Tissue . . . 12
  - 2.4.2 Challenges for Healthcare Providers and Patients . . . 12
  - 2.4.3 The Impact of Overdiagnosis . . . 12
  - 2.4.4 Chronological Evolution . . . 13

3 Technology and Literature Review **14**
- 3.1 Transformers . . . 14
  - 3.1.1 Overview . . . 14
  - 3.1.2 History and Development . . . 14
  - 3.1.3 Self-Attention Mechanism . . . 15
  - 3.1.4 Positional Encoding . . . 16
  - 3.1.5 Encoder-Decoder Architecture . . . 17
  - 3.1.6 Transformer Architecture . . . 18
  - 3.1.7 Training Transformers . . . 18
  - 3.1.8 Applications of Transformers . . . 20
- 3.2 Visual Transformers . . . 20
  - 3.2.1 Overview . . . 20
  - 3.2.2 Comparison with CNNs . . . 20
  - 3.2.3 Vision Transformer (ViT) Architecture . . . 23
  - 3.2.4 Self-Attention in Visual Transformers . . . 23
  - 3.2.5 Training Visual Transformers . . . 25
  - 3.2.6 Applications of Visual Transformers . . . 27
- 3.3 Transfer Learning . . . 27

3.3.1 Feature Extraction . . . 27
3.3.2 Fine-Tuning . . . 28
3.4 CNNs and ViTs for Breast Cancer Classification . . . 28
3.4.1 CNNs for Breast Cancer Classification . . . 29
3.4.2 ViTs for Breast Cancer Classification . . . 30
3.4.3 Other Methods for Mammography Classification . . . 31
3.4.4 Comparative Analysis . . . 31

4 Data, Methodology, and Tools **32**
4.1 Comparison of Clinical Use Cases of Public Mammography Datasets . . . 32
4.2 OPTIMAM Dataset . . . 33
4.2.1 Dataset Overview . . . 34
4.2.2 Analysis of Manufacturer Distribution . . . 35
4.2.3 Image format and data . . . 37
4.3 Image Preprocessing . . . 40
4.4 Reproducibility through Seed Initialization . . . 41
4.5 Preventing Data Leakage . . . 42
4.6 Data Splitting . . . 42
4.7 Data Loading . . . 43
4.8 Evaluation Metrics . . . 44
4.8.1 AUC - ROC Curve . . . 45
4.8.2 Confusion Matrix . . . 46
4.8.3 Binary Classification . . . 48
4.9 Description of the model architecture . . . 50
4.9.1 Models' Input Format . . . 50
4.9.2 Models' Output Format . . . 51
4.9.3 Transfer Learning and Pretrained weights . . . 52

5 Results, Analysis, and Discussion **54**
5.1 Data Modelling . . . 54
5.2 Data Augmentation . . . 55
5.3 Fine-Tuning Hyperparameters . . . 56
5.4 Model's Evaluation and Selection . . . 57
5.5 Class Imbalance . . . 59
5.5.1 Weighted Loss Function . . . 59
5.5.2 Oversampling the minority class . . . 59
5.5.3 Undersampling the majority class . . . 60
5.5.4 Comparison of the results . . . 60
5.6 Evaluation Metrics and Analysis . . . 61
5.6.1 AUC-ROC Curve and Thresholds . . . 62
5.6.2 Confusion Matrix . . . 63
5.6.3 Binary Classification . . . 63

6 Conclusions and Future Work **68**
6.1 Results . . . 68

6.2 Progress, Limits, and Responsibilities 68

Bibliography **69**

A Libraries and Frameworks **80**

B Code and Model Weights **81**

List of Figures

Figure 1.1 Global Cancer Incidence in Women . . . 1
Figure 1.2 Planification of the work. . . . 4

Figure 2.1 Breast anatomy. . . . 7
Figure 2.2 Breast biopsy. . . . 8
Figure 2.3 X-rays mammography. . . . 8
Figure 2.4 Mammography images. . . . 9
Figure 2.5 X-ray imaging. . . . 10
Figure 2.6 Dedicated breast Computed Tomography (CT) drawing and example of a patient's breast CT. . . . 10
Figure 2.7 Magnetic Resonance Imaging (MRI) drawing and example of a patient's breast MRI. . . . 11
Figure 2.8 Breast Ultrasound drawing and example of a patient's breast ultrasound. . . . 11

Figure 3.1 Scaled dot-product attention. . . . 15
Figure 3.2 Multi-head attention. . . . 16
Figure 3.3 The Transformer - model architecture. . . . 17
Figure 3.4 Artificial Neural Network vs Convolutional Networks. . . . 21
Figure 3.5 The process of convolution of a kernel over the image. . . . 21
Figure 3.6 The world through the eyes of CNN. . . . 22
Figure 3.7 Vision Transformer model overview. . . . 24
Figure 3.8 Self-attention mechanism in Vision Transformers. . . . 24
Figure 3.9 Transfer learning from ImageNet. . . . 29

Figure 4.1 Comparison of clinical use cases of mammography datasets. . 33
Figure 4.2 Distribution of mammography images by manufacturer. . . . 35
Figure 4.3 Comparison of mammography images from different manufacturers. . . . 36
Figure 4.4 Age distribution of clients. . . . 37
Figure 4.5 Distribution of lesion status for clients. . . . 38
Figure 4.6 Example images from the OPTIMAM dataset (before cropping). 40
Figure 4.7 Example images from the OPTIMAM dataset (after background cropping). . . . 40
Figure 4.8 Preprocessing of an example image. . . . 41
Figure 4.9 Data Loading Process . . . 44

Figure 4.10 Random patches after loading the data. 44
Figure 4.11 Overlap of test results for two populations. 45
Figure 4.12 Inverse relation between sensitivity and specificity. 46
Figure 4.13 AUC - ROC Curve plot. 47
Figure 4.14 Basic structure of a confusion matrix. 48
Figure 4.15 Examples of pixel intensity value histograms of pre-processed cropped mammography images. 51
Figure 4.16 Model pipeline diagram. 53

Figure 5.1 Nine preprocessed and augmented images. 56
Figure 5.2 Training and Validation Performance Metrics. 58
Figure 5.3 ROC Curve of the Best Model. 62
Figure 5.4 Precision-Recall Curve 64

List of Tables

Table 4.1	Description of columns in the OPTIMAM dataset CSV.	34
Table 4.2	Annotated masses by status and age	38
Table 4.3	Class distribution for Train, Validation, and Test sets.	43
Table 4.4	Binary classification metrics.	50
Table 5.1	Performance of *google/vit-base-patch16-224* on ImageNet.	55
Table 5.2	Hyperparameter tuning results.	56
Table 5.3	Computed class weights.	59
Table 5.4	Comparison of ViT and ResNet50 results using different methods.	61
Table 5.5	Optimal Thresholds for Balanced Accuracy	63
Table 5.6	Confusion Matrix for Threshold 0.5	64
Table 5.7	Confusion Matrix for Threshold 0.16	64
Table 5.8	Sensitivity and Specificity at different thresholds	65
Table 5.9	F1-Score at different thresholds	65
Table 5.10	PPV and NPV at different thresholds	66
Table A.1	Libraries and Frameworks	80

Chapter 1

Introduction

1.1 Contextualization of the problem

Extensive evidence indicates that breast cancer is a major health concern for women globally. It is the most common cancer among women in 157 out of 185 countries. In 2022, approximately 2.3 million women worldwide were diagnosed with the disease, and around 670,000 deaths were attributed to it globally [1].

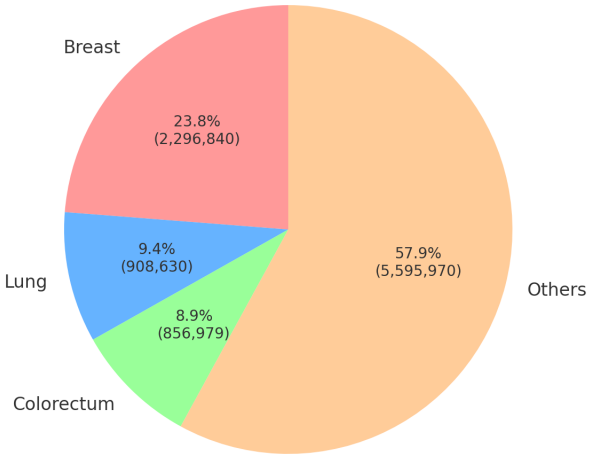

Figure 1.1: Global cancer incidence distribution in women for 2022, showing the top three types: breast, lung, and colorectal cancer, along with other types [2].

Breast cancer burden data normally provides global profiles, with noticeable disparities between human development levels. In developed countries, 1 in 12 women are diagnosed with breast cancer at some point in their lives, and 1 in 71 will die from it. In

contrast, in underdeveloped countries, 1 in 27 women will receive a diagnosis of cancer in their lifetime, but for them, the disease will prove fatal to 1 in 48 women [3].

More screenings programs using X-rays as early detection techniques translate into more diagnosed patients, but it is still hard to detect the cancer in certain cases (i.e., patients with high density breasts), so the number of false positives should however be reduced as this is the root of stress and unnecessary biopsies. The most significant concern though, is the false negatives, as of the lack of evidence, the situation of the cancer will probably evolve further until it is too late to be treatable [4] [5].

Nowadays, the mammography CAD (Computer Aided Detection) software is used for the interpretation of a mammogram, although it did not boost medical research after its introduction in the 90s. On the opposite, it was Deep Learning (DL) that finally started to shine in comparison with human experts when analyzing medical images [6].

This project evaluates the performance of a Machine Learning (ML) system developed to predict cancer likelihood from screenings from the OPTIMAM Medical Database (OMI-DB) [7], which relies on data obtained through the National Health Service Breast Screening Program (NHSBSP), which processes screening images from many breast cancer centers throughout the United Kingdom [8].

1.2 Objectives

The main goal and objective for this thesis is the development of an Artificial Intelligence (AI) based tool differentiating between the benign and malignant nature of breast tumors using screening mammography images. This project involves exploring various ML methods, but the main challenge concerns the analysis of purely image-based data from the OPTIMAM dataset without any extras or other clinical variables. The OPTIMAM screening dataset is provided with no pre-labeled features, such as size, shape, or density markers, which are usually attributed to an increased diagnostic accuracy. The approach taken then needs further exploration for new ways in using ML models to make up for the lack of explicit guidance on the features to optimize a high diagnostic accuracy level, purely stemming from the image.

For that, the thesis will not dwell deeply into Convolutional Neural Networks (CNN) but instead compare them with the latest technology on Vision Transformers (ViT), which make up a new class of DL models that seem to handle sequences and global context within them. The comparison will be made on both CNNs and ViT to independently perceive and classify the visual cues of breast tumor malignancy without clarifying clinical data. Attention inside the comparative analysis will be focused on how well either of these takes the interwoven intricate elements from the images and makes use of them, free of additional information concerning the clinical context. This is important, as the ability of models to perform in a feature-absent mode corresponds to the near real-life task involving the application of the model in clinical situations. Therefore, the ultimate

goal here is to find the best possible approach for accurate tumor classification based on image content alone, as the two sets of models are evaluated under the same conditions.

Additionally, we will further study and evaluate the unique capabilities and possible improvements of ViTs in comparison with traditional CNNs. In medical imaging, small details can sometimes be very critical to the diagnostic process, where the fine-grained details are crucial for an accurate diagnosis. In this regard, the self-attention mechanism of ViTs, just like the one of conventional Transformers, promisingly captures long-range dependencies in images. Therefore, this part of the experiment is going to examine how the models perform with the complex patterns in mammography images and compare their effectiveness in detecting subtle features associated with malignancy.

Moreover, the project will involve a series of experiments in fine-tuning and optimizing the performance of ViTs. These will involve hyperparameter tuning, and data augmentation techniques, to guarantee the robustness and reliability of the results. The objective is to provide clear results on which model will return both the most accurate and reliable predictions on breast tumor classification, and therefore be of potential interest in presenting important insight as to the strength of these AI methods when applied to medical diagnostics.

1.3 Project Timeline

Aiming to accomplish the goals of the thesis, which will focus on ViTs to categorize mammography images between benign and malignant using Portable Network Graphics (PNG) images from the OPTIMAM database, a set of tasks have been defined. In this manner, the study is composed of an analysis of the medical context in relation to breast cancer, a review of applications in breast cancer for AI, ML, and DL, an analysis of ViTs, and a deep analysis of the information present within the OPTIMAM dataset. The timeline and sequence of these activities are shown in 1.2.

1.3.1 Tasks to be Developed

State of the Art

1. Review medical literature and practice related to the clinical context of breast cancer.

2. DL, ML, and AI in breast cancer: Understand their current use for diagnosis of breast cancer, and what can the possible benefits be.

Implementation

1. Organize all the data and images from the OPTIMAM database in an easy-to-access way.

2. Dataset Analysis and Preprocessing: The OPTIMAM dataset has to be analyzed, cleaned, and preprocessed so that it is amenable to the steps further in the preprocessing of images and hence prepared for training the ViT model.
3. Code Implementation: Develop, implement, and test the code that will train the ViT model with the preprocessed data.
4. Result analysis: Assess the performance of the trained model and properly interpret the results so that meaningful conclusions can be made.

Report Writing

1. Report writing: Document the research process, methodology, results and conclusions, in a detailed way.
2. Corrections: Send the suggested corrections back and forth to my thesis supervisor, receive the feedback, and revise again.
3. Submission of the thesis.
4. Thesis Presentation: Elaborate upon the most significant results and findings in the thesis in detail.

1.3.2 Gantt Chart

The Gantt chart graphically represents the time frame and the milestones of the project over a time frame. It is an effective instrument for tracing the course of project work in such a manner that it reaches the desired stages at set deadlines.

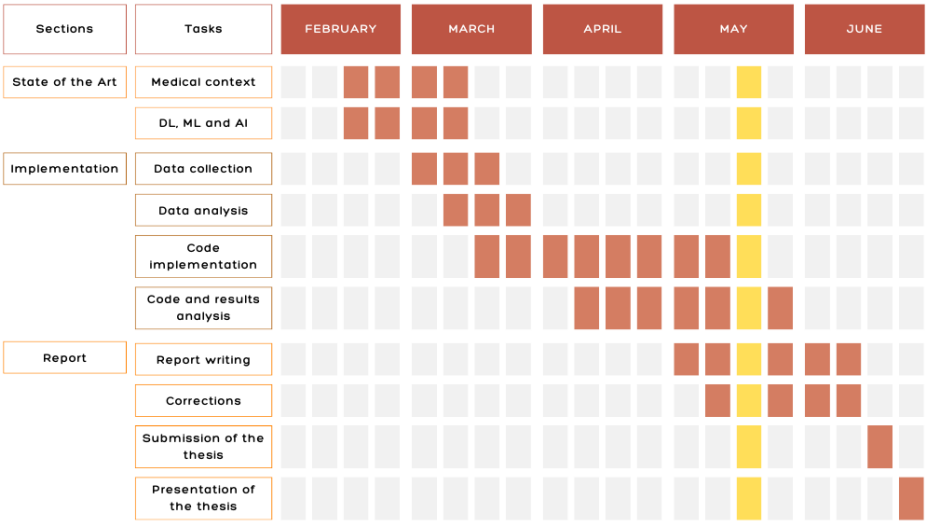

Figure 1.2: Planification of tasks over the semester, detailing the schedule for various sections in the thesis.

Chapter 2

Background and Clinical Context

This chapter paves the way to an understanding of the landscape of breast cancer and the importance of early detection within a clinical environment. We will distinguish between malignant and benign tumors, screening importance across different regions, and the different techniques applied in medical imaging for diagnoses. We will also present problems that high breast density causes, the impact of overdiagnosis, and the chronological evolution of the methods applied to detect breast cancer, pointing out advances from those early mammography images to the most recent, AI-driven technologies. This general overview serves as the background to see the emphasis that is given to the new management of breast cancer and technological improvements that have enhanced diagnostic precision.

2.1 Breast Cancer

There are two main types of cancers[1]: benign and malignant. The benign cancers are less severe since they stay in the place where they originated and do not spread to other locations in the body. Malignant tumors, also known as the cancerous tumors, are more dangerous because they attach themselves and destroy other tissues of the body. Such forms of tumors result from abnormal growth of cells, which destroys the body internal tissues. These malignant tumors, when not treated in time, turn out to be incurable and result in death [10, 11].

The general perception that at the diagnosis of a tumor, one's life is at risk, is not really the case until it is a confirmed to be a malignant tumor. Benign tumors are not life-threatening and can usually be treated with easy dietary adjustments and a change in lifestyle; they are usually encased in a sac created by the immune system, thus easily removable by surgery. Even though it is rarely the case, benign tumors can also upgrade to malignant, hence the need for, as professionals say, monitoring. Malignant tumors, however, grow quickly in an uncontrolled manner and spread to other parts of the body

[1]Cancer is a disease characterized by the uncontrolled growth and spread of abnormal cells, which can invade or spread to other parts of the body [9].

(i.e., metastasis), thus it is a very sensitive area that needs careful attention among health professionals when identified [12].

2.2 Screening Program

The high public health features of breast cancer screening programs call for the early identification of breast cancer in asymptomatic populations. Such programs greatly enhance the prognosis by identifying malignancies at an earlier, more treatable stage. NHS Breast Screening Programme invites women between 50 and 70 for a mammogram every three years. It is targeted at this age group because breast tissue in younger women tends to be denser, making mammograms less effective in detecting abnormalities [13, 14]. Moreover, the incidence of breast cancer increases with age, so regular screening is needed for early detection and successful treatment [15].

In the UK, 18,942 cases of breast cancer were detected through the NHS Breast Screening Programme in the 2022-23 period. The number of diagnoses made during this time refers to a screen population of 1.93 million women, which resulted in an incidence rate of approximately 9.8 cancers per 1,000 women screened. This testifies to the screening program's effectiveness in identifying early-stage breast cancers and increasing treatment outcomes while reducing mortality rates [16, 17].

Though regional variations may exist, Spain's breast cancer screening program invites women between the ages of 50 and 69 for biennial mammograms. The Spanish initiative seeks to lower the death rate from breast cancer by promoting early identification and prompt treatment. The selection of this age range, as in the UK, is motivated by the increased incidence of breast cancer in older women as well as the age-related alterations in the composition of breast tissue. Older women with fatty breast tissue have greater contrast on mammograms, increasing the chance of finding cancers early [18].

The Spanish breast cancer screening program finds about 36,395 new instances of breast cancer a year; a large percentage of these diagnoses are in women between the ages of 45 and 65, when hormonal alterations are most common. An estimated 132 cases per 100,000 people are the incidence rate. Given that breast accounts for over 30% of cancer diagnoses among Spanish women, this program emphasizes the critical role that early diagnosis plays in enhancing treatment outcomes and lowering death rates [19].

These screening programs are of immense significance and save lives through the detection of cancer at an early stage. Many women participating in these programs do not have any symptoms, and seemingly, they may have nothing to be worried about; however, screening may detect early-stage cancers that, at an initial stage, become easily treatable. Data from various studies indicates that, on average, regular screening reduces mortality from breast cancer by about 20 percent among those women who are screened compared to those who are not [20].

Yet, even with the effectiveness of these programs, mammography conducted in less developed regions is still performed using film rather than digital technologies. As a result, film mammography, although still valid, is much lower in efficiency and comparatively less convenient regarding storage and analysis. This discrepancy pinpoints a significant bias in AI models trained on mainly digital mammograms, probably missing possibilities in regions requiring methods based on film. This bias should be addressed when designing diagnostic tools to ensure inclusive and effective testing [1].

Broad access to screening, based on modern technologies, should raise the quality of early detection in the healthcare system aimed at further improvement for all breast cancer patients worldwide.

2.3 Medical Imaging

Mammograms are X-ray images of the breasts used to detect and diagnose breast cancer. This is thanks to the different attenuation properties of the breast tissues, mainly adipose and glandular. The composition of breasts differs between individuals and changes throughout life. Figure 2.1 illustrates the breast anatomy.

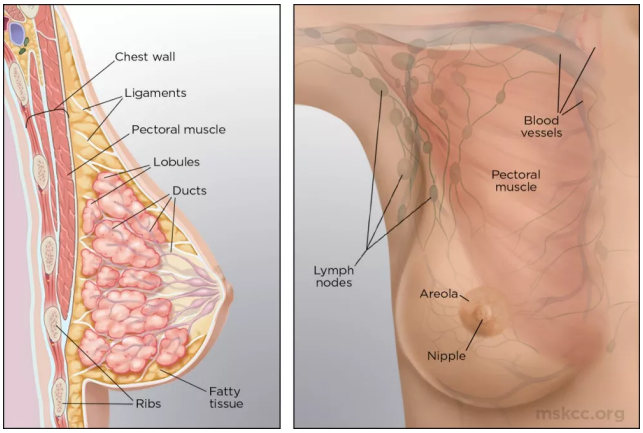

Figure 2.1: Illustration of the breast's anatomy [21].

Besides that, mammograms are used, sometimes in combination with other imaging techniques, to detect anomalies in breast tissue. If an abnormality is found, the clinician will take a tissue biopsy to extract cells and assess if the sample is malignant (cancerous) or benign (non-cancerous). As depicted in Figure 2.2, a core needle biopsy is the most often used technique for doing a breast biopsy. During this process, tissue samples from the breast are taken with a hollow needle and subsequently inspected under a microscope to look for cancerous cells. To precisely target the breast area of concern, imaging techniques like mammography, Magnetic Resonance Imaging (MRI), or ultrasound are typically used as guidance [22].

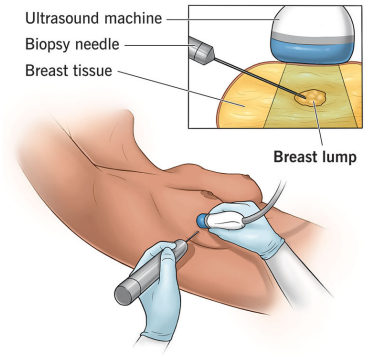

Figure 2.2: Illustration of a breast biopsy procedure showing the use of an ultrasound machine, biopsy needle, breast tissue, and the location of a breast lump [23].

X-rays, Computed Tomography (CT), ultrasound and MRI are the four forms of medical imaging techniques currently most frequently used.

For mammography, women are positioned as shown in Figure 2.3. The breast is compressed between two compression plates, and a planar image is acquired using X-rays.

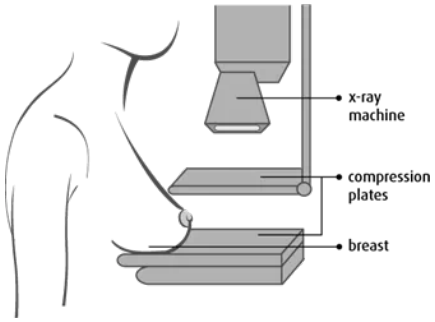

Figure 2.3: Illustration of an X-rays mammography procedure showing the X-ray machine, compression plates, and breast positioning [24].

The results of this procedure are shown in Figure 2.4. Mammography results in two-dimensional (2D) images, exposing details of textures present in breast tissues. Soft tissues that are more glandular and denser will absorb fewer X-ray photons and generally are shown as darker areas in the mammogram. This is the opposite of the brightened tissue in the image, which indeed represents denser (adipose) tissue. Due to the close attenuation properties of both dense tissue and potential cancers (i.e., they look as white on the mammogram) it proves difficult for radiologists to distinguish between these two types of tissue [25, 26, 27].

Furthermore, normal tissue may occasionally appear abnormal due to the compression of the breast tissue during mammography, despite this being necessary to generate quality images. Tissue structures may overlap as a result of this compression, producing shadows or other regions that could be misinterpreted for suspicious lumps or calcifications. This emphasizes how crucial it is to distinguish between benign and malignant findings accurately by using additional imaging modalities and radiologists' skill [28].

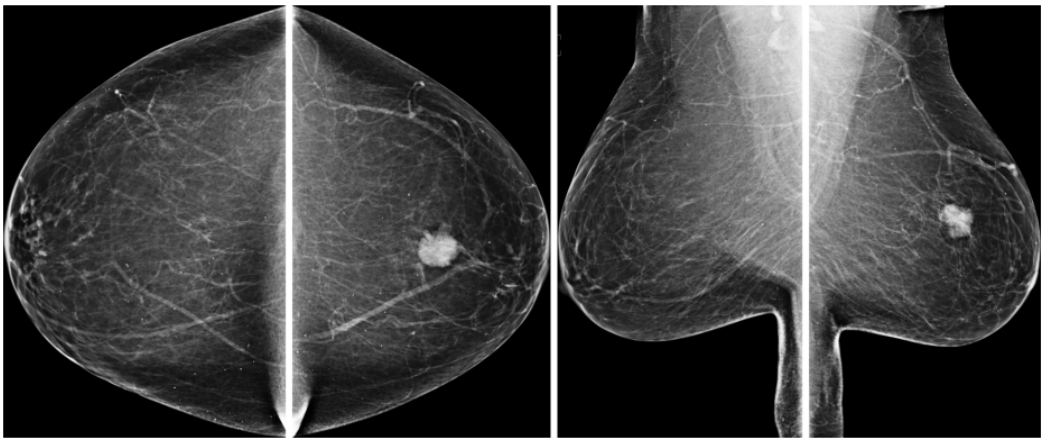

Figure 2.4: Typical screening mammograms views. From left to right: right breast Cranio-Caudal (CC), left breast CC, right breast Mediolateral Oblique (MLO), and left breast MLO [7].

Two mammography views are usually employed for each breast, as shown in Figure 2.4: the CC and the MLO. Four pictures make up a standard screening mammography: two MLO and two CC, one for each breast [29]. Each offers unique, insightful viewpoints:

- **MLO**: An angled view, at 45 degrees, showing the breast and part of the armpit (underarm) area, where lymph nodes reside. This view may become necessary for identifying lymph nodes and upper-outer regions of the breast, which are the most common locations in which breast cancer will develop [30].
- **CC**: This is a view taken from above the breast (0^o), looking downwards. It gives a clear view of the inside to outside (medial to lateral) aspects of the breast. It helps in detecting changes or abnormalities in the center of the breast as well as to compare between the two breasts directly [31].

Combining these two perspectives is crucial for many different reasons and gives the radiologist insightful information. By ensuring more thorough coverage of the breast tissue using both views, it lowers the possibility of overlooking abnormalities. Diagnostic accuracy can be improved by cross-checking abnormalities that are present in one view with the other view. Furthermore, depth is provided by various angles, which aids in differentiating between the breast's surface and deeper components.

A detector, as shown in Figure 2.5, detects the amount of electromagnetic particles received by X-rays from a source as they pass through the patient's breast. The detector then uses this data to create an image. This can make it more difficult for radiologists to interpret calcifications and overlapping soft tissues. CT scans that use X-rays are used to overcome this constraint.

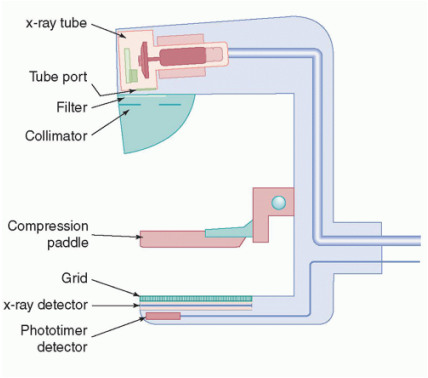

Figure 2.5: Diagram of an X-ray imaging system [32].

CT uses the same X-ray technology, but one more dimension is added to create three-dimensional (3D) volumes, which consist of many 2D projections. The cone-beam geometry is used in dedicated CT scans, so the energy source and the detector rotate together around the patient, as shown in Figure 2.6, taking the projections from different angles [33].

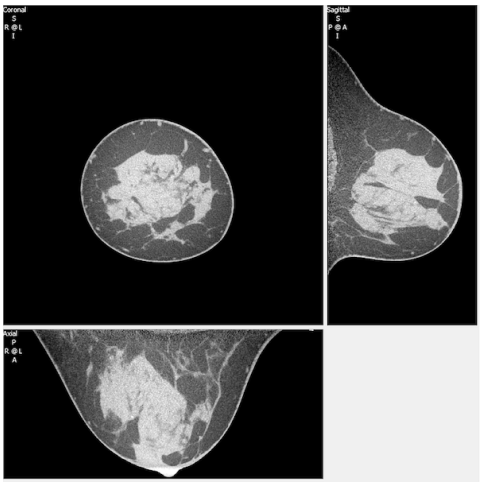

Figure 2.6: (Left) Dedicated breast CT drawing [34]. (Right) Example of a patient's breast CT [35].

MRI scans use strong magnets and radio waves to take detailed images of the body's internal structures. This technique excites a change in the direction of the rotational axis of protons located inside water molecules within living tissues. MRI is most effective

when imaging non-bony parts or soft tissues —such as shoulder ligaments or the brain— since it can provide a high-contrast image of these areas [36]. It is important to notice that although MRI is a method of non-ionizing and very detailed imaging, CT scanning using X-rays is generally more expensive than traditional X-ray imaging [37].

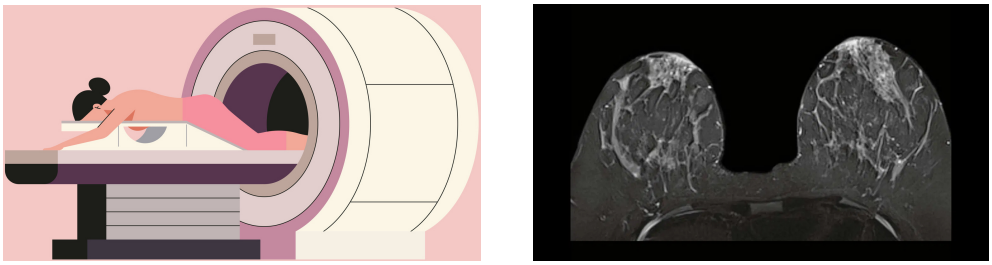

Figure 2.7: (Left) MRI drawing [38]. (Right) Example of a patient's breast MRI [39].

Breast ultrasound is the application of very high-frequency sound waves to make a representation of internal structures within the breast. By moving a handheld transducer over the skin emitting sound waves, these bounce back off tissues to form echoes. These echoes are then represented in real-time as an image on the monitor. Breast ultrasound is most useful for breast lumps, where it can help differentiate solid tumors from fluid-filled cysts[2]. This is a non-invasive technique, which does not use ionizing radiation, and very often becomes an additional tool to mammography, especially for women with dense breast tissue [41].

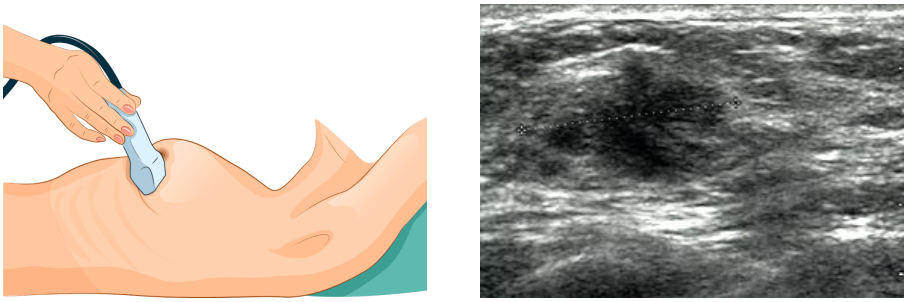

Figure 2.8: (Left) Breast Ultrasound drawing [42]. (Right) Example of a patient's breast ultrasound [43].

2.4 Breast Cancer Detection

Early detection of breast cancer is a significant part of the entire process of care that has a lot to do with the patient's prognosis and the effectiveness of healthcare, among a

[2]Cysts are fluid-filled sacs that can develop in various tissues of the body, including the breast. They are typically benign (non-cancerous) and can range in size from very small to several centimeters. Breast cysts are common and may cause discomfort or pain, especially if they grow or become inflamed [40].

few other essential benefits:

- **Improved patient outcomes**: Breast cancer that is detected early leads to much better patient outcomes, as it is easier to perform timely interventions. At an early, localized stage, the 5-year survival rate of breast cancer is 99% [28]. Early detection means very effective treatments, resulting in less possibility for the spread of the cancer cells and generally high survival rates for the patients [44].

- **Reduced healthcare costs**: Among the most valuable gains is that by minimizing the number of false-positive numbers, early detection prevents the patients from not needed medical procedures, which are equally costly and stressful. Unnecessary procedures do not only reduce the cost of healthcare but also spare the patients from the emotional and physical burdens of unneeded treatments [45].

2.4.1 Challenges of Dense Breast Tissue

Non-calcified lesions are difficult to identify in women with dense breast tissue, which have a higher percentage of glandular and fibrous tissue than adipose tissue. Dense breast tissue makes it difficult to spot abnormalities in mammograms because it appears white, much like how tumors appear. In order to assure proper detection and diagnosis, this intricacy needs the use of modern medical imaging techniques including breast tomosynthesis (stack of 2D images), breast MRI, and breast ultrasound [46].

2.4.2 Challenges for Healthcare Providers and Patients

Following abnormal screening results, healthcare providers frequently have to make difficult decisions. Some of these considerations include deciding whether to request additional testing or to "wait and see". Patients without cancer may experience increased anxiety as a result of this decision-making process, which may enhance sensitivity but reduce specificity. Anxious about their health, patients in the "gray zone" may insist on biopsies, which unintentionally raises the workload of the healthcare system and the number of false positives. Additionally, when a biopsy by itself yields unclear results, there are difficulties in limiting invasive treatments for high-risk individuals.

2.4.3 The Impact of Overdiagnosis

In addition to contributing to the stress and worry of the situation, overdiagnosis, including false positives and false negatives, emotionally burdens patients and their families. While early diagnosis greatly increases the chance of survival and lowers treatment costs from stage 1 to stage 4, overdiagnosis can put patients at risk for needless therapies for lesions that are unlikely to become malignant. This aspect highlights the necessity of precision in diagnostic methods in order to weigh the advantages of early detection against the risks of overdiagnosis [47].

2.4.4 Chronological Evolution

Breast cancer early detection has seen a significant development and revolution in the last few decades, mainly because of technological improvements that have advanced diagnostic accuracy and improved patient outcomes.

Early Techniques and Mammography (1960s-1980s)

Mammography came into its own in the 1960s and remained a gold standard for its potential contribution to early detection of breast cancer, especially calcifications that are early signs of malignancy. Still, limitations exist in mammography, such as in women with dense breast tissue where its sensitivity drops [48].

CAD Systems (1990s)

The introduction of CAD systems during the 1990s was truly commendable. CAD systems aid radiologists by annotating suspicious areas in mammogram images that may go unnoticed during initial reviews. This system increased detection rates and decreased oversight errors, but it also raised the false positive rate, thereby causing unnecessary biopsies and anxiety for the patients [49].

AI and ML (2000s)

AI and ML were integrated in the 2000s. These tools examine huge amounts of imaging data to find patterns and abnormalities that might point to malignancy. It has been proven that AI and ML models can increase sensitivity and specificity while decreasing false positives and negatives [50].

DL and CNNs (2010s)

Medical imaging was greatly influenced by CNNs and DL. In certain situations, these sophisticated models —trained on big datasets— can perform better than radiologists and conventional CAD systems. Particularly in difficult cases with dense breast tissue, DL methods have greatly improved the accuracy of breast cancer detection [51].

Transformers and ViTs (2020s)

Transformers —more specifically, ViTs— have been used in medical imaging recently. ViTs are superior to CNNs in capturing long-range dependencies and obtaining high accuracy with fewer data samples, interpreting images via self-attention mechanisms. According to early studies, ViTs can further enhance breast cancer detection, particularly in more precisely differentiating benign from malignant lesions [52, 53].

Chapter 3

Technology and Literature Review

In this chapter, there is a in-depth overview of the foundational technologies and existing literature relevant to the thesis. We begin with the impact of Transformers on neural network architectures, focusing on the innovative self-attention mechanisms and their breakthrough in tasks ranging from machine translation to image recognition. After that, we examine applications of Transformers in the visual domain and portray their strengths and weaknesses compared to CNNs. The chapter further elaborates on the practical applications of these technologies in breast cancer classification, with particular interest in how improved diagnostic accuracy could be brought through the use of Transformers and CNNs. We end by discussing transfer learning, which shows how pretrained models can be adapted to new tasks. Lastly, we will demonstrate how this shows that the landscape in ML is dynamic and constantly changing as it takes on these harder real-world problems.

3.1 Transformers

3.1.1 Overview

Transformers are a breakthrough in neural network architectures: Vaswani et al. introduced them in 2017 [54]. Unlike conventional sequence transduction models that rely on Recurrent Neural Networks (RNNs) (or even CNNs), the Transformer uses self-attention mechanisms to process the input data in parallel, which makes their approach computationally more efficient, with simultaneous improvements in performance. The architecture of the Transformer is shown in Figure 3.3.

3.1.2 History and Development

One of the driving factors in the development of transformers is that RNNs and CNNs could not handle long-range dependencies or parallelization. Transformers address such issues by using self-attention to capture global dependencies within sequences and, for example, achieve quicker training and better overall performance on tasks such as machine translation [54].

3.1.3 Self-Attention Mechanism

Attention Score Calculation

The self-attention mechanism assigns weights to different input tokens based on their importance. It uses attention scores computed by taking the dot product of queries and keys, then scaling the result and applying a softmax function (this is explained more in detail in section 3.2.5) to obtain the attention weights across all positions. This process is depicted in Figure 3.1.

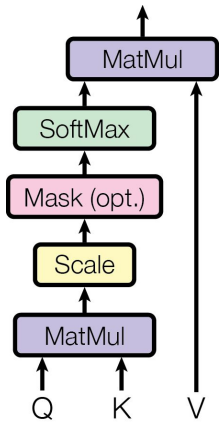

Figure 3.1: Diagram of scaled dot-product attention showing the flow from query (Q), key (K), and value (V) through matrix multiplication, scaling, optional masking, softmax, and final matrix multiplication [54].

Scaled Dot-Product Attention

In scaled dot-product attention, the dot product of the query and key vectors is divided by the square root of the dimension of the keys. This handles the problem of large values, which may not be compatible with training under the vanishing gradients hypothesis [54]. It it represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q is the matrix of queries, K is the matrix of keys, and V is the matrix of values.

The query vectors (Q) represent the elements seeking information, the key vectors (K) are the elements containing the information, and the value vectors (V) are the actual information being retrieved.

Multi-Head Attention

Multi-headed attention enables a model to focus on different input parts with independence in multiple heads. Each head independently applies the scaled dot-product

attention, and then the results are concatenated and linearly transformed into the overall output of attention [54]. Basically, this mechanism comprises many attention layers, which interact in parallel with a set of given inputs. The equation for multi-head attention is:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O,$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$.

This is depicted in Figure 3.2.

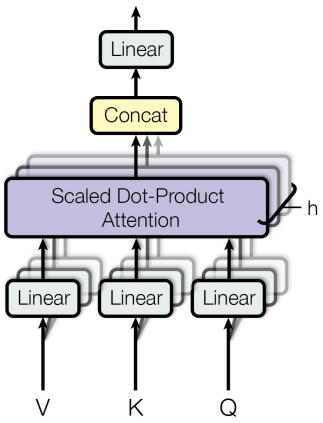

Figure 3.2: Diagram of multi-head attention showing the flow from query (Q), key (K), and value (V) through linear layers, multiple scaled dot-product attention mechanisms, concatenation, and a final linear layer [54].

3.1.4 Positional Encoding

Importance of Positional Information

Transformers, by design, do not have the inherent sequential property of RNN. Positional encodings are added into the embedding of the inputs to provide the model with information about its position relative to the other tokens [54].

Types of Positional Encoding

The positional encoding in Transformers is represented by the sine and cosine functions of different frequencies, allowing the model to generalize relative positions from longer sequences than those observed during training [54]. The equations are:

$$\text{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right),$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right),$$

where pos represents the position of the token in the sequence, and d_{model} is the dimensionality of the model's embeddings.

3.1.5 Encoder-Decoder Architecture

Encoder Components

The Transformer encoder is comprised of a stack of the exact same layers, where each layer has two sub-layers, that is, multiheaded self-attention and position-wise fully connected feed-forward network. In both of these sub-layers, residual connections and layer normalization are applied [54]. The overall architecture is depicted in Figure 3.3.

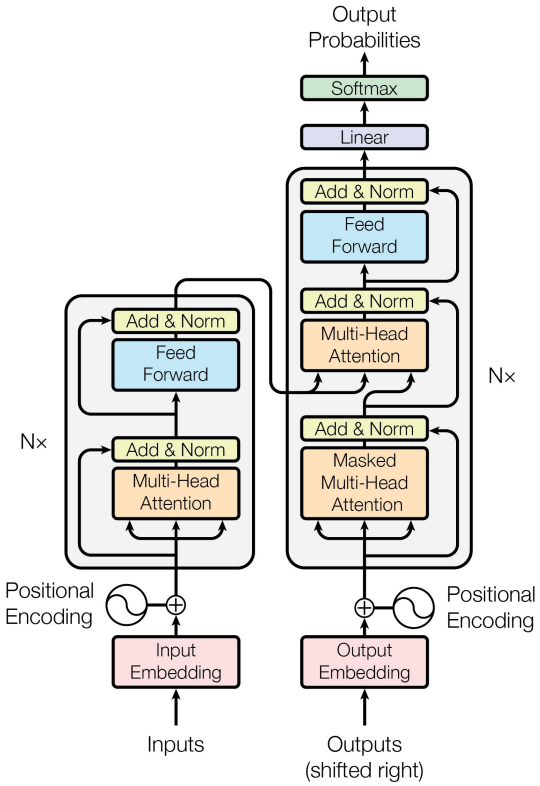

Figure 3.3: Diagram of the Transformer model architecture, showing the input embedding, positional encoding, multi-head attention, feed-forward layers, and output probabilities [54].

Decoder Components

The decoder also consists of a stack of identical layers, but an additional third sub-layer enables it to pay attention to the relevant parts of the input sequence. This multi-

head attention over the encoder's outputs allows the decoder to focus only on the critical regions of the input sequence [54].

Role of Attention in Encoding and Decoding

The attention mechanisms in the encoder enable each position of the input sequence to look at all other positions, capturing the dependencies between them. Attention mechanisms in the decoder enable it to produce output sequences from relevant parts of both the input and the already generated part of the output [54].

3.1.6 Transformer Architecture

Input Representation

The input tokens are first embedded into vectors and then combined with positional encodings. Then, the embeddings are processed using the encoder [54].

Layer Normalization

After adding residual connections inductively, layer normalization stabilizes and accelerates training. It normalizes the output of each sub-layer to have zero mean and unit variance [54].

Residual Connections

Residual connections are put around each sub-layer to assist the gradient flow, allowing the training of deeper networks. This aids in enabling the bypassing of these nonlinear transformations within the sub-layers so that the gradient can avoid the vanishing gradient problem [54].

Feed-Forward Neural Networks

Each layer in the encoder and decoder has a feed-forward neural network that contains two linear operations with a ReLU[1] activation applied in between. It thus enables the model to learn complex representations [54].

3.1.7 Training Transformers

Batch Size

Batch size refers to the number of samples that are going to be processed before the internal parameters of the model get updated. It controls the stability and efficiency of the learning process. Increasing batch sizes will stabilize the training procedure through an

[1]ReLU (Rectified Linear Unit) activation is a widely used function in neural networks, defined as $f(x) = max(0, x)$. It introduces non-linearity by outputting zero for negative inputs and the input value for positive inputs, which helps the model learn complex patterns and accelerates training by mitigating the vanishing gradient problem.

averaging effect on the noise in gradient updates, leading to more reliable convergence. However, this will also require more memory and computing power. Smaller batch sizes introduce more noise in the gradient updates, which might help escape local minima; however, this can lead to less stable training dynamics. The selection of batch size should strike a balance between computation resource limitations and the need for stable and efficient training, which will be analyzed further in 5.3 [55].

Loss Functions

The Transformer uses cross-entropy loss (explained in 3.3) to measure the difference between the predicted and actual output sequences. This loss function is particularly well suited for tasks like sequence transduction [54].

Optimization Techniques

We train the Transformer model using an Adam optimizer with linear warm-up scheduling on the learning rate[2] and then followed by a schedule that decreases proportionally to the inverse square root of the number of steps [54]. The learning rate is given by:

$$\text{lrate} = d_{model}^{-0.5} \cdot \min(\text{step\_num}^{-0.5}, \text{step\_num} \cdot \text{warmup\_steps}^{-1.5})$$

Regularization Strategies

To prevent overfitting[3], the Transformer employs dropouts[4] in the outputs of each sub-layer and also uses label smoothing in training. These techniques improve the generalization ability of the model [54].

Scheduler Techniques

Learning rate schedulers are very important in the training process because they adjust the learning rate to give better performance.

Some of the schedulers used in the literature include StepLR and ReduceLROnPlateau. StepLR reduces the learning rate by a factor after a few epochs, ensuring gradual fine-tuning of the model with progression along the training [58]. On the other hand, ReduceLROnPlateau is a scheduling method that dynamically tunes the learning rates based on how the model is performing: if a plateau[5] in the validation loss is noted, the learning

[2]The learning rate in ML is a hyperparameter that controls the step size during the optimization process. It determines how much the model's weights are adjusted with respect to the loss gradient. A high learning rate can lead to faster convergence but might overshoot the optimal solution, while a low learning rate ensures more precise convergence but may be slower [56].

[3]Overfitting occurs when a ML model learns the details and noise in the training data to an extent that it negatively impacts the performance of the model on new data. This results in a model that performs well on the training data but poorly on unseen data [57].

[4]Dropout involves randomly setting a fraction of the output units to zero during training, which forces the network to learn more robust features and reduces reliance on specific neurons.

[5]In this context, "plateau" refers to a stage where there is no significant change in the validation loss, indicating that the model's performance is no longer improving.

rate is decreased, allowing for dynamic adjustment [59].

Such schedulers optimize the training process, ensure efficient learning, and prevent the model from getting stuck in local minima.

3.1.8 Applications of Transformers

Natural Language Processing (NLP)

Transformers have set state-of-the-art scores across a wide range of NLP benchmarks, including machine translation, text summarization, and question answering, because they can capture long-range dependencies and process sequences in parallel [54].

Text Summarization

Transformers have great importance when used in text summarization to create short summaries of long documents. The self-attention mechanism allows models to focus on the most important parts of texts and provide quality summaries [54].

Question Answering

In question answering tasks, Transformers are good at pulling context and can, therefore, retrieve the correct answers from large bodies of text. Their performance regarding such applications is mainly due to the capability of modeling dependencies between words that are far apart [54].

3.2 Visual Transformers

3.2.1 Overview

ViTs have risen as a powerful alternative to CNNs for image recognition tasks. The ViT model by Dosovitskiy et al. (2021) [52] directly applies the Transformer architecture to sequences of image patches and shows that a pure Transformer can achieve state-of-the-art results in image classification.

3.2.2 Comparison with CNNs

Convolutional Neural Networks

CNNs provide the basic building block for visual recognition within a DL framework. Through repetitive patterns, CNNs extract features at various levels by applying the convolution operation to a given image.

Convolutional Networks (ConvNets) are designed for image processing. In contrast to the traditional neural networks, in which the weights are scalar connections between neurons, in ConvNets, weights are like filters between convolution operations. Neurons

in hidden layers are not connected with all preceding neurons but they form 3D volumes characterized by width, height, and depth. Additionally, the layers in ConvNets are composed of different sub-layers, including the Convolution Layers, Pooling Layers, and Fully Connected Layers.

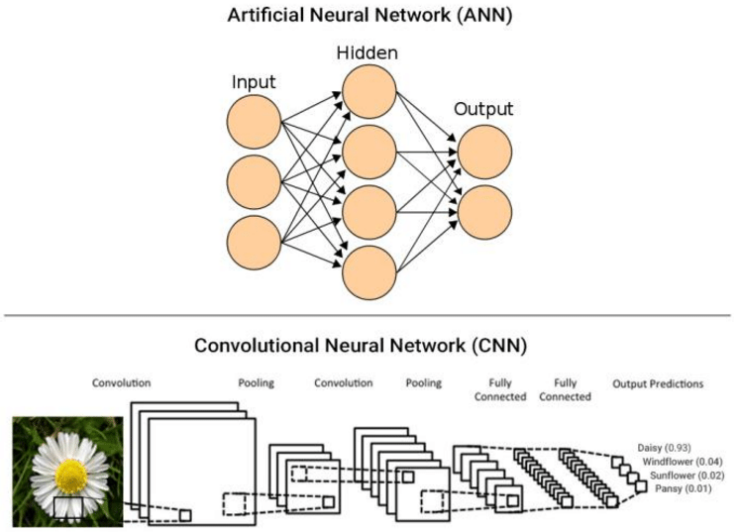

Figure 3.4: Comparison between an Artificial Neural Network (top) and a CNN (bottom), illustrating their different structures and processes for image classification [60].

In simple words, convolution is used to extract features from images. For this, a small window, known as a kernel, moves across an image to find lines, shapes, objects, etc. The kernel (i.e., a matrix) is actually multiplying the pixel values for producing an output image where certain features are highlighted.

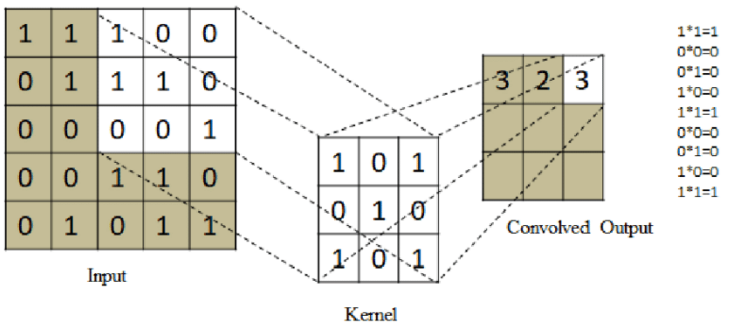

Figure 3.5: Illustration of the convolution process, showing the application of a kernel to an input image to produce a convolved output [61].

Figure 3.6 shows the feature maps' evolution for different CNN layers, which can be adapted to the case of classifying breast cancer.

- **Low-Level Feature Maps**: In the initial layers, the network captures basic features such as edges and orientations.
- **Mid-Level Feature Maps**: In the middle layers, the network extracts more complex patterns and textures from the input data.
- **High-Level Feature Maps**: In the top layers, the network identifies parts of objects and their overall structure.
- **Trainable Classifier**: These high-level features are then fed into a classifier to make the final prediction.

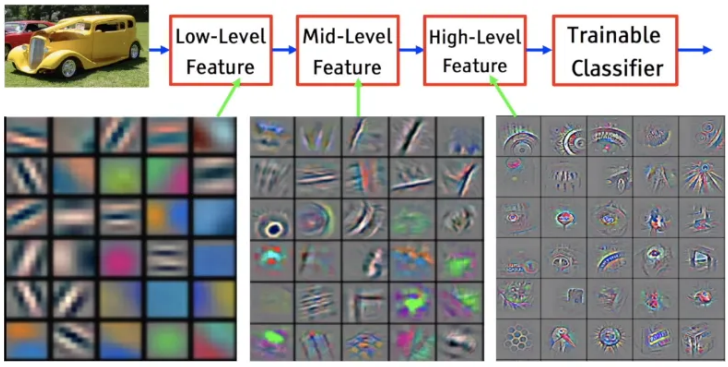

Figure 3.6: Visualization of how a CNN processes an image through different levels of feature extraction, from low-level features to a trainable classifier [62].

CNNs work effectively in capturing spatial hierarchies in an image due to its ability to perform convolution and pooling operations. Hence they are very effective in image classification tasks.

However, ViTs apply Transformer's architecture (depicted in Figure 3.7), over sequences of image patches; they are hence not tied to spatial hierarchies, as is the case with CNNs. CNNs focus on local spatial features via their convolution, while Transformers capture long-range dependencies via self-attention mechanisms.

While CNNs use convolutional layers to extract features hierarchically, ViTs break an image into fixed-size patches, linearly embed them, and then process them using a standard Transformer encoder [52].

ViTs require large datasets and huge computational resources for training to perform better than CNNs and achieve state-of-the-art performance. Still, they can potentially outperform CNNs in accuracy because of their capability to model global relationships in data [52].

Transformers are more versatile over various data types, making them suitable for a huge number of different kinds of tasks, not only image classification.

The use of both ViTs and CNNs in breast cancer classification leverages the strengths of the two types of models. CNNs became well-established in medical imaging due to their built-in ability to capture local features through convolutional layers. This helped in detecting patterns in mammograms. Conversely, self-attention mechanisms are used in ViTs to efficiently capture the global context and long-range dependencies for feature representation in images. Some studies have identified that ViTs outperform CNNs in some tasks because of their ability to deal with much more complex features [52, 53].

Hence, for this project on mammography image classification, the use of ViTs was considered to take advantage of their state-of-the-art handling of complex visual data and thus providing better overall performance than traditional CNNs.

3.2.3 Vision Transformer (ViT) Architecture

Input Representation

In ViT, an image is split into a sequence of fixed-size patches, linearly embedded after flattening them individually. These patch embeddings are then combined with positional encodings to retain spatial information [52].

Patch Embedding

Each image is first separated into non-overlapping patches with dimensions of 16×16 pixels. The patches are then flattened and projected into a lower-dimensional space using a linear layer. The resulting embeddings of the patches can then be used as input tokens for the Transformer model [52].

Positional Embedding

Learnable positional embeddings are then added to the patch embeddings while still ensuring that spatial information from the patches is retained, which is meant to ensure that the model can infer relative patch positions in the image [52].

3.2.4 Self-Attention in Visual Transformers

Application of Self-Attention to Image Data

The ViT model is endowed with self-attention under this mechanism, making different patches have appropriate weights, grasping long-range dependencies and global context. This is of paramount importance in tasks where spatial relationships are essential [52]. Figure 3.8 illustrates the self-attention mechanism in ViT.

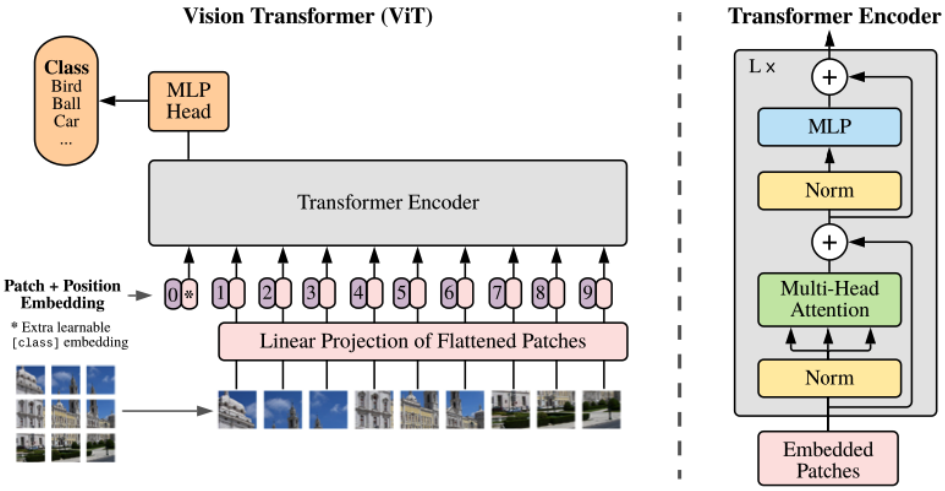

Figure 3.7: Overview of the ViT model, showing the process from patch embedding to transformer encoding and the final classification [52].

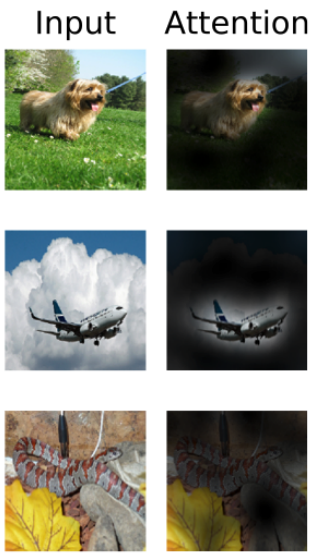

Figure 3.8: Visualization of the self-attention mechanism in ViTs, showing the input images and their corresponding attention maps [52].

Benefits and Challenges

ViTs have the advantage of making global-context models without an inductive bias similar to those made by CNNs. However, they need large-scale pretrained datasets to perform well and can be computationally intensive due to the self-attention mechanism's quadratic complexity [52].

3.2.5 Training Visual Transformers

Loss Functions for Visual Data

Generally, a cross-entropy loss is used for image classification tasks in ViTs. The choice of the loss function is also more or less equivalent to what is used in training CNNs, wherein the stress measures the slightest difference between a predicted and a real class label [52]. One of the most critical metrics that need to be observed during the training of models is the loss, which measures the extent of prediction error and quantifies how well a model is performing. Thereby, the output of a specific function can be choosen as this loss function.

Softmax Function

This project aims to predict whether a patient with abnormal screenings is most likely to suffer from cancer. These normalized class probabilities are obtained from the softmax function, which extends a binary sigmoid function to the multiple classes. In this section, we briefly explain the general principle and motivation behind the softmax function.

First, we need to define a loss function that computes a measure of classification error. The softmax classifier uses a generalization of the logistic function as its loss function, extended to handle multiple classes. This should not be confused with the sigmoid function, which is the binary equivalent of the logistic function.

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{3.1}$$

In this context, x represents the output from the linear function

$$f(x_i; W) = Wx_i + b, \tag{3.2}$$

which uses W as the weights, x_i as the input value of each neuron, and b as the bias. This function provides a probabilistic interpretation and assigns the probability to correctly or incorrectly classify an image, where the total of these probabilities equals one.

Cross-entropy Loss

Cross-entropy loss, often popularly known as log loss, is a measurement of how well a classification model that produces a value from 0 to 1 representing a probability actually matches a categorical label. This loss increases proportionately to the deviation of the predicted probability from the actual label. For instance, it would give a significant loss in the case of a prediction of 0.012 when the actual label has to be 1, meaning poor model performance. A perfect model, on the other hand, would have a log loss of 0, evidencing its perfect accuracy in all its predictions [63].

The equation for Cross-entropy or log loss, where y is the true label and $\hat{y}$ is the predicted probability class, is given by:

$$\mathcal{L}_{CE}(y, \hat{y}) = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]. \tag{3.3}$$

Using the sigmoid function, the predicted class probability $\hat{y}$ can be expressed as:

$$\hat{y} = \sigma(w \cdot x + b) \tag{3.4}$$

Now, the cross-entropy loss in terms of the weights w, bias b, and input x is written as:

$$\mathcal{L}_{CE}(w, b) = -\left[y \log(\sigma(w \cdot x + b)) + (1 - y) \log(1 - \sigma(w \cdot x + b))\right]. \tag{3.5}$$

Another prevalent issue, which our dataset also has and which is familiar in many ML problems, is class imbalance. We discuss more on that in 5.5.1; for now, this imbalance is commonly handled by modifying the standard cross-entropy loss and introducing weights to each class. This approach enhances sensitivity towards underrepresented classes during training of models.

Optimization Techniques

Effective techniques of optimization are required to optimize the training of ViTs. One of the most popular optimizers, AdamW, was developed to solve the shortcomings of the original Adam optimizer, particularly weight decay. AdamW decouples weight decay from the gradient update, and it offers much better regularization and, consequently, improvement in generalization performance [64]. The adaptive learning rates and momentum of AdamW have made it suitable for large-scale models like ViTs.

AdamW updates the parameters with respect to the gradients from the loss function but in an adaptive manner, lowering the learning rate of different parameters differently. Such an adaptive approach helps these models converge faster and more stably, especially for complex tasks such as image recognition.

Another optimization technique commonly used in training ViTs, especially at fine-tuning stages, is Stochastic Gradient Descent (SGD) with momentum. SGD updates the parameters using only a randomly selected subset, or mini-batch, of data at each iteration. It helps to lessen the computational demands and introduces beneficial stochasticity for avoiding local minimums and exploring through the parameter space more thoroughly [65]. Momentum in SGD helps accelerate the gradient vectors, thus promoting faster convergence and smoother optimization trajectories.

Some of the vital optimization techniques in training ViTs include AdamW and SGD with momentum. While AdamW has shined because of its adaptive learning rates and regularization benefits, SGD with momentum is still computationally efficient and allows better parameter space exploration. Both techniques will be further compared during fine-tuning of the hyperparameters used in the training of the model in section 5.3.

Data Augmentation Strategies

Effective data augmentation practices are vital for training ViTs. These include random cropping, horizontal flipping, and color jittering to boost the generalization capacity of the model [52]. All such strategies are carried out in section 5.2.

3.2.6 Applications of Visual Transformers

Image Classification

ViTs have shown state-of-the-art performance on image classifiers over a list of benchmarks, including ImageNet[6]. Global context modeling aids Vit in achieving higher accuracy and outperforming traditional CNNs [52].

Object Detection

Although ViTs were developed for classification tasks, they are generalized for solving object detection. By utilizing the self-attention module, ViTs can easily detect and localize objects within an image [67].

Image Segmentation

In the same line of work, other projects have applied these models to image segmentation tasks and further proved the power of ViTs in partitioning an image into meaningful segments, thanks to its global context modeling which aids in accurate complex scene segmentation [67].

3.3 Transfer Learning

Transfer learning (TL), in a broad sense, is a ML method where a model developed for a task is reused as the starting point for a model on a second (different) task. This becomes very useful when the dataset for a new task is too small to train full models without overfitting [68]. TL is a method that enables solving new problems more effectively based on knowledge obtained previously, as it is depicted in Figure 3.9.

3.3.1 Feature Extraction

In feature extraction, it is the use of a pretrained model what pulls valuable features out of a new dataset. The basic idea is just to use the lower layers of the network but fine-tune them accordingly. Those low-level features, such as edges and textures, that are general and thus transferable across different tasks or datasets, have to be adapted without changing all the other layers (i.e., freezing them) [68].

To make it more concrete, convolutional layers in CNNs pretrained on an extensive dataset like ImageNet can be considered as feature extractors with fixed feature parameters on the targeted dataset. The features extracted through it are further fed to a new classifier, commonly some fully connected layers, to adapt to the specific task [68].

[6]The ImageNet dataset is a large visual database designed for being used in visual object recognition research. It contains millions of annotated images belonging to thousands of categories [66].

When pretrained on large datasets, ViTs can act as robust feature extractors and support the transfer of learned representations to new datasets. The process involves extracting generalized features with the use of lower layers of the ViT and fine-tuning the upper layers specific to the task for which it has been developed [69].

3.3.2 Fine-Tuning

Fine-tuning means making a slight adjustment to a pre-trained model by further training it on the new dataset. The process fine-tunes the network weights to be closer to new data and yet maintain the features learned from the original dataset [68]. Fine-tuning generally focuses on the last few layers of a pre-trained model, whereas earlier layers are left frozen.

When fine-tuning, it is crucial to:

- **Freeze some layers**: Often, initial layers capturing general features are frozen, and only the final layers are retrained.
- **Adjust the learning rate**: A lower learning rate is used to prevent drastic changes to the pretrained weights, allowing the model to adapt smoothly to the new data [68].

Importance of Transfer Learning

TL is important, especially when it is not possible to collect and annotate a large dataset because of the resources available (e.g., time or costs). This makes it possible to leverage models already pretrained on big datasets to perform well on tasks for which only few data might be available [68].

Practical Considerations

TL works best when the source and target datasets are somewhat similar. For instance, transferring weights from a model pretrained on ImageNet to a medical imaging dataset might require further fine-tuning because data is different [68].

It is important to ensure that the architecture of the pretrained model is compatible with the new task. Changes may be required if the number of layers or input dimensions do not match [68].

3.4 CNNs and ViTs for Breast Cancer Classification

Breast cancer is one of the leading causes of cancer deaths in women worldwide. Stressing the importance of early detection plays a central role, and different imaging modalities such as mammography and ultrasound work as a crucial source to detect them. Both CNNs and ViTs have been extensively investigated for their potential to enhance breast cancer classification accuracy in recent years.

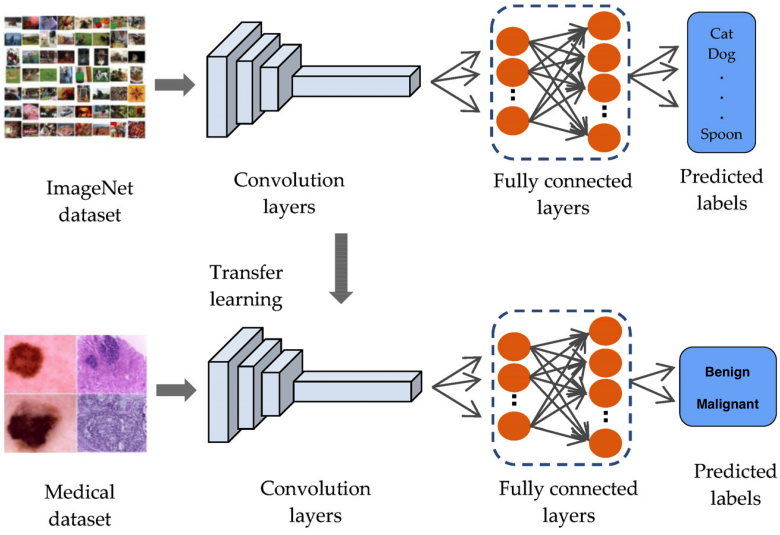

Figure 3.9: Illustration of transfer learning process: pre-training a CNN on the ImageNet dataset and then fine-tuning it on a medical dataset to predict labels such as benign or malignant [70].

3.4.1 CNNs for Breast Cancer Classification

CNNs achieved enormous success in the analysis of medical images, beginning with breast cancer detection and classification and extending to other illnesses. CNNs are preferred specifically for their automatic feature extraction capabilities that are hierarchical in nature.

Performance of CNN Models

Several authors have reported the efficiency of specific CNN architectures in breast cancer classification tasks. For example, Boukaache et al. in 2023 [71] evaluated breast ultrasound images with ResNet18, ResNet50, and VGG16, and all three provided high accuracy rates. Their study found that ResNet18, based on small datasets, reached a top value of 99.98% accuracy, while further data augmentation increased the outcomes for VGG16 and ResNet50.

In another study, Thirumalaisamy et al. also in 2023 [72], proposed an enhanced optimization algorithm and synthesized CNN model for diagnosing breast cancer. The researchers presented a comprehensive classification technique which applied the methodology of Enhanced Ant Colony Optimization (EACO) to find the best hyperparameter values in designing the architecture of the CNN. The accuracy of the proposed EACO-ResNet101 model was found to be 98.63% on the CBIS-DDSM dataset, and 99.15% on the MIAS dataset, demonstrating its great performance compared to conventional methodologies.

Challenges and Limitations

While CNNs perform well, they have several difficulties in the case of breast cancer classification. One fundamental limitation is using many convolutional layers, leading to high computational complexity. Besides, very often, solving problems like dataset imbalance and low image quality requires very sophisticated data preprocessing and data augmentation for CNNs [71].

3.4.2 ViTs for Breast Cancer Classification

Recently, ViTs have emerged as a potential alternative to CNNs, leveraging the self-attention mechanism for modeling global dependencies within the images and its capability to capture intricate patterns. It offers possibly even more significant advantages in the context of mammography, where minimal differences in tissue appearance can be crucial for making a diagnosis. This section focuses on the applications of ViTs in breast cancer classification, specifically in mammography and ultrasound imaging.

Application and Performance of ViTs

Gheflati and Rivaz [73] investigated the potential for using ViTs to classify breast ultrasound images. The performance obtained showed that ViTs can get equal or even better performance than state-of-the-art CNNs, with ViTs reaching over 85% accuracy in classification, and getting average AUCs[7] close to 0.95.

Ayana et al. (2023) [53] developed transfer learning in mammogram classification using a ViT-based method. Their method outperforms traditional CNN-based models, with an AUC-ROC[8] reaching 1, thus illustrating the potential of ViTs in enhancing diagnostic accuracy in clinical scenarios.

Advantages and Challenges of ViTs

These algorithms provide several advantages over CNNs, including reduced inductive bias and the ability to model long dependencies, but they come at the cost of needing large-scale datasets for training. Precisely, these can be a limiting factor in medical image data, where annotated data is often not abundant [53]. What's more, ViTs are computationally demanding and require efficient implementation strategies.

[7] AUC (Area Under the Curve) measures the performance of a binary classification model. It ranges from 0 to 1, with 1 indicating perfect classification and 0.5 indicating random guessing. Higher AUC values mean better model performance.

[8] ROC (Receiver Operating Characteristic) is a graphical plot that illustrates a binary classification model's performance. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The ROC curve helps evaluate the model's ability to distinguish between positive and negative classes.

3.4.3 Other Methods for Mammography Classification

Besides CNNs and ViTs, other advanced methods have been proposed for mammography image classification. One example is using diffusion models, such as MedSegDiff [74], which uses the denoising process to derive high-quality segmentation, which will further improve the detection of cancerous tissues in mammograms. This kind of architecture is robust enough to handle noise and variability present in medical images [75].

Generative Adversarial Networks (GANs) have also found applications in improving training data and classification performance. GANs can generate realistic, synthetic images of mammograms. In this manner, they aid in the training of a more accurate and robust model for a better classification. More recently, Yi et al. (2019) [76] reported that GANs have shown great potential, in particular, for breast cancer classification in medical imaging. Thus, GANs can be used to create high-quality synthetic mammograms to overcome the pressing issue of scarce annotated data and improve the performance of the classification models. Therefore, synthetic images will diversify the database for training, leading to higher generalization and better detection of cancerous tissues in mammograms.

3.4.4 Comparative Analysis

Both CNNs and ViTs have proved to be highly promising toward the classification of breast cancer with unique advantages and disadvantages, showing that DL models will improve diagnostic accuracy. Although CNNs excel in tasks for which the data is abundant and local features are well defined, ViTs possess unchallenged predominance in capturing global context and dealing with complex patterns. The preference for these predominantly depends on the peculiar requirement of the task and the available resources.

Chapter 4

Data, Methodology, and Tools

In this chapter, we will properly introduce the basis and background of our research, activity data, methodology, and tools in use. We begin with a summary comparison of different clinical implementations of mammography datasets based on their applications in lesion segmentation, detection, and classification functions. This will then be followed by an in-depth look of the OPTIMAM dataset, a very critical resource in this thesis, detailing its composition and structure and the preprocessing techniques applied in its preparation for further analysis. We also explain how to ensure reproducibility and avoid the possibility of data leakage while splitting the dataset into training, validation, and test sets correctly. Finally, we outline the preprocessing steps, data loading procedures, and the evaluation metrics used to assess the models to set up all the necessary and corresponding stages for a thorough analysis and robust conclusion in the next chapter.

4.1 Comparison of Clinical Use Cases of Public Mammography Datasets

Figure 4.1 provides a glimpse of the scenarios in which public mammography datasets could apply based on the respective applications concerning breast lesion segmentation and detection, background tissue assessment, and whole mammogram classification. These datasets can still be segmented further based on the imaging modality applied: film screen, or fully digital.

The diagram (4.1) underscores the heterogeneity in the purpose of publicly available datasets and emphasizes how each dataset can be engaged successfully for different clinical applications. For example, the OPTIMAM dataset [7] which is further analyzed, contains all the essential information to be successfully used for segmenting and detecting lesions with the use of digital mammography images. This makes it very suitable for developing strong AI models that are targeted at classifying mammograms into malignant and benign categories.

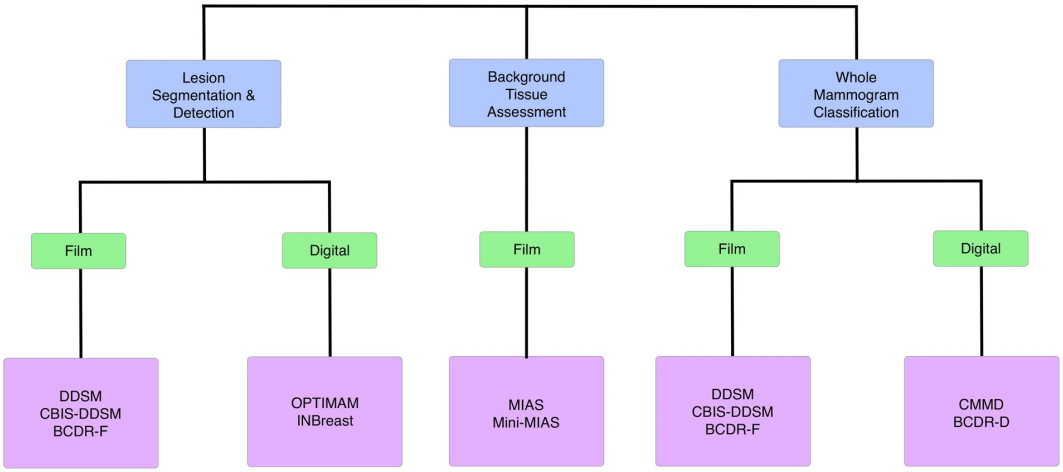

Figure 4.1: Comparison of clinical use cases of mammography datasets, categorizing datasets based on lesion segmentation and detection, background tissue assessment, and whole mammogram classification for both film and digital formats [77].

4.2 OPTIMAM Dataset

The OPTIMAM Mammography Image Database (OMI-DB) is a large dataset developed to support the development and evaluation of CAD systems and other ML models in mammography. It incorporates a collection of over 2.5 million images from 173,319 women over a decade through a pooling of various NHS breast screening centers scattered across the UK [7]. This includes images of normal breasts, benign findings, screen-detected cancers (malignant), and interval cancers[1]. This wide-ranging collection is maintained with regular updates, so that it remains a vital resource in any research quest [7].

Cancer Research UK funds the OPTIMAM study and underpins a large amount of work on detecting and diagnosing breast cancer [78]. Images are annotated at a very high quality by experienced radiologists, forming valuable data both for training AI models and the evaluation of these models. Shared under agreed-upon restrictions in use with academic and commercial research groups, this further allows for widespread research and innovation in mammography [7].

Fully annotated centralized data assists not only in developing AI algorithms but also in a myriad of research studies aimed at the understanding and optimizing processes for screening and diagnosis of breast cancer. Continuous data collection and updating clinical

[1]In mammography, **normal** refers to breasts with no signs of cancer or other abnormalities. **Benign** findings are non-cancerous abnormalities. **Screen-detected cancers** (**malignant**) are cancers identified during routine screening mammograms. **Interval cancers** are cancers that are diagnosed between regular screening intervals, typically due to symptoms or clinical findings.

data from the already established and newly built screening sites will ensure the dataset is relevant and comprehensive enough for all future research needs [7].

4.2.1 Dataset Overview

Together with the images from the OPTIMAM dataset is an associated CSV [2] file that provides detailed information to identify and classify the images for each patient (i.e., client). This file includes a variety of columns, each with specific information relevant to the patient's mammographic screening. Table 4.1 describes each column in the CSV file.

Table 4.1: Description of columns in the OPTIMAM dataset CSV, detailing the information for each patient, image, lesion, ROI, and manufacturer.

Column	Description
client_id	Unique identifier for each patient.
status	Indicates whether the lesion is benign or malignant.
study_id	Identifier for the study.
serie_id	Identifier for the series of images.
image_id	Unique identifier for each image.
view	The view of the mammographic image (e.g., CC, MLO).
laterality	Side of the body (left or right) where the image was taken.
age	Age of the patient.
mark_id	Identifier for the mark indicating a lesion (if present).
lesion_id	Identifier for the lesion (if present).
conspicuity	Visibility or prominence of the lesion.
x1	X-coordinate of the top-left corner of the lesion bounding box.
x2	X-coordinate of the bottom-right corner of the lesion bounding box.
y1	Y-coordinate of the top-left corner of the lesion bounding box.
y2	Y-coordinate of the bottom-right corner of the lesion bounding box.
pathologies	Type of pathology identified in the image.
manufacturer	Manufacturer of the imaging device.
pixel_spacing	Pixel spacing information of the image.
magnification_factor	Magnification factor used in the image.
implant	Indicates the presence of implants.
xmin_cropped	X-coordinate of the top-left corner of the cropped image.
xmax_cropped	X-coordinate of the bottom-right corner of the cropped image.
ymin_cropped	Y-coordinate of the top-left corner of the cropped image.
ymax_cropped	Y-coordinate of the bottom-right corner of the cropped image.

[2] CSV (Comma-Separated Values) is a simple file format used to store tabular data. Each line in a CSV file represents a row, with values separated by commas. CSV files are widely used for data exchange between different applications due to their simplicity and ease of use.

This level of detailed information allows for a comprehensive analysis and classification of mammographic images, aiding the setting up of robust models capable of discriminating between benign and malignant lesions.

For this thesis, a subset of the OPTIMAM dataset containing images from 6,000 unique clients was used. From this dataset, we had 47,542 images distributed among different manufacturers. The number of episodes each manufacturer contributed to is shown in the following bar chart 4.2.

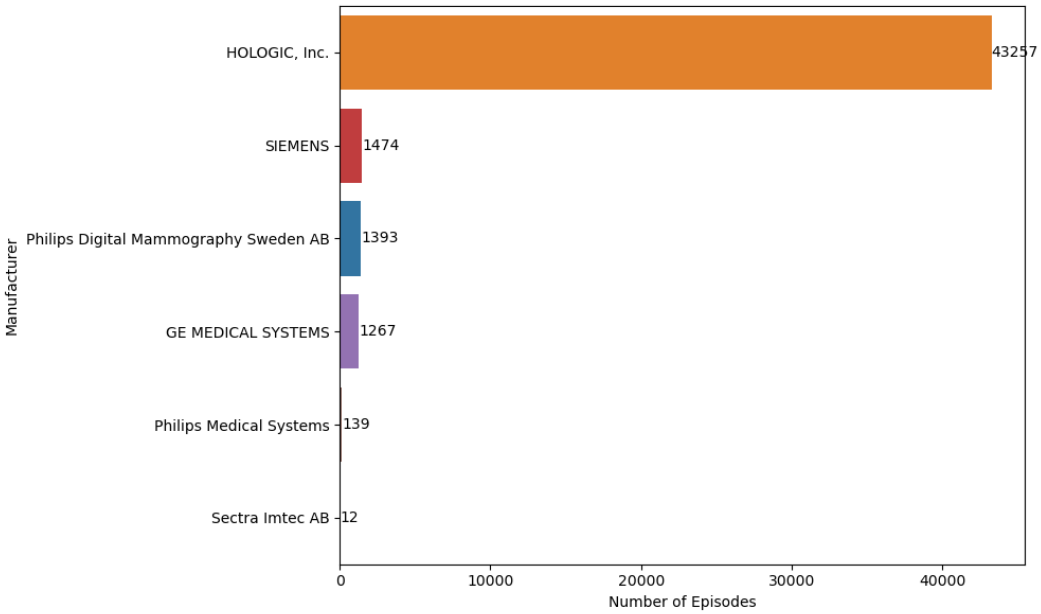

Figure 4.2: Bar chart showing the distribution of mammography images by manufacturer, with counts of episodes for each manufacturer.

4.2.2 Analysis of Manufacturer Distribution

As Figure 4.2 shows, the spread of the various mammography manufacturers in the dataset is highly uneven. The majority of the images are from HOLOGIC, Inc. [79], which makes up for 43,257 episodes. This is higher compared to the rest of the manufacturers.

This brings several issues within the dominance of the dataset by a single manufacturer:

- **Potential Bias:** The significant skew in numbers toward images from HOLOGIC, Inc. might introduce a bias into the analysis because the dataset would not be able to fully represent the variability and features of images from other manufacturers.
- **Device-Specific Features:** Manufacturers use different technologies and protocols, leading to differences in the appearance and quality of images. This concentration

of data from one manufacturer might limit the generalizability of the findings to images from other devices.

- **Analysis Robustness:** Any developed models or diagnostic tools on this dataset may work well on images from HOLOGIC, Inc., yet they could face issues when using other manufacturers' datasets owing to differences in image characteristics.

The evidence on manufacturer distribution considerations is taken into the interpretation of results and building models so that the derived findings are robust and generalizable across various imaging devices and manufacturers.

An example image from each manufacturer is illustrated in Figure 4.3 to provide a visual representation. Each image shows the typical quality and characteristic of mammogram images delivered by the respective manufacturer and illustrates the diversity within the dataset.

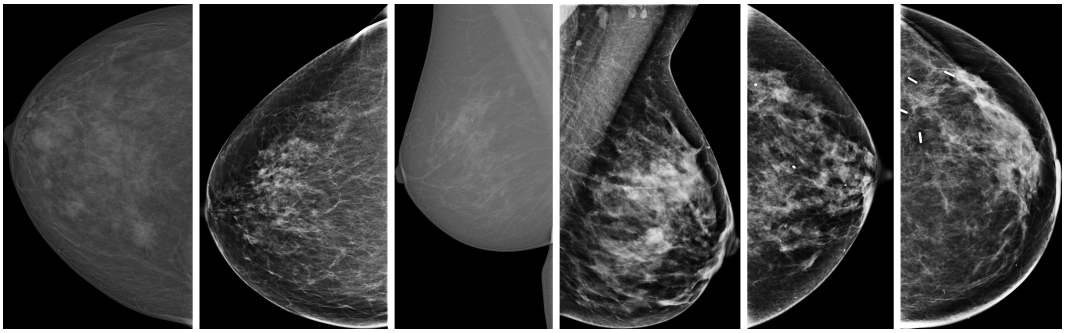

Figure 4.3: Sample mammography images from the OPTIMAM dataset. From left to right: Philips Digital Mammography Sweden AB, HOLOGIC, Inc., Sectra Imtec AB, SIEMENS, GE MEDICAL SYSTEMS, Philips Medical Systems.

The histogram of Figure 4.4 shows that most of the subjects fit in the age range of 50 to 70, centered around the 55 to 60 age category. This is a typical pattern for the screening population, as mammography screening programs are conventionally aimed at this age bracket for women, to ensure early detection of breast cancer when it is less advanced and hence more treatable. It is also targeted for this age group due to the density of the breast tissue. Women's breasts are denser when they are younger, so abnormalities may be more complex to pinpoint on a mammogram. As women become older, breast tissue will, in general, become less dense, and therefore, mammography is more effective for screening. The distribution tapers off for both younger and older age groups, which aligns with standard screening practices focusing on middle-aged to older women who are at a higher risk for breast cancer.

The plot in Figure 4.5 shows that the highest number of lesions is classified as malignant, with a total of 3,968 cases. The next most significant is the category "Normal", with 1,030 and then 970 benign lesions. There is also a class for 'Interval Cancer', but

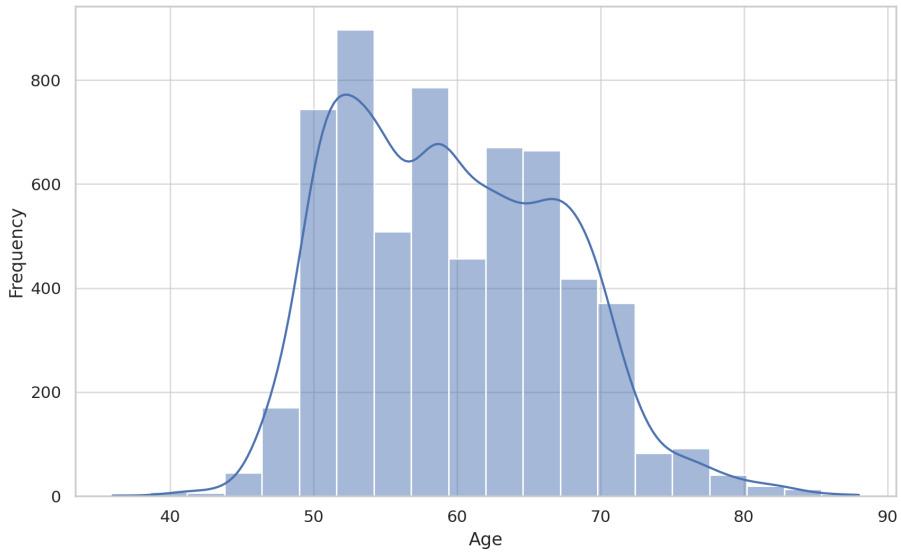

Figure 4.4: Histogram showing the age distribution of patients in the OPTIMAM dataset.

only with 32 cases. That suggests the data set is skewed toward malignant cases, which will be valuable in training strong classification models. However, this feature is not very normal because in real-life screening programs there are many more benign cases than malignant ones, and the model can become biased if proper attention is not given to this issue [80, 81]. Normal and benign cases exist in a balanced dataset, ensuring that the model can really distinguish between various kinds of lesions. However, in this thesis, we will concentrate on distinguishing between malignant and benign mammography images.

An overview and analysis of the dataset presented in the CSV file is presented in Table 4.2, showing the distribution of annotated masses by mass status (benign or malignant) and patient age group (<50, 50–60, 60–70, >70) across each manufacturer. Percentages are computed from the total number of masses per manufacturer.

4.2.3 Image format and data

Each mammography image in the OMI-DB is stored in PNG format. This format is widely used due to its lossless compression; that is, no data loss occurs in the image while converting from the original format. However, this feature of the PNG format does not inherently store image metadata as DICOM[3] formatted images do, and for some kinds of

[3]DICOM (Digital Imaging and Communications in Medicine) is a standard format for storing and transmitting medical images. It includes not only the image data but also metadata about the image, such as patient information, acquisition parameters, and device settings, which are crucial for accurate diagnostic analysis and interoperability across different medical systems [82].

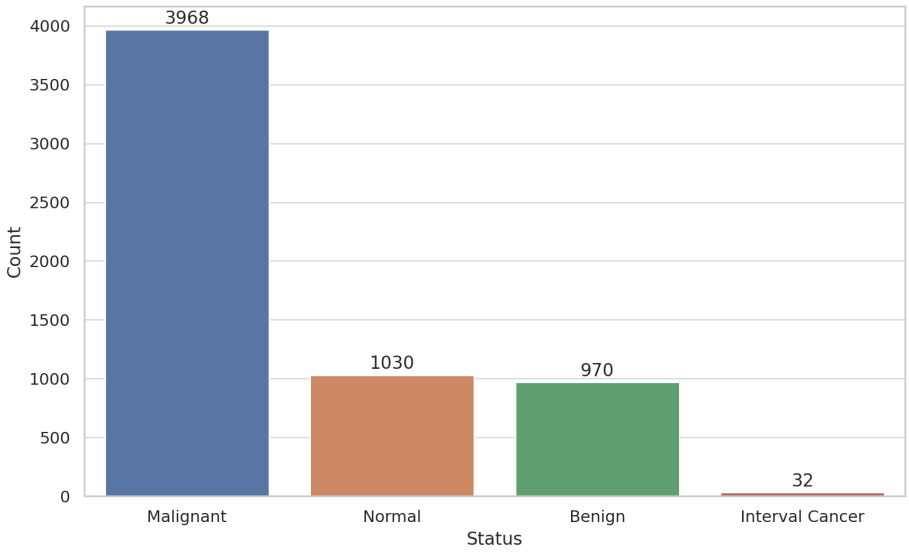

Figure 4.5: Bar chart showing the distribution of lesion status for clients in the OPTIMAM dataset, including counts of malignant, normal, benign, and interval cancer cases.

Table 4.2: Distribution of annotated masses in the OPTIMAM dataset classified by mass status and patient age. Each row contains the percentage and the total number of masses in each category.

Manufacturer	**Malignant**	**Benign**	**< 50**	**50–60**	**60–70**	**> 70**	**Total**
GE MEDICAL SYSTEMS	79.3%	20.7%	7.5%	58.5%	50.3%	11.7%	1267
HOLOGIC, Inc.	73.0%	27.0%	9.6%	63.9%	49.0%	8.5%	43257
Philips Digital Sweden	100.0%	0.0%	3.3%	39.2%	44.7%	12.8%	1393
Philips Medical Systems	50.0%	50.0%	20.9%	91.9%	48.8%	0.0%	139
SIEMENS	67.8%	32.2%	7.2%	66.6%	43.5%	6.5%	1474
Sectra Imtec AB	100.0%	0.0%	0.0%	0.0%	66.7%	33.3%	12

diagnostic analysis and ML models, acquisition parameters may be critical.

The choice of PNG format over DICOM has a number of effects:

- **Loss of Metadata**: All such necessary metadata regarding patient demographics, imaging parameter information, and device information are lost in the conversion to PNG format. Such metadata has clinical significance for being able to make a

holistic analysis of this nature and is also essential for training ML models that are dependent on contextual data.

- **Standardization Issues**: DICOM is a standardized format in medical imaging, hence allowing uniformity and consistency, as well as total interoperability. Adoption of the PNG format can introduce variability in terms of image quality and characteristics and, therefore, result in possible inconsistencies.
- **Analysis Limitations**: The lack of metadata restricts the possibility to apply some advanced analyses that depend on the conditions under which the images were taken. This could affect, for instance, the development and validation of diagnostic tools.
- **Advantages of PNG:**
  - **Simplified Handling**: PNG is a simpler format to handle and manipulate compared to DICOM, making it easier for non-specialized applications to process and display images.
  - **Wide Compatibility**: PNG format is supported on a vast majority of platforms and different software thus guaranteeing broad accessibility and ease of use.
  - **High-Quality Visuals**: The lossless compression of PNG ensures high-quality image preservation, which is beneficial for visual inspections.

Directory structure

The OMI-DB originally stored the data as follows. A main folder called *breast_images*, inside of which we have one folder per client. Then we have inner folders following the structure below for each of the studies and images carried out with each of the clients:

```
breast_images
└── client_id
    └── study_id
        └── serie_id
            └── image_id.png
```

Since all that information could be accessed through the CSV file mentioned earlier (section 4.2.1) with all the data regarding the dataset, we reordered the directories structure into a simple and more convinent one:

```
breast_images
└── client_id
    └── image_id.png
```

This way we could access easier each of the client's images for further data splitting.

While performing the reordering of the data, we also cropped all the images (Figure 4.6 and 4.7) using the aforementioned columns from the CSV file: **xmin_cropped**,

xmax_cropped, **ymin_cropped**, and **ymax_cropped**. This will be useful later, so that we only focus on the ROI[4] of the image, and we do not require any patch extraction techniques.

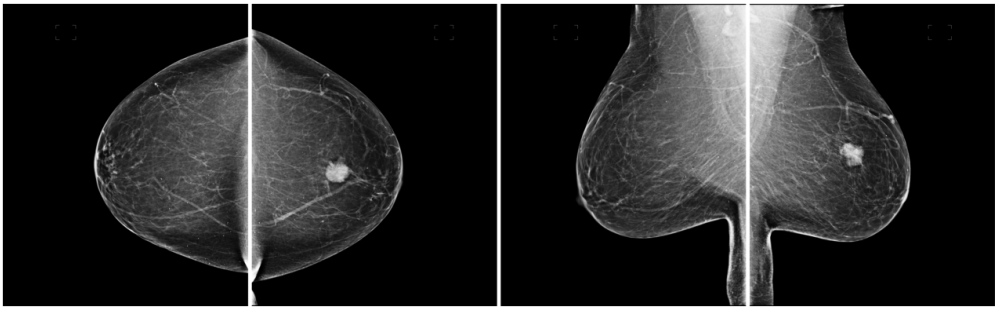

Figure 4.6: Example images from the OPTIMAM dataset (before cropping).

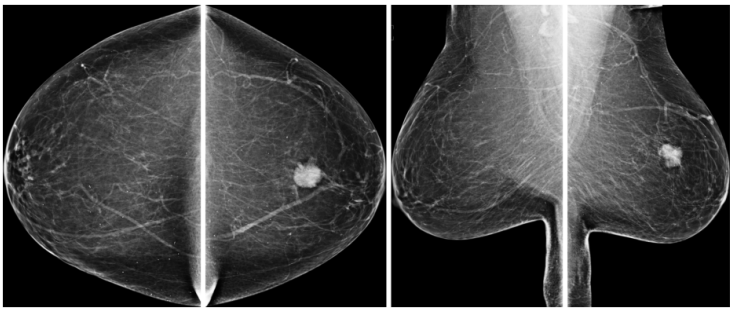

Figure 4.7: Example images from the OPTIMAM dataset (after background cropping).

4.3 Image Preprocessing

Preprocessing steps further enhance the quality and robustness of the image data for its classification applicability. These steps are shown in Figure 4.8 and include:

1. **Cropping**: This can be done by extracting relevant ROIs from the images using bounding box coordinates. The model would then focus only on those core regions rather than all the other regions containing irrelevant information.

2. **Noise Reduction**: Median filtering is employed to reduce image noise, since it can interfere with feature extraction and classification accuracy.

[4]Region of Interest

3. **Contrast Enhancement**: CLAHE[5] is applied to enhance the contrast of the images. This technique enhances the visibility of details, making it more notable for a classifier to distinguish various classes.

4. **Resizing**: The images are resized to a standard size of (224, 224), allowing consistency and minimizing computational complexity within ViT models. It also helps the features in the image remain intact across different images.

5. **Normalization**: Normalization of pixel intensity values within the range [0, 1] is adopted to ensure that all features have a similar scale, primarily those that belong to different manufacturers. This way, some features do not dominate over others in the training process, and it also helps in faster convergence of the optimization algorithm.

CLAHE is particularly well-suited for classification problems because it adapts the contrast enhancement locally based on the image content [83]. It means that regions with different contrast levels are normalized differently so that no critical information is lost at the preprocessing stage. Normalization further increases robustness by centering all input features around zero and making them standard on a common scale. This helps the optimization algorithm in its convergence and thus further enhances classifier performance.

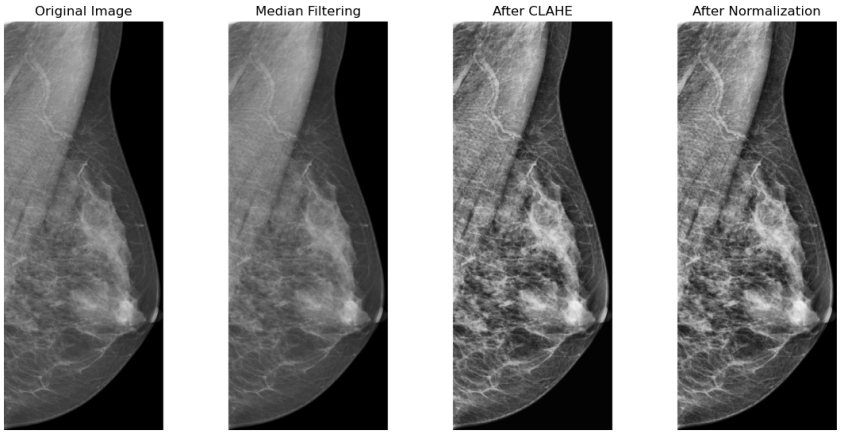

Figure 4.8: Preprocessing of an example image, showing the continuous process starting from the original image, applying median filtering, then CLAHE, and finally normalization.

4.4 Reproducibility through Seed Initialization

Setting a seed is paramount in ML to ensure that the results are reproducible. By controlling the randomness in data shuffling, weight initialization, and many other stochastic

[5]Contrast Limited Adaptive Histogram Equalization

processes, consistent results can be achieved. This is of particular importance in applications in medical imaging —say, classifying mammography images with ViTs like our particular case.

Reproducibility ensures that model results will be consistent across all different runs. It helps debugging and development, enables fair benchmarking, and supports scientific integrity. For medical purposes, results must be reproducible since decision-making depend on this idea.

The importance of reproducibility is recognized in the scientific community. The "reproducibility crisis" underlines general guidelines on how to reproduce findings, thereby pointing out the actual need for shared methodologies and data [84]. Best practices with ML, such as reproducibility, require seeds and systematic documentation of experiments to ensure responsible research [85]. This is also extremely important for AI applications in medical imaging, both to guarantee the reliability of the models in practice and to determine the impact they may have on patient care pathways [86].

4.5 Preventing Data Leakage

The data was split into the training, validation, and test sets based on "client_id," so no patients were overlapped among either the train, test or the validation sets. This careful separation is crucial to avoid data leakage, where the model possibly memorizes specific patient outcomes during the training process and then further recognizes those very patients when tested again, hence affecting the final testing performance. Consequently, due to this categorization, each patient was assigned exclusively to one of the three subsets to maintain the integrity of the evaluation process.

4.6 Data Splitting

We began each model training by splitting the data into 70% for training, 10% for validation, and 20% for testing. At this stage, ensuring that images for the same client fall in the same set is vital so that there will be no data leakage and that the generalization of models can be guaranteed, as we mentioned earlier.

The training subset has 70% of the initial data that is used to fit the model parameters. This subset feeds the model with sufficient data to learn patterns and relationships within the data. The validation set, taking 10% of the initial data, is used during training time; it is the subset on which hyperparameters are tuned and where decisions on when to stop training so as not to overfit are taken. Ultimately, the test subset, comprising 20% of the initial data, allows evaluation of the model's performance. The test set provides an unbiased estimate of how well the model generalizes to new, unseen data [87].

Here, a 70/10/20 split is chosen so that the available dataset gives enough data to train the model and still retains good portions for validation and testing to assess model

performance. This is commonly used in ML because it maintains the balance between sufficient data for training and the unbiased evaluation needed for testing. The recent full medical imaging domain generalization work by Garrucho et al. (2022) [88] states that a robust data-splitting strategy is expected to handle domain shifts in model generalization across diverse clinical environments.

The process we follow to split the data is as follows:

1. **Group data by client**: Firstly, images with their corresponding data from each client are grouped together to ensure that all the information from one particular client is under the same dataset (i.e., training, validation, or testing).

2. **Splitting strategy**: We employ a stratified split, considering the distribution of labels (i.e., Benign, Malignant) across clients. This makes sure that in each resultant set, the distribution is equal. We also use this step to omit the clients with a "Normal" or "Interval Cancer" status, since they are of no use for this thesis.

3. **Assign sets**: Each client group is then assigned to one of the sets (i.e., training, validation, or testing).

4. **Ensuring consistency across splits**: We guarantee that each client's images and their corresponding data are only in one set for consistency across splits.

Table 4.3: Class distribution for Train, Validation, and Test sets, showing the count and percentage of malignant and benign cases in each set.

Class Distribution	Count	Percentage
Train set (70% of total)		
Malignant	2777	80.35%
Benign	679	19.65%
Validation set (10% of total)		
Malignant	397	80.36%
Benign	97	19.64%
Test set (10% of total)		
Malignant	794	80.36%
Benign	194	19.64%

4.7 Data Loading

To load the images split into training, validation, and test sets, we use a function that reads image data and associated bounding boxes from the aforementioned CSV file and

processes them into a usable format for model training.

The process of loading the data is detailed in Figure 4.9:

Figure 4.9: Process of loading the data for model training.

When iterating through all the clients we construct the path to each client's image directory, and recursively walk through the directory to find all the images.

Then for each image we get its path and bounding box data associated. If this data exists, we process the image as detailed in section 4.3. Some of the patches loaded can be seen as an example in Figure 4.10.

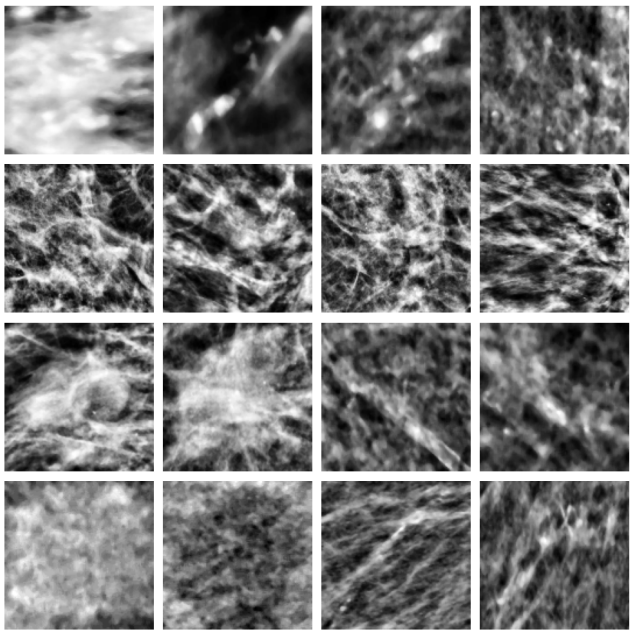

Figure 4.10: Random patches extracted from the dataset after loading the data.

4.8 Evaluation Metrics

In ML, predictive models and classifiers are evaluated based on particular metrics that generally allow for the comparison of results among themselves [89]. This is a pivotal section since choosing the proper evaluation metrics used in assessing results for a particular

classification problem is very important.

4.8.1 AUC - ROC Curve

The AUC–ROC Curve is a general norm when evaluating binary classification problems. The performance of the ROC Curve will be measured by the AUC. This section aims to explain this metric, laying down relationships between sensitivity, specificity, and thresholds.

AUC-ROC is usually applied in binary classification to summarize the general performance of a model at all available evaluation thresholds into a single value. This is particularly useful because the dataset is highly imbalanced, with very few positive cases (malignant) compared to negative cases (benign). AUC-ROC evaluates how well a model can discriminate between classes without consideration of the different thresholds. This is important in medical diagnosis because the cost for incorrect predictions of false negatives will be missing malignant cases. For the false positives, it will be incorrectly flagging benign cases. The superiority of AUC-ROC in handling class imbalance has been noted in various works, making it a preferred choice for our evaluation [90, 91, 92, 53].

Terms and Relations

The threshold is where observations get classified into class 0 or class 1; it is a probability value from 0 to 1 that bisects the two Probability Density Functions in binary classification, each modeling a class label. The values within the confusion matrix are determined by this threshold, which then affects sensitivity and specificity. These metrics change inversely as the threshold is adjusted from 0 to 1.

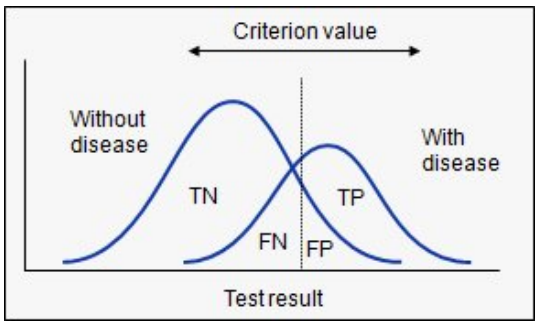

Figure 4.11: Illustration of the overlap of test results for two populations, showing true negatives (TN), false negatives (FN), true positives (TP), and false positives (FP) in relation to the criterion value [93].

For example, moving the threshold down makes the region that is considered 'with disease' more significant and so sensitivity increases. Similarly, moving the threshold up makes specificity more significant, as seen in Figure 4.11.

A model with an AUC close to 1 is considered perfect; such models lead to minimal misclassifications. On the other hand, a model in which AUC is close to 0 almost always means that the observations are misclassified. An AUC of 0.5 denotes a model that performs no better than random choice.

As evidenced by Figure 4.12, for an increasing value of the criterion, the sensitivity goes down and the specificity goes up. This can be attributed to the higher number of true negatives and lower number of false positives; hence, it establishes an inverse relationship between the two measures, with respect particularly to when the two classes are equiprobable and equivariant Gaussians.

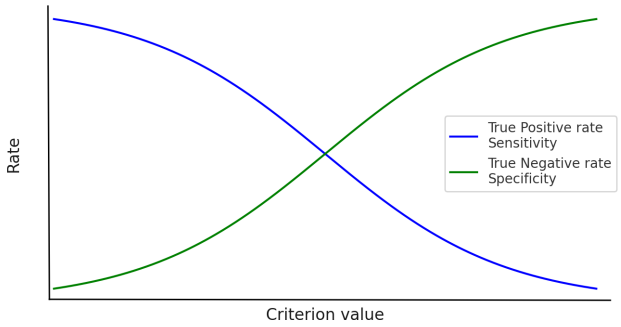

Figure 4.12: Graph illustrating the inverse relationship between sensitivity (true positive rate) and specificity (true negative rate) as a function of the criterion value.

It can be observed in Figure 4.13 how the blue line for the optimal trend would stay as close as possible to the top left corner of the plot; this position illustrates at the same time the best performance of both sensitivity and specificity, that is the rate of false positives and false negatives in the confusion matrix, set by the threshold.

The sensitivity is plotted on the vertical axis against 1-specificity on the vertical axis in the ROC plot. There is also the red dashed line representing the random classifier; for this case, it would run through the graph's origin since the AUC equals 0.5.

4.8.2 Confusion Matrix

The Confusion Matrix combines all the predictions from the dataset into four categories through the cross-referencing of the actual class with the predicted class, illustrated in Figure 4.14.

A confusion matrix gives a detailed account of the model's performance through the counts of true positives, true negatives, false positives, and false negatives. In medical imaging, an itemized breakdown and understanding of the source and types of classification errors are essential tools for model improvement and clinical decision-making [95].

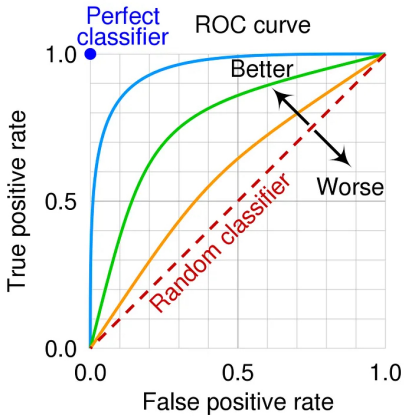

Figure 4.13: Graph showing the ROC curve, illustrating the true positive rate versus the false positive rate for different classifiers, highlighting the performance of a perfect classifier, a random classifier, and better or worse classifiers [94].

In 2022, Chaudhury et al. [96] revealed how to analyze the performance of various ML techniques in defining features to aid the most reliable diagnosis of breast cancer based on the confusion matrix. The research findings indicated that analysis using a different classifier depends on the confusion matrix. Notably, it would help pinpoint strengths and weaknesses in the model's performance by availing a comprehensive visualization of the model's predictions to help in effecting targeted improvements toward augmenting overall diagnostic accuracy and reliability.

These categories in the matrix not only indicate the accuracy of the predictions but also the nature of the prediction, and are defined as follows:

- **True Positives (TP)**: Correct predictions where the class is positively identified. For example, in mammography screening, this refers to malignant tumors correctly identified as malignant.
- **True Negatives (TN)**: Correct predictions where the class is negatively identified. In the context of mammography, this means benign conditions correctly identified as not malignant.
- **False Positives (FP)**: Incorrect predictions where a negative class is identified as positive. In mammography screening, this occurs when benign tumors are incorrectly identified as malignant, leading to unnecessary additional testing.
- **False Negatives (FN)**: Incorrect predictions where a positive class is identified as negative. In mammography, this refers to malignant tumors that are incorrectly identified as benign, potentially delaying essential treatment.

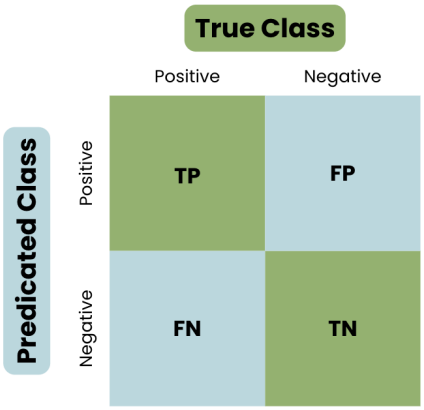

Figure 4.14: Diagram showing the basic structure of a confusion matrix with TP, FP, FN, and TN classifications [97].

4.8.3 Binary Classification

Binary classification metrics are essentially derived from the information produced by the confusion matrix elements, and evaluate the classifier's performance. Table 4.4 provides a summary list of these metrics, its formulas, and descriptions.

Accuracy

Accuracy is a broad and standard metric to look for the overall performance of a classifier, defined as the ratio of correct predictions out of the total predictions made. But in an imbalanced data situation, like medical diagnostics, it can be misleading. According to some studies, this leads to high accuracy even when the model fails to detect minority class observations (in this instance, benign cases) [95, 98].

Precision

Precision is then the measure of how many of the positively identified cases were actually positive. It is of particular importance in medical diagnostics, where false positive errors can cause unnecessary follow-up tests and treatments. Precision is critical when the cost of false positives is high [98, 99].

Sensitivity and Specificity

Sensitivity (recall) is the degree to which the model can successfully find all actual positive cases (malignant cases). High sensitivity is desired in medical applications so that the cases that are genuinely malignant will not be missed. It is considered more important in early-stage cancer detection, because prompt treatment could dramatically change the patient's outcome [98, 100].

Specificity is the ability of the model to identify negative cases correctly (benign cases). It allows for the reduction of false positives, hence minimizing unnecessary patient anxieties and follow-up procedures. It is essential for maintaining trust and reducing costs in healthcare [98, 101].

F1-Score

The F1-Score is the harmonic mean of precision and recall, balancing the two factors. It is most advantageous when imbalanced datasets are used, where a balance in precision and recall needs to be taken care of in order not to let the levels go very high in a falsely positive or negative manner [98].

Positive Predictive Value (PPV)

PPV is a measure of the proportion of positive results that are true positives. It indicates how high is the possibility that a positive testing result represents a malignant case, which testifies to the importance of assessing the diagnostic value of the test [98, 101].

Negative Predictive Value (NPV)

The NPV represents the proportion of accurate negative results relative to the total number of negative test results, and quantifies the probability that a negative test result truly indicates a benign case. This is important for patient reassurance and avoiding unnecessary treatments [98, 101].

The significant difference between PPV and NPV, with sensitivity and specificity, relates to what each set of metrics emphasizes: PPV and NPV focus on the test results' outcomes, either positive or negative, and thereby reflects how reliable the results are in confirming or excluding the disease. Sensitivity and specificity, however, focus on the overall ability of the outcome of the test to give an actual positive value and accurate negative results, thus provide much more scope of the accuracy of the test.

False Negative Rate (FNR)

FNR is the fraction of actual positive cases that are misclassified as negative. High FNR in medical diagnosis exhibits missed malignant cases, leading to delayed treatment and worse conditions for patient's recovery [98, 102].

False Positive Rate (FPR)

FPR measures the real negative cases that were incorrectly classified as positive. High FPR may result in unnecessary anxiety, follow-up tests, and more procedures for the patients. Hence, in practice, the level of FPR should be low to avoid these issues [98, 102].

Table 4.4: Binary classification metrics used to evaluate the model, including accuracy, precision, recall (sensitivity), specificity, F1-score, PPV, NPV, FNR, FPR.

Metrics	Formula	Description
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	Correct predictions over the total classifications.
Precision	$\frac{TP}{TP+FP}$	Positives correctly classified over total predicted as positives.
Recall (Sensitivity)	$\frac{TP}{TP+FN}$	Positives correctly classified over total positives.
Specificity	$\frac{TN}{TN+FP}$	Negatives correctly classified over total negatives.
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$	Equal importance between precision and recall.
PPV	$\frac{True\ Positives}{True\ Positives+False\ Positives}$	The proportion of positive results that are true positives.
NPV	$\frac{True\ Negatives}{True\ Negatives+False\ Negatives}$	The proportion of negative results that are true negatives.
FNR	$\frac{\text{False Negatives}}{\text{False Negatives}+\text{True Positives}}$	The proportion of actual positives incorrectly classified as negatives.
FPR	$\frac{\text{False Positives}}{\text{False Positives}+\text{True Negatives}}$	The proportion of actual negatives incorrectly classified as positives.

4.9 Description of the model architecture

4.9.1 Models' Input Format

The input format is essential regarding a system aiming to classify mammography images into a binary class of either benign or malignant. The original images are to be considered in PNG format, and are supposed to come from a medical context where such images have been deemed, by either a radiologist or another software, to be suspicious of having a tumor. Images go through a series of preprocessing steps described in section 4.3. Here, we focus only on the input format specific to the model and the model pipeline described in Figure 4.16.

Each preprocessed image is transformed into a 3D tensor, with dimensions (224, 224, 3), meaning it resizes all the images into 224×224 pixels and three color channels (Red-Green-Blue, RGB). Even though mammography images are initially in single-channel grayscale, they are triplicated to represent an RGB format. This is because the ViT model expects input images in RGB format.

These images are batch-processed during both training and inference. In that way, we

get the shape of model input as *(batch, 224, 224, 3)*. This batch dimension helps in efficiently processing and training multiple images at once using parallel processing features provided by modern hardware.

In summary, the input images to our ViT model have been preprocessed to 224 $\times$ 224 pixels and standardized as RGB images. This uniform, common format of input ensures smooth operations of the model's architecture in training, facilitating that all the mammography images can be appropriately classified into benign and malignant classes.

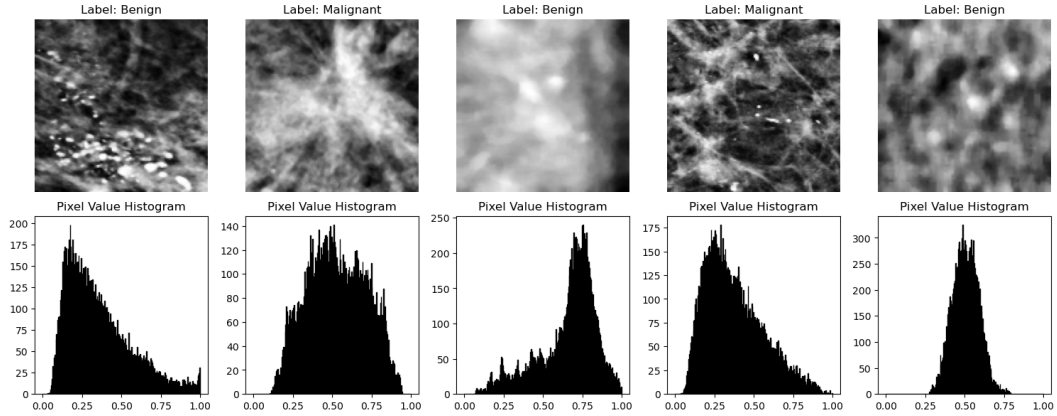

Figure 4.15: Examples of pixel intensity value histograms (bottom row) of preprocessed cropped mammography images (top row) with labels indicating benign and malignant cases.

4.9.2 Models' Output Format

The output of the ViT model, designed to classify mammography images as benign or malignant, is crucial for evaluating its performance and accuracy. After the preprocessed images are fed into the model, the following steps outline how the predictions are generated and interpreted.

The model's architecture concludes with a fully connected layer (classifier) that outputs logits corresponding to the two classes: benign and malignant. Logits are the raw, unnormalized scores, which the model computes as an output. As detailed in the subsequent steps, logits are further processed to obtain probabilities and predictions [103, 104].

During inference, the outputs of the model are obtained as follows:

1. **Batch Processing**: During inference, images are processed in batches. Each batch contains multiple images, which allows efficient parallel processing.
2. **Model Inference**: For each batch, the images are passed through the model to obtain the raw logits. Because the model type is a ViT, the logits are accessed via `net_final(inputs).logits`.

3. **Softmax Transformation**: The logits are converted to probabilities using the softmax function. This transformation ensures that the outputs are normalized and can be interpreted as probabilities of the images belonging to the benign or malignant class.

4. **Class Prediction**: The class with the highest logit value is selected as the predicted class. This is achieved by using the `max` function on the logits.

4.9.3 Transfer Learning and Pretrained weights

TL is considered one of the most critical factors in ML, especially when data is limited. Pretrained models support using transferable features for the best results for our specific task and lessen the need for extensive data.

In this study, we employed a ViT. The ViT architecture, known for its effectiveness in image classification tasks, was initialized with pretrained weights from the large ImageNet dataset. These pretrained weights carry a massive set of learned features, giving a perfect initialization point to our model to capture essential patterns and features right from the start. This is particularly beneficial when our dataset is limited, as it reduces the risk of overfitting and improves generalization.

Through experimentation, we observed that the performance of our model was significantly influenced by how we handled the pretrained layers. We considered freezing all layers except for the final one so that only the last layer was adapted to our specific task, but such step did not maximize the results. Instead, we achieved better performance and results by not freezing any layers, enabling the entire model to fine-tune itself based on our dataset.

This fine-tuning approach allows the model to adjust all its parameters, both low-level and high-level, to better capture the nuances of mammography images. The results of these experiments, and related performance improvements, will be further discussed in Chapter 5.

TL with ViT, and not freezing any layers during fine-tuning, provided us with the most effective strategy for our mammography classification task. This approach tackles the power of pretrained models while allowing for full adaptability to our specific dataset.

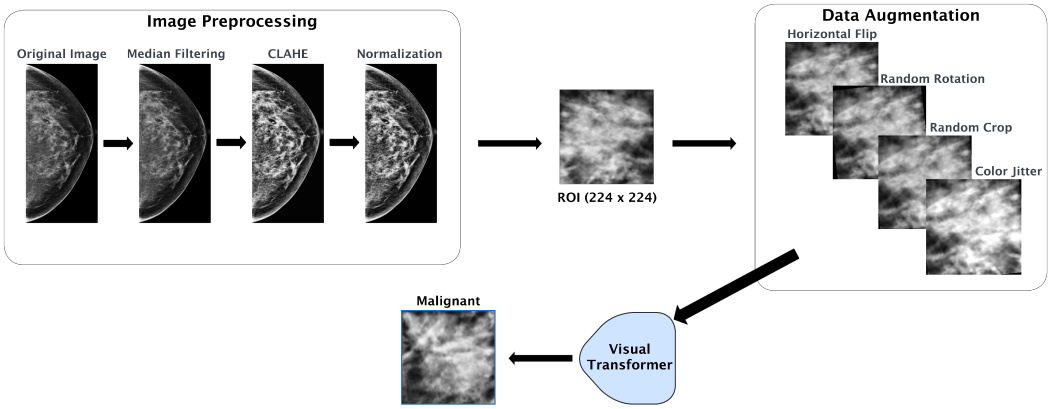

Figure 4.16: Model pipeline diagram showing the steps from image preprocessing, ROI extraction, data augmentation, and final classification using a ViT.

Chapter 5

Results, Analysis, and Discussion

This chapter focuses on a detailed discussion of data modeling processes, including the necessity and significance of keeping consistent datasets to ensure proper hyperparameter tuning and model evaluation. It will highlight the selection of the ViT model for its exceptional performance in image recognition tasks. Additionally, this chapter will explore data augmentation techniques used to enhance model generalization, discuss the fine-tuning of hyperparameters, and evaluate the model's performance using various metrics. Furthermore, it will address class imbalance issues and compare different methods to mitigate their impact. Finally, the chapter will present an extensive analysis of the evaluation metrics, including the AUC-ROC curve, confusion matrix, and precision-recall metrics, providing a comprehensive understanding of the model's effectiveness in breast cancer diagnosis.

5.1 Data Modelling

It is essential to make use of consistent data between different loops, so that in enables fair comparisons and, thereby, enable correct identification of the optimal hyperparameters. Changes in data splits, while tuning hyperparameters, can provide a wrong intuition for performance ranking within a model. Therefore, to facilitate fair comparisons of the trained models, the images in the datasets have been kept identical.

The ViT used model, google/vit-base-patch16-224 [105], was chosen based on its state-of-the-art performance with large-scale image recognition tasks. This architecture is built with a transformer, which has shown tremendous improvements, and has excelled in different vision tasks, particularly on datasets such as ImageNet. Being quite robust and efficient, the model is great for benchmarking and further experiments.

The following table shows the performance of the ViT used on the ImageNet validation dataset [1].

[1]The top-1 and top-5 accuracy refers to the model's performance on the ImageNet validation dataset.

Table 5.1: Performance metrics of the google/vit-base-patch16-224 model on ImageNet, including top-1 accuracy, top-5 accuracy, and the number of parameters.

Model	Top-1 Accuracy	Top-5 Accuracy	Number of Parameters
google/vit-base-patch16-224	77.9%	93.6%	86M

5.2 Data Augmentation

Real-world datasets are often too large to fit into memory and require data augmentation to increase generalization ability and reduce overfitting. The most common misconception is that the dataset after augmentation has a size greater than that of the original one, but this is not accurate. For this project, the original training set contained 5,909 observations: 715 benign and 5,194 malignant. When we perform data augmentations, the original data is transformed, but the training size remains the same. Thus, the model sees 5,909 training images per epoch, but these images differ for each epoch due to augmentation. As a result, with multiple epochs, the model is trained on many different augmented images, instead of just the original ones.

This is achieved by the transforms object that the PyTorch library uses, as it can do on-the-fly augmentations like translations, rotations, resizing, flipping, and zooming. The mechanism also normalizes the input by doing a simple division by its standard deviation. Changing the channels format, shuffle, and resize of inputs are built-in parameters as well. Thus, the DataLoader function does data augmentation over each batch, iterating over a whole epoch. This is suitable for large datasets that require batch data augmentation.

It is important to use different augmentations for training, validation, and test sets to avoid data leakage since normalization uses per-batch statistics which cannot be shared between training and testing sets.

The training transformations include:

- **Random Horizontal Flip**: This simulates the natural variability in medical images, such as differences in orientation, which helps the model become more robust to such changes.
- **Random Rotation**: Small rotations can help the model learn to recognize features that may appear at slightly different angles, improving its generalization ability.
- **Random Resized Crop**: This focuses the model on different parts of the images, helping it to learn important features that are not always centered.
- **Color Jitter**: This accounts for variations in imaging conditions, such as lighting and contrast, making the model more resilient to such changes.

- **Resize**: This ensures all images are of a uniform size, which is necessary for batch processing in neural networks.

These transformations (shown in Figure 5.1) are particularly valuable in the medical context and breast cancer classification, because allow the model to learn to recognize abnormalities, such as tumors, under changing conditions of imaging and changing anatomy or other pathologies of the patient.

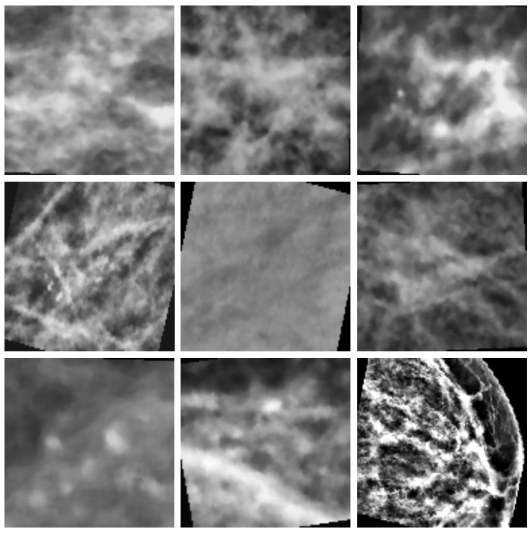

Figure 5.1: Nine examples of preprocessed and augmented mammography images, showing various transformations applied for data augmentation.

5.3 Fine-Tuning Hyperparameters

The following hyperparameters were fine-tuned during training and validation in combination with each other, rather than independently, to account for their interactions, as shown in Table 5.2. The validation set has been used to adjust the hyperparameters of the model, and to find the best number of epochs that would ensure it did not overfit. The test set has been left for use only in the evaluation process and was not employed in building the model.

Table 5.2: Results of hyperparameter tuning, including the modes/values tested and the best result obtained for each hyperparameter.

Hyperparameters	Mode/Values	Best Result
Learning Rate	0.0005, 0.00005	0.00005
Batch Size	16, 32	32

Hyperparameters	Mode/Values	Best Result
Optimizer	SGD, AdamW	AdamW
Scheduler	StepLR, ReduceLROnPlateau	StepLR

For ease of reference, we presented performance metrics for only four combinations of the different possible hyperparameters. For the sake of model training, we set a patience of 10, which means that if the AUC score does not further improve after ten epochs, it stops the training early to avoid overfitting.

Figure 5.2 represents the values of training and validation accuracies and their corresponding losses over epochs for different sets of hyperparameters. In the top plot, it was clear that models trained with a lower learning rate (0.00005) consistently showed higher values of validation accuracy, which means better generalization. These were the models that used the StepLR scheduler with a batch size of 32, and proved to be the most stable and produced the highest validation accuracy.

The bottom plot highlights the training and validation losses. Models with a learning rate of 0.00005 and using the StepLR scheduler achieved the lowest validation losses. This suggests that while StepLR is effective in enhancing accuracy, it can also minimize loss when paired with the right learning rate and batch size.

These results highlight the hyperparameter fine-tuning combination because interactions between learning rate, batch size, optimizer, and scheduler significantly affect the performance of a model. The patience value used here makes sure that the process of training stops in due time when there is no improvement in the AUC score (hence increasing model generalization), and not in later epochs which could lead to overfitting.

5.4 Model's Evaluation and Selection

The best model will be evaluated and selected based on the AUC-ROC Curve, as it considers both sensitivity and specificity, giving an overall measure of how the model performed. To thoroughly evaluate the model, we will also study other metrics, including the optimal threshold that maximizes the balance between sensitivity and specificity, the confusion matrix to provide a detailed view of TP, TN, FP, and FN, and the PPV and NPV to assess the accuracy of the model's positive and negative predictions. Sensitivity and specificity will be key factors, especially in the context of disease prediction, where high sensitivity is essential to avoid false negatives. The precision-recall curve will be analyzed to understand the trade-offs between precision (PPV) and recall (sensitivity). Finally, the F1-score will be used to balance precision and recall, which is important in cases of imbalanced classes.

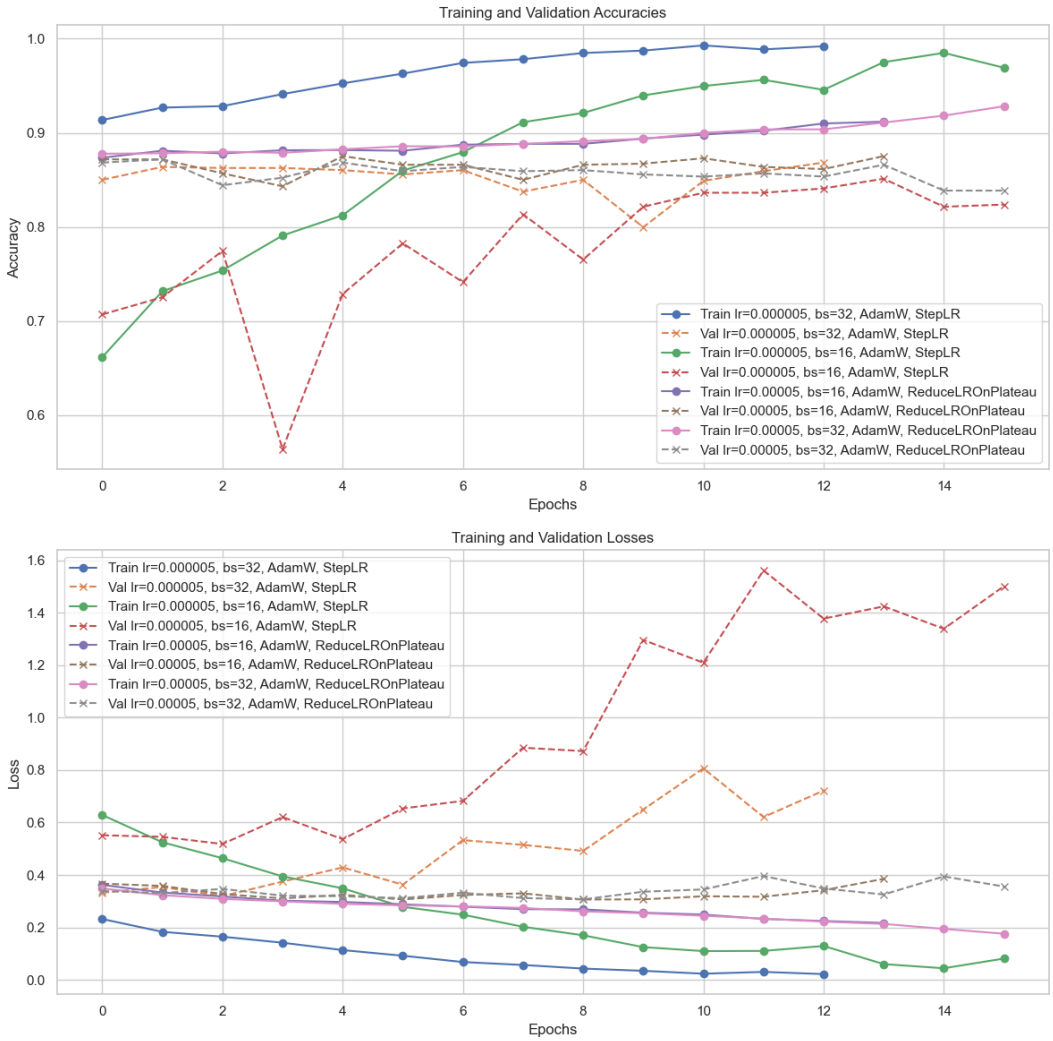

Figure 5.2: Training and validation accuracies (top) and losses (bottom) over epochs for different hyperparameter settings.

Ideally, we would want a model with high recall for both benign and malignant cases. This is important because:

- **Benign Cases**: High recall ensures that benign cases are correctly recognized, hence reducing unwanted anxiety, further testing, and probable overtreatment for those patients incorrectly reported as malignant cases.

- **Malignant Cases**: A high recall for malignant cases is essential. This is to ensure that most, if not all, cases of malignancy are detected. Missing a case with malignancy would be very serious because the consequence would be that treatment might be delayed, the condition might advance, and the patient might have worse outcomes.

Keeping high recall in both malignant and benign cases will, therefore, balance the trade-offs. This way, the model minimizes the risk of false negatives in malignant instances while at the same time increasing the accuracy of detecting benign cases, which leads to better clinical decisions and, ultimately, better patient care.

5.5 Class Imbalance

Class imbalance happens when class labels are unequally distributed, leading to biased loss during training and testing. To address this, three validated and common techniques are tested: weighted loss function, oversampling, and undersampling [106].

5.5.1 Weighted Loss Function

Class Weights

The goal is to emphasize the under-represented class by assigning it a larger weight. Each class gets a weight based on its frequency in the dataset. The minority class gets a higher weight, typically calculated as the ratio of the majority class to the total observations [107]. As shown in Table 5.3, the computed class weights are:

Table 5.3: Computed class weights for the negative (benign) and positive (malignant) classes.

Class	Weight
Negative (Benign)	4.13
Positive (Malignant)	0.57

Incorporating class weights into the loss function modifies the standard cross-entropy loss to account for class imbalance. This weighted cross-entropy loss can be formulated as:

$$\mathcal{L}_{WCE}(y, \hat{y}) = -\sum_{i=1}^{N} w_i \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{5.1}$$

where w_i is the weight for class i, y_i is the true label, and $\hat{y}_i$ is the predicted probability for class i. By assigning higher weights to the minority class, the loss function penalizes misclassifications of the under-represented class more heavily, thus encouraging the model to perform better on these samples. This technique is particularly useful in medical diagnosis tasks, where the cost of misclassifying a rare but critical condition (such as malignancy) is high.

5.5.2 Oversampling the minority class

The minority class (benign) is oversampled using *medigan* [108] to reach an equally training set.

Medigan stands for Medical Generative Adversarial Networks. It provides a user-friendly platform for medical image synthesis, allowing users to select from a variety of pretrained generative models to create synthetic datasets. These datasets can be used to train or adapt AI models for clinical tasks such as lesion classification, segmentation, and detection.

Instead of developing a generative model from scratch, a pretrained model from *medigan* is employed to generate synthetic data. The model used to generate images, which are subsequently preprocessed in the same way as the original images from OMI-DB, is the so-called *Conditional DCGAN Model for Patch Generation of Mammogram Masses Conditioned on Biopsy Proven Malignancy Status*, trained on the CBIS-DDSM dataset. This model is a class-conditional deep convolutional generative adversarial network (DCGAN) that generates mammogram mass patches conditioned to be either benign (1) or malignant (0). The generated patches have pixel dimensions of 128x128. The Cond-DCGAN model was trained on mammogram patches from the CBIS-DDSM dataset [108].

5.5.3 Undersampling the majority class

The technique should be applied to balance the dataset in which the count of instances of the majority class has to be reduced, achieving equal distribution of benign and malignant samples in the training set. The impact that this technique can have on the training of the model is the following:

1. **Balanced Training Set**: Prevents bias towards the majority class and provides fair learning from both classes.

2. **Improved Performance Metrics**: This implies increased sensitivity and specificity from testing and validation, thus leading to better generalization.

3. **Reduced Overfitting**: Tackles the problem of overfitting on the majority class and helps the classifier perform better with unseen data.

4. **Loss of Information**: Can be a potential disadvantage due to the deletion of many samples of the majority class, which may not lead to comprehensive learning.

5.5.4 Comparison of the results

For this task, we used the google/vit-base-patch16-224 model, which is a ViT model designed by Google [105]. This model was chosen due to its superior performance in image classification tasks, making it suitable for distinguishing between malignant and benign cases in mammography screenings. All parameters were fixed except for the approach used to address class imbalance. The initial results are listed in Table 5.4.

[2]The RadImageNet database is an open-access medical imaging database designed to improve transfer learning performance on downstream medical imaging applications. It includes 1.35 million annotated CT, MRI, and ultrasound images covering a wide range of pathologies and anatomical regions [109].

Table 5.4: Comparison of ViT and ResNet50 (pretrained on RadImageNet[2]) results using different methods, showing the AUC for weighted loss function, oversampling, and undersampling techniques.

Model	Method	AUC
ViT	Weighted Loss Function	**0.91**
	Oversampling	0.81
	Undersampling	0.77
ResNet50	Weighted Loss Function	**0.76**
	Oversampling	0.75
	Undersampling	0.72

The results show that the ViT model achieves the highest AUC of **0.91** when using the weighted loss function. Good performance was also shown in oversampling and undersampling, with AUC values of 0.81 and 0.77, respectively. Based on these results, we will focus on using the weighted loss function for further analysis and model development.

The ResNet50 model pretrained with RadImageNet gave AUC scores of **0.76** using a weighted loss function, 0.75 when oversampling was used, and 0.72 when undersampling is performed. These metrics indicate that while RadImageNet pretraining can provide some benefits, it is the ViT architecture, in combination with the weighted loss function, that performs best for this specific task.

The model shall perform better with pretrained weights on RadImageNet than with ImageNet due to closer domain relevance, since RadImageNet includes several medical imaging modalities. This closer domain alignment brings out the transfer learning capability more effectively for our specific task.

It is worth mentioning that this comparison to ResNet50 was done purely for benchmarking our ViT model, followed by the comparison made in 3.2.2. All results and analyses are going to be based on the ViT model, given its demonstrated superior performance in this task.

5.6 Evaluation Metrics and Analysis

This section presents the evaluation metrics used to assess the malignancy predictions in patients with abnormal breast cancer screenings. It is important to note that the evaluation was performed on the test set, which was not exposed to the model either during fine-tuning of the hyperparameters or during training.

5.6.1 AUC-ROC Curve and Thresholds

Setting the threshold at which a prediction will be made adjusts the ROC curve. The traditional ROC Curve assumes that predictions under a threshold of 0.5 are classified as negative, and above it are positive. In medical literature, the AUC-ROC is described as the probability of a randomly selected patient with the disease being correctly diagnosed as having the disease, as depicted in Figure 5.3.

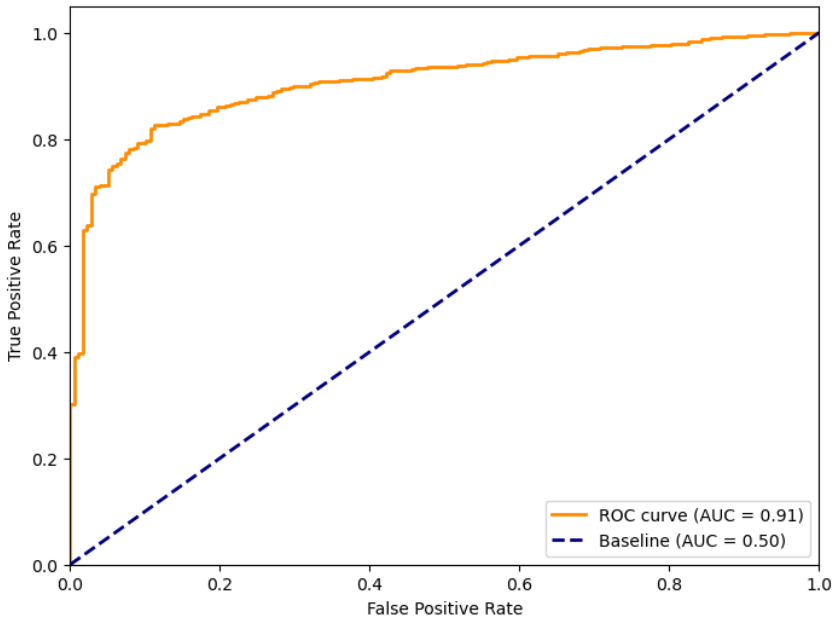

Figure 5.3: ROC Curve (including AUC) of the best-performing model.

The ROC curve obtained, with an AUC result of **0.91**, demonstrates that the model has good discrimination between the positive and negative classes; hence, it shows high accuracy for diagnoses. This high AUC value reveals the model as significantly better than a completely random one, which would have a value of 0.5.

Since the curve lies more to the top-left corner, it means that there is a certain optimal threshold, characterized by maximizing the true positive rate and, at the same time, keeping a relatively low false positive rate. This is an essential characteristic of diagnostic models in medical settings, ensuring high sensitivity (minimal missed true cases) and specificity (minimal false alarms).

Our model performs quite competitively when considering the AUC of 0.91 compared to the study by Gheflati and Rivaz (2022) [110], who used ViTs in the context of breast ultrasound image classification. They reported a close performance with an AUC of 0.95, slightly higher than ours, indicating the effectiveness of ViTs in achieving high diagnostic accuracy in medical imaging.

Optimal Threshold

A threshold value of 0.5 is often used in the literature on binary classification, as it is an intuitive and balanced point to start from. If the predicted probability of the positive class is greater than or equal to 0.5, an instance is classified as being of the positive class; otherwise, as being from the negative class. This is selected as default because it is at the center of the probability range, thus making a linear and fair decision rule without putting initial bias toward any class [111].

However, for some applications, this threshold might not be appropriate. Modifications could be needed to reduce FPs or FNs as much as possible, particularly in the case of medical diagnosis, where the cost of misclassification is very high. The optimal threshold can be determined by testing for different thresholds and then choosing the one at which the balanced accuracy is maximized.

We test it with various thresholds to find the best decision boundary where the model gives us maximum performance. We then calculate a performance metric, in our case the balanced accuracy, at each threshold, and from this, we select the one that produced the highest value in the chosen metric. This method ensures that we choose a threshold according to our needs and constraints within our task. The optimal threshold is shown in Table 5.5.

Table 5.5: Optimal thresholds to maximize the balanced accuracy.

Threshold	Accuracy
0.5	0.6637
0.16	0.8234

5.6.2 Confusion Matrix

After we choose the threshold, the confusion matrix tells us how images were classified compared to their actual class label. The results can be seen for both thresholds in Table 5.6 and 5.7.

In breast cancer diagnosis, reducing the number of false negatives is a major challenge since breast cancer screenings are usually performed once a year. However, with the reduction of false positives, there is a possibility of avoiding a lot of unnecessary surgeries, biopsies, and radiation for patients, which are harmful, stressful, and expensive.

5.6.3 Binary Classification

Precision-Recall Curve

This metric comes in handy when the classes are very imbalanced. Indeed, precision can correspond to the PPV that gives the probability of finding true positives among the

Table 5.6: Confusion matrix values when the threshold of decision is set to 0.5.

	Predicted Benign	Predicted Malignant
Actual Benign	175	3
Actual Malignant	557	930

Table 5.7: Confusion matrix values when the threshold of decision is set to 0.16.

	Predicted Benign	Predicted Malignant
Actual Benign	159	19
Actual Malignant	275	1212

group of patients classified as positive. Recall, on the other hand, gives the probability of detecting the positive cases.

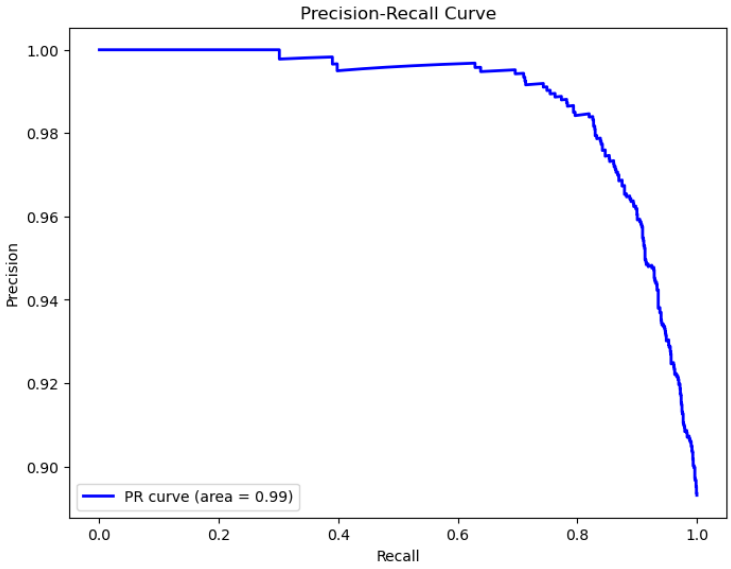

Figure 5.4: Precision-Recall Curve for the best trained model, with an AUC of 0.99.

In the precision-recall curve shown in Figure 5.4, precision is very high, close to 1.0, for a large part of recall values. It means that the classifier works quite well in separating positive cases from negative ones. Aditionally, the AUC comes out to be **0.99**, which again brings the point to the high performance of the classifier. As recall increases, precision slightly decreases but remains above 0.90, demonstrating that the classifier maintains a

high level of accuracy even as it identifies more positive cases.

Our results align closely with those obtained by similar studies using ViT models for breast cancer detection. For instance, a study by Se-Woon Choe et al. in 2023 [53] achieved an AUC of 1.00 using a ViT model for mammogram classification, highlighting the efficacy of transformer models in this domain. This comparison underscores the robustness and reliability of ViT models in achieving high precision and recall in breast cancer classification tasks, further validating the potential of our approach.

Recall (Sensitivity) and Specificity

The optimal threshold is the point that shows the highest percent of sensitivity and specificity. Referring to Table 5.8, it is the sensitivity that yields 81.51% in identifying positive cases correctly, thereby reducing the number of false negatives, and its specificity rate accounts for 98.31% in detecting negative cases correctly, thus minimizing the scope for false positives. These metrics are essential for assuring precision and reliability of diagnostics, mostly in minimizing missed positive cases and unnecessary biopsies for negative cases.

Table 5.8: Sensitivity and Specificity values when the decision threshold is set to 0.5 and 0.16, respectively.

	Threshold 0.5	**Threshold 0.16**
Sensitivity	0.6254	0.8151
Specificity	0.9831	0.8933

On a different dataset but the same model architecture of ViT, Shiri et al. (2024) [112] obtained a specificity of 89.71% and a sensitivity of 87.51%. This demonstrates the effectiveness of the ViT architecture in maintaining high diagnostic accuracy and reducing false positives, which is critical for clinical applications.

F1-Score

In this context, an F1 score of 0.89 (using the optimal threshold) indicates that the classifier has a good balance between precision and recall, performing well in identifying the positive cases while maintaining a low rate of false positives.

Table 5.9: F1-Score values when the decision threshold is set to 0.5 and 0.16, respectively.

	Threshold 0.5	**Threshold 0.16**
F1-Score	0.77	0.89

It is this performance that makes a mark when compared to the other studies using ViT for breast cancer classification. For instance, Wang et al. (2022) [113] developed a semi-

supervised learning model based on ViT with ultrasound and histopathology datasets, showing an F1 score of 96.15%, outperforming CNN models such as DenseNet201, ResNet101, or VGG19.

Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

In the context of diagnostics, sensitivity and specificity are less relevant since they presume that the patient already has or does not have the disease. Instead, the PPV and NPV give an indication of whether a patient has the disease, thus making more sense diagnostically.

Table 5.10: PPV and NPV at decision thresholds of 0.5 and 0.16, respectively.

	Threshold 0.5	**Threshold 0.16**
PPV	0.9968	0.9846
NPV	0.2391	0.3663

At a threshold of 0.5, the PPV is 0.9968, meaning that 99.68% of patients with a positive screening test are true positives. The NPV at this threshold is 0.2391, indicating that 23.91% of patients with a negative screening test are true negatives.

At a threshold of 0.16, the PPV is 0.9846, meaning that 98.46% of patients with a positive screening test are true positives, which is almost all patients, making it highly effective for confirming the disease. The NPV at this threshold is 0.3663, indicating that 36.63% of patients with a negative screening test are true negatives, highlighting a potential area for improvement in reducing false negatives. This is an essential distinction for clinical decision-making —especially in reassuring patients while awaiting biopsy— because the strength of the test lies in confirming rather than excluding the disease.

False Negative Rate (FNR) and False Positive Rate (FPR)

There is a trade-off between the number of false negatives and false positives when choosing a threshold. This trade-off can be decided by computing a weighted sum of the false positives and false negatives as the cost of misclassifying, where each term has its own weight cost associated. Setting such parameter values is subjective and depends upon what we want to prioritize, between minimizing the amount of patients that got sent to biopsy with no malignancy (false positive) or getting a cancer diagnosis (false negative).

For example, at a threshold of 0.50, we have 557 false negatives and 3 false positives. This results in a FNR of:

$$\text{FNR} = \frac{557}{557 + 930} = 0.374,$$

and a FPR of:

$$\text{FPR} = \frac{3}{3 + 175} = 0.017.$$

At the optimal threshold of 0.16, we have 275 false negatives and 19 false positives. This results in a FNR of:

$$\text{FNR} = \frac{275}{275 + 1212} = 0.185,$$

and a FPR of:

$$\text{FPR} = \frac{19}{19 + 159} = 0.107.$$

Setting a threshold at 0.16 implies that 18.5% of the people with malignant tumors (TP and FN) would be classified as not having the disease (FNR). On the other hand, 10.7% of people without malignant tumors (TN and FP) would be classified as having the disease (FPR).

This threshold aims to balance reducing the FPR to avoid unnecessary emotional anxiety and financial burden for patients while also managing the FNR to ensure early diagnosis and treatment.

It is essential to remember that this classifier's objective is to output the likelihood of having a malignant tumor based on patients with an abnormal screening. False negatives are only temporary until the results of the biopsy are disclosed. Accordingly, the classifier's aim is more one of relief for the patients than create unnecessary anxiety. The priority is therefore put on a low FPR and high confidence in truly not having a disease.

Chapter 6

Conclusions and Future Work

6.1 Results

The main goal of this thesis was to develop an AI-based tool that would effectively differentiate between the malignant and benign states of breast tumors using screening mammography images. The scope of our study was focused on examining the applicability of recent Deep Learning models, such as Vision Transformers, with high capabilities to capture global context through self-attention mechanisms, in comparison with traditional CNNs.

In this regard, the objective has been reached and explored in detail using the OPTIMAM Medical Database, which has led to obtain a classifier able to predict malignancy from abnormal breast cancer screening with an AUC of 0.91, a sensitivity of 81%, specificity of 89%, and an F1-Score of 0.89, with an optimal threshold of prediction set at 0.16.

Before the patient undergoes a biopsy and while they wait for the results, it would now be possible to share with them the likelihood of malignancy by extracting the predicted value from a mammogram sent into the ViT classifier. Women with low probability scores would feel less anxious as a result, and unnecessary, costly follow-up exams could be avoided.

6.2 Progress, Limits, and Responsibilities

Future research would, therefore, need to be focused on testing the developed model in a variety of domains and datasets to be able to prove generalizability and robustness. This can show what the model is capable of over different mammography images from available data in different geographical regions and medical settings. Additionally, expanding the model scope to treat multiclass classification, not only binary, would enable much more detailed work in providing diagnostic insights. This extension would allow the model to classify different types of breast tumors and potentially other related abnormalities, increasing its scope for clinical utility.

Another important direction for further work would be improving the model based on the limitations observed during this study, these being the need for large-scale annotated datasets for practical training of ViTs. Collaboration with institutions that provide healthcare services to access higher volumes of data and have them annotated would improve the performance and generalizability of the model. Secondly, incorporating more clinical variables along with image data using the DICOM format could have also improved the diagnostic accuracy of the model.

Last but not least, making sure that the use of AI in clinical contexts is ethical and protects patients' privacy is the responsibility of researchers and developers. We need not only transparency and interpretability of the AI models developed, but is also pivotal to involve clinicians along the whole process, thus aligning the AI-tools with real-world medical needs. There must be a mechanism to allow the continuous monitoring and validation of AI models in medicine to guarantee efficacy and safety while using them in an actual clinical setup.

Bibliography

[1] Breast Cancer, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Absolute Numbers, Incidence, Females, in 2022, 2022. [Online]. Available: https://gco.iarc.fr/today/en/dataviz/pie?mode=cancer&group_populations=1&sexes=2

[3] Iranian Journal of Public Health, *Disparities in Incidence and Mortality of Breast Cancer*, 2016.

[4] Mammograms: What to Know About False-Positive Results, 2015. [Online]. Available: https://newsnetwork.mayoclinic.org/discussion/mammograms-what-to-know-about-false-positive-results/

[5] Artificial Intelligence Improves Breast Cancer Detection on Mammograms in Early Research, 2020. [Online]. Available: https://news.northwestern.edu/stories/2020/01/ai-breast-cancer/

[6] RG Bar, ZZCJ, *Probably Benign Lesions at Screening Breast US*, RSNA, pp. 701-712, 2013.

[7] M. Halling-Brown, L. M. Warren, D. Ward, E. Lewis, A. Mackenzie, M. G. Wallis, L. Wilkinson, R. M. Given-Wilson, R. McAvinchey, and K. C. Young, "Optimam mammography image database: a large-scale resource of mammography images and clinical data," *Radiology: Artificial Intelligence*, vol. 3, no. 1, Art. no. e200103, 2020.

[8] MT Scott Mayer McKinney, SJG, *International evaluation of an AI system for breast cancer screening*, Nature, January 2020.

[9] What Is Cancer?, 2021. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[10] Different Kinds of Breast Lumps, 2024. [Online]. Available: https://cancer.stonybrookmedicine.edu/breast-cancer-team/patients/bse/breastlumps

[11] Breast Masses: Cancerous Tumor or Benign Lump?, 2022. [Online]. Available: https://www.verywellhealth.com/breast-cancer-tumors-or-benign-masses-430277

[12] Breast Tumors, 2024. [Online]. Available: https://www.nationalbreastcancer.org/breast-tumors/

[13] What is Dense Breast Tissue, What Do You Need to Know?, 2021. [Online]. Available: https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/what-is-dense-breast-tissue-what-do-you-need-to-know

[14] Breast Density and Supplemental Screening, 2017. [Online]. Available: https://www.sbi-online.org/white-papers/breast-density-and-supplemental-screening

[15] Breast Screening (Mammogram), 2024. [Online]. Available: https://www.nhs.uk/conditions/breast-screening-mammogram/

[16] Breast Screening Programme, England, 2022-23, 2024. [Online]. Available: https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme/england---2022-23

[17] New Breast Screening Figures Prompt Fresh Uptake Appeal, 2024. [Online]. Available: https://www.england.nhs.uk/2024/01/new-breast-screening-figures-prompt-fresh-uptake-appeal/

[18] La situación del cáncer en España, 2024. [Online]. Available: https://www.sanidad.gob.es/ciudadanos/enfLesiones/enfNoTransmisibles/docs/situacionCancer.pdf

[19] Las cifras del cáncer en España, 2024. [Online]. Available: https://www.geicam.org/wp-content/uploads/2024/03/LAS_CIFRAS_2024.pdf

[20] Breast Screening, 2023. [Online]. Available: https://about-cancer.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/screening-breast

[21] Memorial Sloan Kettering Cancer Center, *Anatomy of the Breast*, 2024. [Online]. Available: https://www.mskcc.org/cancer-care/types/breast/anatomy-breast

[22] Mayo Clinic, *Breast Biopsy*, 2024. [Online]. Available: https://www.mayoclinic.org/tests-procedures/breast-biopsy/about/pac-20384812

[23] Cleveland Clinic, *Breast Biopsy*, 2024. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/24204-breast-biopsy-overview

[24] University Hospitals Birmingham NHS Foundation Trust, *The Warwickshire, Solihull & Coventry Breast Screening Service*, 2016. [Online]. Available: https://hgs.uhb.nhs.uk/wp-content/uploads/Presentation-for-gps-November-2016-no-22.pdf

[25] Breast Cancer Research Foundation, *Dense Breast Tissue: What It Means and What to Know*, 2023. [Online]. Available: https://www.bcrf.org/blog/dense-breast-tissue-what-it-means-and-what-to-know/

[26] Cleveland Clinic, *Dense Breast Tissue*, 2022. [Online]. Available: https://my.clevelandclinic.org/health/diseases/21169-dense-breast-tissue

[27] Mayo Clinic, *Dense Breast Tissue: What it Means to Have Dense Breasts*, 2024. [Online]. Available: https://www.mayoclinic.org/tests-procedures/mammogram/in-depth/dense-breast-tissue/art-20123968

[28] Survival for Breast Cancer, 2023. [Online]. Available: https://www.cancerresearchuk.org/about-cancer/breast-cancer/survival

[29] M. Brahim, K. Westerkamp, L. Hempel, R. Lehmann, D. Hempel, and P. Philipp, "Automated Assessment of Breast Positioning Quality in Screening Mammography," *Cancers*, vol. 14, no. 19, Art. no. 4704, 2022. DOI: https://doi.org/10.3390/cancers14194704

[30] Mediolateral Oblique View, 2023. [Online]. Available: https://radiopaedia.org/articles/mediolateral-oblique-view

[31] Positioning and Technique, 2024. [Online]. Available: https://www.uclahealth.org/departments/radiology/education/breast-imaging-teaching-resources/screening-mammogram/positioning-and-technique

[32] Radiology Key, *Breast Imaging: Mammography*, 2021. [Online]. Available: https://radiologykey.com/breast-imaging-mammography/

[33] Wikipedia, *CT Scan*, 2024. [Online]. Available: https://en.wikipedia.org/wiki/CT_scan

[34] Revolutionary Koning Vera Breast CT, 2024. [Online]. Available: https://www.koninghealth.com/news-insights/press-releases/koning-announces-first-new-york-metropolitan-area-installation-of-revolutionary-koning-vera-breast-ct-at-community-radiology-ny

[35] Dedicated Breast CT Research, 2024. [Online]. Available: https://health.ucdavis.edu/radiology/research/bcti.html

[36] Magnetic Resonance Imaging (MRI), 2024. [Online]. Available: https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri

[37] CT Scan vs. MRI: How They Work and What They Show, 2023. [Online]. Available: https://health.clevelandclinic.org/ct-scan-vs-mri

[38] Breast MRI, 2024. [Online]. Available: https://www.facs.org/for-patients/the-day-of-your-surgery/breast-cancer-surgery/preoperative-tests-and-imaging/breast-mri/

[39] Siemens Healthineers, *Breast MRI: High Sensitivity and Specificity to Support Cancer Imaging*, 2024. [Online]. Available: https://www.siemens-healthineers.com/en-au/magnetic-resonance-imaging/clinical-specialities/breast-mri

[40] Breast Cysts, 2024. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/breast-cysts/symptoms-causes/syc-20370284

[41] Breast Ultrasound, 2024. [Online]. Available: https://www.breastcancer.org/screening-testing/ultrasound

[42] A Start-To-Finish Guide To Performing A Breast Ultrasound, 2022. [Online]. Available: https://sonographyminutes.com/a-start-to-finish-guide-to-performing-a-breast-ultrasound/

[43] Breast Ultrasound, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Breast_ultrasound

[44] N. Hawkes, "Cancer survival data emphasise importance of early diagnosis," *BMJ*, vol. 364, Art. no. l408, Jan. 2019. DOI: 10.1136/bmj.l408. PMID: 30683652.

[45] P. R. Eby, S. Ghate, and R. Hooley, "The Benefits of Early Detection: Evidence From Modern International Mammography Service Screening Programs," *Journal of Breast Imaging*, vol. 4, no. 4, pp. 346-356, Jun. 2022. DOI: 10.1093/jbi/wbac041

[46] Understanding Dense Breast Tissue, 2024. [Online]. Available: https://www.ucealth.com/en/conditions/dense-breast-tissue

[47] J. H. Tanne, "Breast cancer is overdiagnosed in one in six or seven cases, finds large US study," *BMJ*, vol. 376, Art. no. o581, Mar. 2022. DOI: 10.1136/bmj.o581. PMID: 35246450.

[48] O. Peart, *Mammography and Breast Imaging: Just The Facts*, 1st ed., McGraw-Hill Medical, Apr. 2005, ISBN: 978-0071431200.

[49] J. Fenton, L. Abraham, S. Taplin, B. Geller, P. Carney, C. D'Orsi, J. Elmore, and W. Barlow, "Effectiveness of Computer-Aided Detection in Community Mammography Practice," *Journal of the National Cancer Institute*, vol. 103, pp. 1152-1161, Aug. 2011. DOI: 10.1093/jnci/djr206

[50] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017. DOI: https://doi.org/10.1016/j.media.2017.07.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841517301135

[51] K. Geras, S. Wolfson, S. Kim, L. Moy, and K. Cho, "High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks," Mar. 2017.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[53] G. Ayana, K. Dese, Y. Dereje, Y. Kebede, H. Barki, D. Amdissa, N. Husen, F. Mulugeta, B. Habtamu, and S.-W. Choe, "Vision-Transformer-Based Transfer Learning for Mammogram Classification," *Diagnostics*, vol. 13, no. 178, 2023.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L·Laiser, and I. Polosukhin, "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017, pp. 5998-6008.

[55] P. Radiuk, "Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets," *Information Technology and Management Science*, vol. 20, pp. 20-24, Dec. 2017. DOI: 10.1515/itms-2017-0003

[56] Understand the Impact of Learning Rate on Neural Network Performance, 2020. [Online]. Available: https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/

[57] P. Charilaou and R. Battat, "Machine learning models and over-fitting considerations," *World Journal of Gastroenterology*, vol. 28, no. 5, pp. 605-607, Feb. 2022. DOI: 10.3748/wjg.v28.i5.605. PMID: 35316964; PMCID: PMC8905023.

[58] StepLR, 2023. [Online]. Available: `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.StepLR.html`

[59] ReduceLROnPlateau, 2023. [Online]. Available: `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html`

[60] I. Gogul and S. Kumar, "Flower Species Recognition System Using Convolution Neural Networks and Transfer Learning," in *Proceedings of the International Conference on Communication and Signal Processing (ICSCN)*, Mar. 2017, pp. 1-6. DOI: 10.1109/ICSCN.2017.8085675

[61] R. Dhiman, G. Joshi, and R. Challa, "A Deep Learning Approach for Indian Sign Language Gestures Classification with Different Backgrounds," Journal of Physics: Conference Series, vol. 1950, no. 1, pp. 012020, Aug. 2021. DOI: 10.1088/1742-6596/1950/1/012020

[62] The World Through the Eyes of CNN, 2020. [Online]. Available: `https://medium.com/analytics-vidhya/the-world-through-the-eyes-of-cnn-5a52c034dbeb`

[63] Cross-Entropy Loss, 2017. [Online]. Available: `https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html`

[64] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv preprint arXiv:1711.05101*, 2019.

[65] Stochastic Gradient Descent Clearly Explained, 2019. [Online]. Available: `https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31`

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.

[67] C. Yu, J. Wu, C. Song, H. Zhu, Z. Li, and J. Sun, "Exploring Plain Vision Transformer Backbones for Object Detection," *arXiv preprint arXiv:2203.16527*, 2022.

[68] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: A friendly introduction," *Journal of Big Data*, vol. 9, no. 102, 2022.

[69] J. Kim, K. Shim, J. Kim, and B. Shim, "Vision Transformer-based Feature Extraction for Generalized Zero-Shot Learning," *arXiv preprint arXiv:2302.00875*, 2023.

[70] A. A. Mukhlif, B. Al-Khateeb, and M. A. Mohammed, "Incorporating a Novel Dual Transfer Learning Approach for Medical Images," *Sensors*, vol. 23, no. 2, Art. no. 570, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/2/570

[71] A. Boukaache, N. E. Benhassine, and D. Boudjehem, "Breast Cancer Image Classification Using Convolutional Neural Networks (CNN) Models," *International Journal of Informatics and Applied Mathematics*, vol. 6, no. 2, pp. 20-34, 2023.

[72] S. Thirumalaisamy, K. Thangavilou, H. Rajadurai, O. Saidani, N. Alturki, S. K. Mathivanan, P. Jayagopal, and S. Gochhait, "Breast Cancer Classification Using Synthesized Deep Learning Model with Metaheuristic Optimization Algorithm," *Diagnostics*, vol. 13, no. 2925, 2023.

[73] B. Gheflati and H. Rivaz, "Vision Transformers for Classification of Breast Ultrasound Images," *arXiv preprint arXiv:2110.14731*, 2021.

[74] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "MedSegDiff-V2: Diffusion based Medical Image Segmentation with Transformer," *arXiv preprint arXiv:2301.11798*, 2023.

[75] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *arXiv preprint arXiv:2006.11239*, 2020.

[76] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 2019. DOI: https://doi.org/10.1016/j.media.2019.101552. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518308430

[77] J. Logan, P. J. Kennedy, and D. Catchpoole, "A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future," *Scientific Data*, vol. 10, no. 595, Sep. 2023. DOI: https://doi.org/10.1038/s41597-023-02430-6

[78] Cancer Research UK, 2024. [Online]. Available: https://www.cancerresearchuk.org

[79] Hologic, 2024. [Online]. Available: https://www.hologic.com

[80] M. Al-Balas, H. Al-Balas, Z. AlAmer, et al., "Clinical outcomes of screening and diagnostic mammography in a limited resource healthcare system," *BMC Women's Health*, vol. 24, Art. no. 191, Mar. 2024. DOI: https://doi.org/10.1186/s12905-024-03007-0

[81] Breast Cancer Screening, 2024. [Online]. Available: https://www.cancer.gov/types/breast/hp/breast-screening-pdq

[82] DICOM Standard, 2024. [Online]. Available: https://www.dicomstandard.org

[83] A. M. Reza, "Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 38, no. 1, pp. 35-44, Aug. 2004. DOI: 10.1023/B:VLSI.0000028532.53893.82

[84] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, pp. 452-454, May 2016. DOI: 10.1038/533452a. PMID: 27225100.

[85] O. E. Gundersen and S. Kjensmo, "State of the Art: Reproducibility in Artificial Intelligence," in *Proceedings of the 2018 Conference*, Feb. 2018.

[86] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle, "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)," *arXiv preprint arXiv:2003.12206*, 2020.

[87] Q. Nguyen, H.-B. Ly, H. Lanh, N. Al-Ansari, H. Le, T. V. Quan, I. Prakash, and T. Phm, "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," *Mathematical Problems in Engineering*, vol. 2021, Art. no. 4832864, Feb. 2021. DOI: 10.1155/2021/4832864

[88] L. Garrucho, K. Kushibar, S. Jouide, O. Diaz, L. Igual, and K. Lekadir, "Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study," *Artificial Intelligence in Medicine*, vol. 132, pp. 102386, 2022. DOI: https://doi.org/10.1016/j.artmed.2022.102386

[89] L. Maier-Hein, A. Reinke, P. Godau, et al., "Metrics reloaded: recommendations for image analysis validation," *Nature Methods*, vol. 21, pp. 195-212, Feb. 2024. DOI: https://doi.org/10.1038/s41592-023-02151-z

[90] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29-36, Apr. 1982. DOI: 10.1148/radiology.143.1.7063747. PMID: 7063747.

[91] T. Fawcett, "Introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861-874, Jun. 2006. DOI: 10.1016/j.patrec.2005.10.010

[92] M. S. A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, A. T. Azar, and A. Shaikh, "Enhancing Breast Cancer Detection and Classification Using Advanced Multi-Model Features and Ensemble Machine Learning Techniques," *Life*, vol. 13, no. 10, Art. no. 2093, 2023. DOI: https://doi.org/10.3390/life13102093

[93] S. Hashem, S. Habashy, W. Elakel, S. Raouf, G. Esmat, M. Eladawy, and M. Elhefnawi, "A simple multi-linear regression model for predicting fibrosis scores in chronic Egyptian hepatitis C virus patients," *International Journal of Bio-Technology and Research (IJBTR)*, vol. 4, pp. 37-46, Jun. 2014.

[94] ROC Curve and AUC: Evaluating Model Performance, 2023. [Online]. Available: https://medium.com/@ilyurek/roc-curve-and-auc-evaluating-model-performance-c2178008b02

[95] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009. DOI: https://doi.org/10.1016/j.ipm.2009.03.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457309000259

[96] S. Chaudhury, A. Krishna, S. Gupta, K. Sankaran, K. Sau, A. Raghuvanshi, and F. Sammy, "Effective Image Processing and Segmentation-Based Machine Learning Techniques for Diagnosis of Breast Cancer," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1-6, Apr. 2022. DOI: 10.1155/2022/6841334

[97] What is a Confusion Matrix in Machine Learning? The Model Evaluation Tool Explained, 2023. [Online]. Available: https://www.datacamp.com/tutorial/what-is-a-confusion-matrix-in-machine-learning

[98] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1, pp. 5979, Apr. 2022. DOI: 10.1038/s41598-022-09954-8. PMID: 35395867; PMCID: PMC8993826.

[99] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[100] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, pp. e0118432, Mar. 2015. DOI: 10.1371/journal.pone.0118432. PMID: 25738806; PMCID: PMC4349800.

[101] Making Sense of Sensitivity, Specificity and Predictive Value: A Guide for Patients, Clinicians and Policymakers, 2023. [Online]. Available: https://blogs.imperial.ac.uk/medical-centre/2023/10/10/making-sense-of-sensitivity-specificity-and-predictive-value-a-guide-for-patients-clinicians-and-policymakers/

[102] P. Olliaro and E. Torreele, "Managing the risks of making the wrong diagnosis: First, do no harm," *International Journal of Infectious Diseases*, vol. 106, pp. 382-385, May 2021. DOI: 10.1016/j.ijid.2021.04.004. PMID: 33845195; PMCID: PMC8752462.

[103] Keras, *Image classification with Vision Transformer*, 2023. [Online]. Available: https://keras.io/examples/vision/image_classification_with_vision_transformer/

[104] Hugging Face, *Vision Transformer (ViT)*, 2023. [Online]. Available: https://huggingface.co/transformers/model_doc/vit.html

[105] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual Transformers: Token-based Image Representation and Processing for Computer Vision," *arXiv preprint arXiv:2006.03677*, 2020.

[106] C. Yang, E. A. Fridgeirsson, J. A. Kors, et al., "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data," *Journal of Big Data*, vol. 11, no. 7, Jan. 2024. DOI: https://doi.org/10.1186/s40537-023-00857-7

[107] Use Weighted Loss Function to Solve Imbalanced Data Classification Problems, 2023. [Online]. Available: https://medium.com/@zergtant/use-weighted-loss-function-to-solve-imbalanced-data-classification-problems-749237f38b75

[108] R. Osuala, G. Skorupko, N. Lazrak, L. Garrucho, E. García, S. Joshi, S. Jouide, M. Rutherford, F. Prior, K. Kushibar, et al., "medigan: a Python library of pretrained generative models for medical image synthesis," *Journal of Medical Imaging*, vol. 10, no. 6, pp. 061403, 2023.

[109] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning," *Radiology: Artificial Intelligence*, vol. 0, no. ja, pp. e210315, 2023. DOI: https://doi.org/10.1148/ryai.210315

[110] B. Gheflati and H. Rivaz, "Vision Transformers for Classification of Breast Ultrasound Images," in *Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 480-483, 2022. DOI: 10.1109/EMBC48229.2022.9871809

[111] A Gentle Introduction to Threshold-Moving for Imbalanced Classification, 2021. [Online]. Available: https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/

[112] M. Shiri, M. P. Reddy, and J. Sun, "Supervised Contrastive Vision Transformer for Breast Histopathological Image Classification," *arXiv preprint arXiv:2404.11052v2*, Apr. 2024.

[113] W. Wang, R. Jiang, N. Cui, Q. Li, F. Yuan, and Z. Xiao, "Semi-supervised vision transformer with adaptive token sampling for breast cancer classification," *Frontiers in Pharmacology*, vol. 13, Art. no. 929755, Jul. 2022. DOI: 10.3389/fphar.2022.929755. PMID: 35935827; PMCID: PMC9353650.

[114] Pandas, 2024. [Online]. Available: https://pypi.org/project/pandas/

[115] NumPy, 2024. [Online]. Available: https://numpy.org

[116] Scikit-Learn, 2024. [Online]. Available: https://scikit-learn.org/stable/

[117] Pillow, 2024. [Online]. Available: https://pypi.org/project/pillow/

[118] OpenCV-Python, 2024. [Online]. Available: https://pypi.org/project/opencv-python/

[119] PyTorch, 2024. [Online]. Available: https://pytorch.org

[120] Torchvision, 2024. [Online]. Available: https://pytorch.org/vision/stable/index.html

[121] Matplotlib, 2024. [Online]. Available: https://matplotlib.org

[122] Seaborn, 2024. [Online]. Available: https://seaborn.pydata.org

[123] Vision Transformer, 2024. [Online]. Available: https://pytorch.org/vision/main/models/vision_transformer.html

[124] Tqdm, 2024. [Online]. Available: https://tqdm.github.io

[125] Weights Biases, 2024. [Online]. Available: https://docs.wandb.ai

Appendix A

Libraries and Frameworks

This appendix provides an overview of the libraries and frameworks used to develop, train, and test the ViT model.

Table A.1: Libraries and frameworks used for training the ViT model.

Library/Tool	**Version**	**Usage**
pandas [114]	2.2.1	Data analysis and manipulation
numpy [115]	1.26.4	Numerical operations
scikit-learn [116]	1.3.0	Data splitting, and evaluation metrics
Pillow [117]	10.3.0	Image processing
opencv-python [118]	4.9.0	Computer vision tasks
torch [119]	2.4.0.dev20240422	PyTorch library for DL
torchvision [120]	0.19.0.dev20240422	Image transformations and utilities
matplotlib [121]	3.8.4	Data visualization
seaborn [122]	0.13.2	Statistical data visualization
tqdm [124]	4.65.0	Progress bar for loops and operations
transformers [123]	4.40.2	ViT model
wandb [125]	0.17.0	Experiment tracking and logging

The development environment included Python 3.10.14, managed using a Conda environment, and the code was developed using Visual Studio Code 1.90.0.

Appendix B

Code and Model Weights

The code used to train and test the ViT model, along with the weights of the best model can be accessed through:

`https://github.com/xvxnoah/TFG-Noah-ViT-Breast-Cancer-Classifier`