FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A transfer learning approach for predicting breast cancer hormonal receptor status with MRI data

Ana Isabel Pontes Silva

Master in Electrical and Computer Engineering

Supervisor: Hélder Filipe Pinto de Oliveira, PhD

Co-Supervisor: Tânia Maria Pereira Lopes, PhD

February 28, 2023

Abstract

Breast cancer develops from breast tissue. Nowadays the diagnosis of breast cancer is confirmed by taking a biopsy of the concerning tissue, which is an invasive intervention. Creating virtual biopsies to overcome the limitations of invasive biopsies may be feasible due to recent technological advances that allow integration of imaging and molecular data.

Clinical MRI scanners today are capable of providing excellent temporal and spatial resolution, which allows extraction of numerous imaging features. Meanwhile, advances in breast cancer genetics research have resulted in the identification of promising genes associated with cancer outcomes. In addition, validated genomic signatures have been developed that allow categorization of breast cancers into distinct molecular subtypes as well as predict the risk of cancer recurrence and response to therapy.

This study proposes the use of a CNN, specifically the VGG16 model, to predict the presence of hormonal receptors using MRI images. Initially, the model was trained from scratch, however, it did not yield satisfactory results. Therefore, a transfer learning approach was also proposed, utilizing a pre-trained VGG16 model with the ImageNet dataset. This approach improved the model performance. The highest balanced binary accuracy on test for predicting the presence of the HER2 hormonal receptor was 58%, as for the PR hormonal receptor, the highest balanced binary accuracy was 53%. In both cases data augmentation was used. One possible cause of this low model performance may be related to the use of an unbalanced dataset, and a pre-trained model with ImageNet.

Resumo

O cancro da mama desenvolve-se a partir do tecido mamário. Actualmente o diagnóstico de cancro da mama é confirmado através da realização de uma biópsia do tecido em questão, que é uma intervenção invasiva. A criação de biópsias virtuais para ultrapassar as limitações das biópsias invasivas pode ser viável devido aos recentes avanços tecnológicos que permitem a integração de imagens e dados moleculares.

Os scanners de MRI são hoje capazes de proporcionar uma excelente resolução temporal e espacial, o que permite a extracção de numerosas características da imagem. Entretanto, os avanços na investigação genética do cancro da mama resultaram na identificação de genes promissores associados aos resultados do cancro. Além disso, foram desenvolvidas assinaturas genómicas validadas que permitem categorizar o cancro da mama em subtipos moleculares distintos, bem como prever o risco de reaparecimentto do cancro e a resposta à terapia.

Este estudo propõe a utilização de uma CNN, especificamente o modelo VGG16, para prever a presença de receptores hormonais utilizando imagens de ressonância magnética. Inicialmente, o modelo foi treinado de raiz, no entanto, não produziu resultados satisfatórios. Por conseguinte, foi também proposta uma abordagem de transfer learning, utilizando um modelo VGG16 pré-treinado com o dataset ImageNet. Esta abordagem melhorou o desempenho do modelo. A maior precisão binária equilibrada no teste para prever a presença do receptor hormonal HER2 foi de 58%, e para o receptor hormonal PR, a maior precisão binária equilibrada foi de 53%. Em ambos os casos, foi utilizado o aumento de dados. Uma possível causa deste baixo desempenho do modelo pode estar relacionada com a utilização de um conjunto de dados desequilibrado, e de um modelo pré-treinado com ImageNet.

Acknowledgments

I would like to express my gratitude for being able to integrate the Visual Computing and Machine Intelligence Research Group of the Centre for Telecommunications and Multimedia at INESC TEC, to develop my thesis. I had the opportunity to meet people that contributed to this work. To my supervisors, Prof. Dr. Hélder Filipe Pinto de Oliveira, and Dr. Tânia Maria Pereira Lopes, I would like to express my gratitude for all the support and patience throughout this work. Despite not being part of the supervising team, I would also like to express my deepest gratitude to MSc Francisco Silva, for all the knowledge and support that he gave me during my work, and all the patience.

To all the professors that taught me during these years, I would like to express my gratitude for the knowledge that they gave and for all the challenges that I had to pass through so that I could grow up as a student and as a person.

To all my friends that I was lucky enough to meet in Faculty of Engineering of the University of Porto, and who have been by my side through all these 5 + 1 years. I will never forget all the precious moments that we had together, since 2016, when we first met. I hope to continue to build memories with them in the following years.

To my dear boyfriend who never stopped believing in my potential and my capacities, even when I just wanted to give up. Thank you for always being by my side, to help me stand up in the worst times and to always have the right words that I needed to hear and the comfort.

And last but not least I would like to express my gratitude and love to my family, especially my parents, for all the support, sacrifices, and love they gave through these years, so that I can be here today. A last thank you to my brother, who also gave me his support and advices when I most needed them.

Finally I apologize all other unnamed people who helped me in my life. Any omission in this brief acknowledgment does not mean lack of gratitude.

Ana Isabel Pontes Silva

"A person who never made a mistake never tried anything new."

Albert Einstein

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 2
- 1.3 Objectives . . . 3
- 1.4 Report Structure . . . 4

2 Background **5**
- 2.1 Breast Cancer Pathology . . . 5
  - 2.1.1 Histological Subtypes . . . 5
  - 2.1.2 Breast Cancer Classification . . . 6
- 2.2 Magnetic Resonance Imaging . . . 7
- 2.3 Summary . . . 8

3 Literature Review **9**
- 3.1 Machine Learning and Computer Vision . . . 9
- 3.2 Deep Learning . . . 11
  - 3.2.1 Convolutional Neural Networks . . . 11
- 3.3 Summary . . . 15

4 Methodology **17**
- 4.1 Dataset . . . 17
- 4.2 Data Preparation . . . 18
  - 4.2.1 DICOM images pre-processing . . . 19
  - 4.2.2 Data Augmentation . . . 21
  - 4.2.3 Data Normalization . . . 22
- 4.3 Solution implementation . . . 22
  - 4.3.1 Environment . . . 23
  - 4.3.2 Training from Scratch using VGG16 . . . 23
  - 4.3.3 Transfer Learning using VGG16 . . . 25
- 4.4 Summary . . . 26

5 Results and Discussion **29**
- 5.1 Training from Scratch . . . 29
  - 5.1.1 Data augmentation . . . 29
  - 5.1.2 Dropout with PR hormonal receptor . . . 30
  - 5.1.3 Single patient slice with HER2 hormonal receptor . . . 32
- 5.2 Transfer Learning . . . 35
  - 5.2.1 Using weight decay in hormonal receptor HER2 and PR . . . 35

5.3 Summary 37

6 Conclusions and Future Work **39**

References **41**

List of Figures

1.1 Estimated number of new cases of cancer in 2020 . . . 2
1.2 A fifty-five-year-old breast from woman with microcalcifications in screening mammogram . . . 3
1.3 An example of using MRI functional tumor volume (FTV) to monitor a breast cancer patient during neoadjuvant chemotherapy . . . 4

2.1 Molecular subtypes of breast cancer . . . 6
2.2 Therapeutic strategies for early breast cancer, according to molecular subtype . . . 7

3.1 A CNN sequence to classify handwritten digits . . . 12
3.2 Network architecture of a CNN . . . 14

4.1 Distribution of patients according to MRI Manufacturer . . . 18
4.2 DICOM from MRI from Duke University dataset . . . 19
4.3 Box graph for values of row and column . . . 20
4.4 Crop of 200x170 from DICOM image . . . 20
4.5 Crop 100x100 in MRI from a patient . . . 21
4.6 Data augmentation on patient with ID 132 . . . 22
4.7 Vgg16 model with cropped image as input . . . 24
4.8 Exemplification RGB image transformation . . . 26
4.9 Cross Validation used . . . 27

5.1 Results for PR, with augmentation, but with different values for batch size and Learning rate . . . 31
5.2 Loos and accuracy with 1 dropout with PR hormonal receptor . . . 33
5.3 Loos and accuracy with 2 dropouts with PR hormonal receptor . . . 33
5.4 Loos and accuracy with 3 dropouts with PR hormonal receptor . . . 33
5.5 Loos and accuracy in experiment with only one slice in hormonal receptor HER2 34
5.6 Loos and accuracy in experiment with only one slice and with data augmentation in hormonal receptor HER2 . . . 34
5.7 Loos and accuracy in experiment with transfer learning on hormonal receptor PR 38
5.8 Loos and accuracy in experiment with transfer learning on hormonal receptor HER2 38

List of Tables

2.1 Classification of molecular subtypes of breast cancer. 6

3.1 Classification results of the deep neural networks for genomics, radiomics, and radiogenomics data. Performance measurement is the Area Under Receiver Operating Characteristics (AUC) curves. [1] 13

4.1 Annotation of lesions file 18
4.2 Distribution of patients according to breast cancer hormonal receptor status . . . 21
4.3 Hyper parameters tested during model training 25
4.4 Distribution of images according to label, for HER2, PR and ER 25
4.5 Cross validation using 5-fold, with distribution of images for train, validation and test in label HER2 27
4.6 Cross validation using 5-fold, with distribution of images for train, validation and test in label PR 27
4.7 Cross validation using 5-fold, with distribution of images for train, validation and test in label ER 27

5.1 Test set Balanced binary accuracy results for the PR hormonal receptor 29
5.2 Test set Balanced binary accuracy results for the HER2 hormonal receptor 30
5.3 Test set Balanced binary accuracy results for the ER hormonal receptor 30
5.4 Results for balanced binary after experiments with dropout in hormonal receptor PR 32
5.5 Balanced binary accuracy in five experiments for the PR hormonal receptor . . . 35
5.6 Binary accuracy in five experiments for the HER2 hormonal receptor 36
5.7 Mean performance on all the experiments done, with PR and HER2 hormonal receptors 37

Abbreviations and Symbols

AI	Artificial Intelligence
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic curve
BRCA	BReast CAncer gene
CAD	Computer-aided Diagnosis
CI	Confidence Interval
CNN	Convolutional Neural Network
DCE-MRI	Dynamic Contrast-enhanced Magnetic Resonance Imaging
DCIS	Ductal Carcinoma In Situ
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DNN	Deep Neural Networks
ER	Estrogen Receptor
FISH	Fluorescence in situ hybridization
FN	False Negatives
FNA	Fine Needle Aspiration
FP	False Positives
GSEA	Gene Set Enrichment Analysis
HER2	Human Epidermal growth factor Receptor 2
IHC	Immunohistochemistry
ILC	Invasive Lobular Carcinoma
LCIS	Lobular Carcinoma In Situ
LR	Learning Rate
NST	No Special Type
ML	Machine Learning
MRI	Magnetic Resonance Imaging
pCR	pathological complete response
PR	Progesterone Receptor
ReLU	Rectified Linear Unit
RGB	Red, green, blue
ROC	Receiver Operating Curve
SVMs	Suport Vector machines
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive
TN	True Negative
TNBC	Triple-Negative breast cancer
TP	True Positives
VGG	Visual Geometry Group
WHO	World Health Organization

Chapter 1

Introduction

1.1 Context

Cancer is a generic term for a large group of diseases that can affect any part of the body. Other terms used are malignant tumors and neoplasms. The main feature of cancer is the abnormal proliferation of cells. Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 [2]. The most common cases of cancer in 2020 were: breast, lung, colorectum, prostate, skin (non-melanoma) and stomach (see in Figure 2.1).

In 2020, there were 2.3 million women worldwide diagnosed with breast cancer. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer. The World Health Organization (WHO) estimates that by the end of the year 2030, more 400 000 cases of breast cancer in both sexes and on all ages will be diagnosed, compared to 2020 [4]. Breast cancer is the most common type of cancer among women, and it is the second leading cause of cancer death in women [5]. Breast cancer is one of the diseases with the greatest impact on our society, not only because it is very common, and associated with a very serious image, but also because it attacks an organ full of symbolism, in motherhood and femininity [5]. When identified early, cancer is more likely to respond to treatment and can result in a greater probability of survival, as well as less expensive treatment [2].

Breast cancer detection can be done through self examination, clinical exams and imaging exams. A clinical exam may include, a doctor examination, and lab analysis. The doctor palpates the breasts in different positions: while standing, sitting and while lying down [6]. The doctor may ask the patient to raise the arms above the head, let them hang down, or press the hands against the haunch. The doctor looks for any differences between the breasts, including any unusual differences in size or shape. On the skin, the presence of redness, skin depressions or other abnormal signs is checked. Nipples should be pressed to check for any secretion or fluid leakage. In Portugal, every two years, an invitation is made to all women between 50 and 69 years old, enrolled in health centres, to have a free mammogram for breast cancer screening [7]. Mammogram often shows a lump in the breast before it can be felt or palpated. It may also show an aggregation of

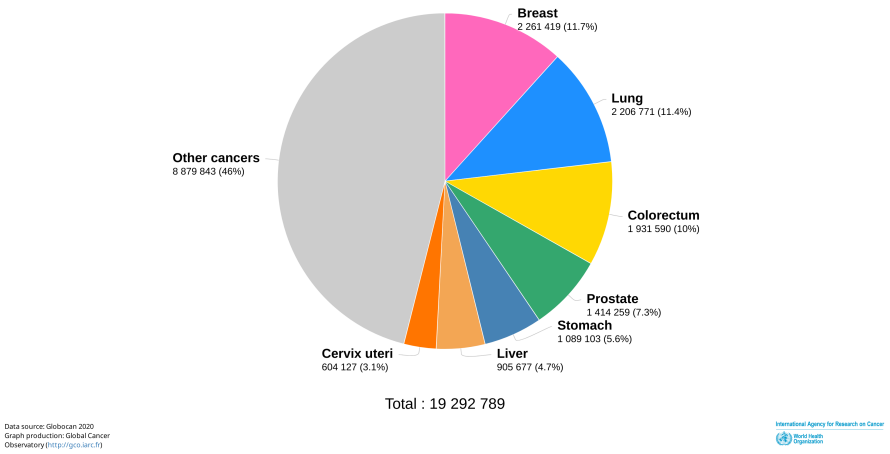


Figure 1.1: Estimated number of new cases of cancer in 2020, worldwide, both sexes, all ages [3].

small calcium particles [6] which are defined as microcalcifications, (see in Figure 1.2) and both can be signs of cancer. The mammogram to the left, in Figure 1.2 a), demonstrates clusters of microcalcifications with overlapping glandular tissue, with these being more conspicuous in the image on the right 1.2 b). This radiological exam uses small doses of radiation and if there is any suspicion of breast cancer, an anatomohistopathological characterization biopsy of the breast tissue is needed. If the radiologist has any doubts about the diagnosis, complementary exams are made, such as Magnetic Resonance Imaging (MRI).

The MRI uses magnetic fields and radio waves to produce detailed images of the inside the body. It may be used when patients have a family history of breast cancer, BRCA (Breast Cancer gene) mutations, breast implants, lobular cancers, if multiple tumors are suspected or if the results of other imaging techniques are inconclusive. MRI is also used to see if a tumor has responded to treatment and to plan next therapy. In Figure 1.3 is an example of breast cancer monitoring, for a patient during neoadjuvant chemotherapy.

Treatment of breast cancer often consists of a combination of surgical removal, radiation therapy and medication (e.g., hormonal therapy, chemotherapy and/or targeted biological therapy) to treat the cancer that has spread from the breast tumor through the blood.

1.2 Motivation

There are a lot of cancers that can be cured if detected early and treated effectively. To characterize and to confirm that a tumor is malignant, a biopsy is needed. There are several types of biopsies procedures, that can be done depending on tumor location and size, or the area of concern: 1) Fine needle aspiration (FNA) biopsy; 2) Core needle biopsy; 3) Open (surgical) biopsy; 4) Stereotactic biopsy; 5) Vacuum-assisted core biopsy; and 6) Ultrasound-guided biopsy. In those procedures a doctor removes small pieces from the suspicious area for later analysis in the laboratory to check the existence of cancer cells [9]. Nevertheless, biopsies are an invasive procedure

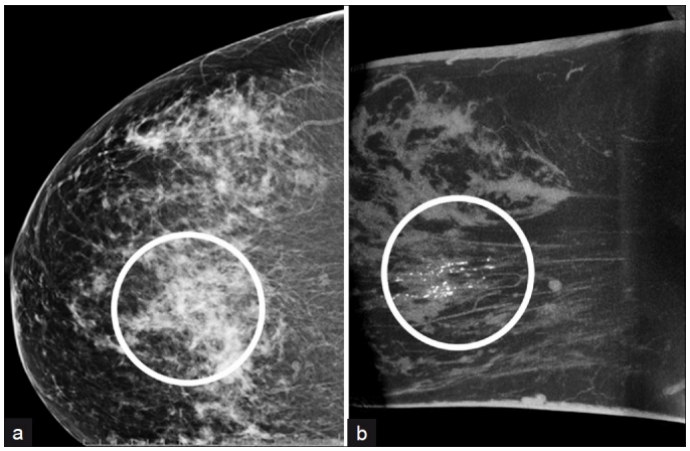


Figure 1.2: A fifty-five-year-old breast from woman with microcalcifications in screening mammogram [8].

and have some limitations. During and after the biopsy procedure, the patient may have pain, bruising, and bleeding at the biopsy site.

To overcome this, the classification of breast cancer may be possible through image analysis with the use of Convolutional Neural Netwoks (CNN), to classify the images obtained from the MRI and consequently the different types of tumor, since treatment is also different according to the tumor detected. The presence or not of the different breast cancer hormonal receptors Estrogen Receptor (ER), Progesterone Receptor (PR) and the Human Epidermal growth factor Receptor 2 (HER2) may reveal different features and so the different types of molecular subtypes could be identified. The use of MRI to do this predictions is fundamental, since it is a noninvasive procedure and does not uses radiation to produce the images, contrary to the mammogram. The high quality of tissue, can differentiate better than mammogram between fat, muscle, etc, so we can extract features. MRI features may have the possibility to describe tumor characteristics that are the result of biologic processes in the tumor and hence can aid in investigating the mechanisms that drive tumor progression or therapy resistance [10].

1.3 Objectives

The main goal of this thesis is to develop predictive models that relate MRI scan image features (phenotypes), with genotype signatures based on retrospective databases (radiogenomics approach).

To achieve this goal, several objectives have been defined:

- Identify public datasets that are suited for this work;
- Identify state of the art solutions, and its limitations;
- Develop our own solution and compare it with existing ones.

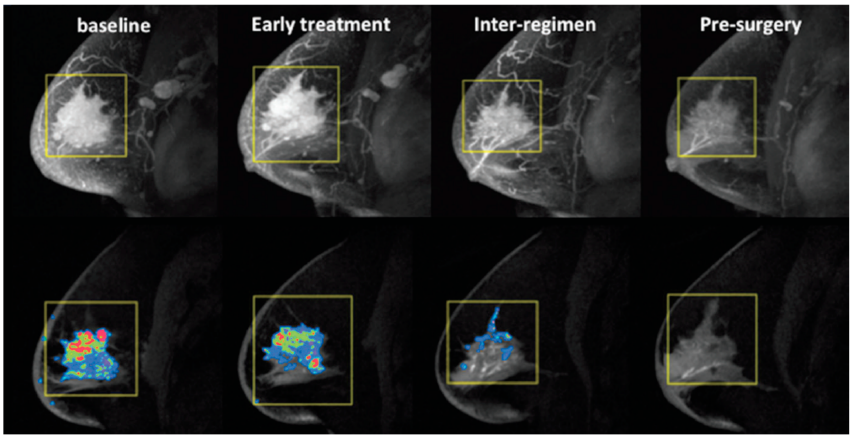


Figure 1.3: An example of using MRI functional tumor volume (FTV) to monitor a breast cancer patient during neoadjuvant chemotherapy. From left to right, FTV measurements were 48.5, 35.4, 5.6, and 0 cm3 at baseline, early treatment, inter-regimen, and pre-surgery time points, respectively [11].

With this objectives in mind we developed our solution, to predict the hormonal receptors of breast cancer tumors, with Magnetic Resonance Imaging, using deep learning techniques, in order to improve the quality of life of patients, and also make the treatment and diagnosis of this disease efficient.

1.4 Report Structure

This document is structured in five chapters. Chapter 2 focuses in the Background. Chapter 3 presents the Literature Review. Chapter 4 describes the methodology used for the development of this thesis. Chapter 5 presents the results and discussion for all the work done. Chapter 6 presents the main conclusions and provides future work directions.

Chapter 2

Background

In this chapter, a clinical background is given. First, a clinical view, about how breast cancer can be classified. Then a screening approach about how this malignant cells can be seen on the breast. Finally, an overview of the chapter is provided in the summary section.

2.1 Breast Cancer Pathology

Personalized medicine for the treatment of cancer was conceived more than 30 years ago [12] and exemplified, in the context of breast cancer, by the discovery of the ER and its use as a target therapeutic. Subsequently, the fundamental role of HER2 was discovered and the development of therapeutic strategies that have changed radically the prognosis of patients with breast cancer [13].

2.1.1 Histological Subtypes

The breast is a gland that produces milk. For this, the glandular part of the breast consists of **lobules** and **ducts**. The most common type of breast cancer is carcinoma arising from epithelial cells [14]. The breast epithelium is composed of two known cell types, an inner layer of secretory luminal cells and an outer layer of basal/myoepithelial cells. This tissue is constantly being renewed through cell division and is more often exposed to carcinogens in the environment. Breast cancer can be classified as:

- **ductal carcinoma** - the most frequent, originating from the cells of the ducts;
- **lobular carcinoma** - carcinoma originating from lobular cells is called lobular.

The cells of the ducts and lobules are separated from the other cells of the breast by a membrane, called the basement membrane, since it is at the base of these cells. Malignant cells, that proliferate in a disorganized way without dying, and acquire the ability to invade, are capable of going beyond this basement membrane [15]. When this happens, carcinomas are referred to as invasive ductal carcinoma – No Special Type(NST) – or lobular carcinoma (ILC). If the cancer has not yet crossed the basement membrane, it is called carcinoma in situ or preinvasive – ductal carcinoma in situ or lobular carcinoma in situ – because it is "in its anatomical site". Figure 2.1 explains

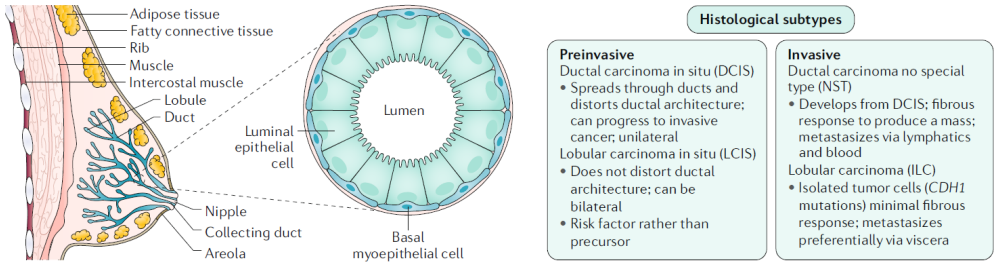


Figure 2.1: The histological subtypes described in this picture are the most frequent subtypes of breast cancer: ductal carcinoma (now referred to as NST) and lobular carcinoma are the invasive lesions. Their preinvasive counterparts are ductal carcinoma in situ and lobular carcinoma in situ (or lobular neoplasia), respectively [16].

where this two types of breast cancer are localized, and provides a more detailed explanation of the histological subtypes.

2.1.2 Breast Cancer Classification

The surrogate intrinsic subtypes are typically used clinically and are based on histology expression of key proteins: **ER**, **PR**, **HER2** and the proliferation marker **Ki67**. Table 2.1 presents the breast cancer classification, according to the biomarkers. We currently distinguish three major biological subtypes of breast cancer. Tumors expressing ER and/or PR are termed **'hormone receptor- positive'**; tumors not expressing ER, PR and HER2 are called **'triple- negative'(TNBC)**, and tumor expressing only HER2 positive are called **'HER2 positive'**.

The molecular subtypes of cancer, can be discovered with the help of biopsy, where a piece of tissue or liquid from the breast is analysed, as explained in section 1.2. After the type of molecular subtype is discovered, the best treatment approach is chosen accordingly. Figure 2.2, summarises the therapeutic strategies for early breast cancer. Management of the early breast cancer is based on tumor burden and subtype. All patients with ER-positive disease receive adjuvant endocrine therapy after surgery. If patients are at high risk of recurrence (for example, owing to high-risk gene expression signature results with 0 to 3 involved lymph nodes, involvement of $\geq$ 4 lymph

Molecular Subtypes	TNBC	Hormonal Receptor Positive			HER2 positive
		Luminal A	Luminal B (HER2-)	Luminal B (HER2+)	
Biomarkers	ER- PR- HER2- Ki67 high	ER+ PR+ HER2- Ki67 low	ER+ PR- HER2- Ki67 high	ER+ PR+/- HER2+ Ki67 low/high	ER- PR- HER2+ Ki67 high

Table 2.1: Classification of molecular subtypes of breast cancer.

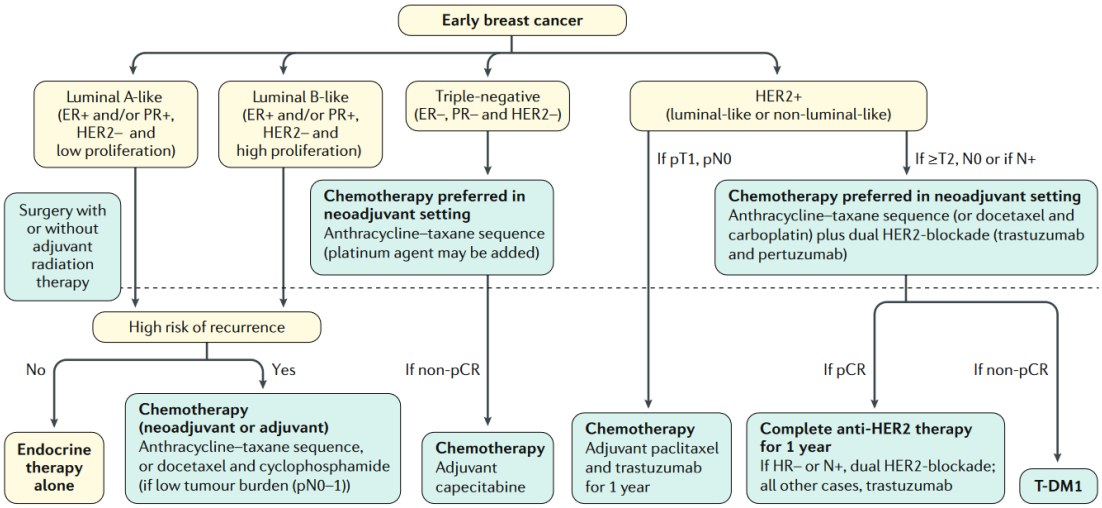


Figure 2.2: Therapeutic strategies for early breast cancer, according to molecular subtype [16].

nodes or a $>$ 10% risk of breast cancer-specific mortality at 10 years), chemotherapy needs to be recommended as well. In triple-negative and HER2 positive early breast cancer, neoadjuvant subtype-specific systemic therapy is standard, followed by surgery. In the case that pathological complete response (pCR) is not achieved, systemic therapy can be escalated. Bisphosphonates are an additional adjuvant therapy option for all postmenopausal patients and premenopausal patients receiving ovarian suppression; they also conserve bone density. If indicated, radiation therapy can be administered after surgery. The management algorithm takes evidence-based registered therapy options into account. Availability and reimbursement of individual diagnostic or therapeutic options may differ regionally and require adjustments of the treatment concepts outlined here.

2.2 Magnetic Resonance Imaging

MRI is a non-invasive imaging technology that produces three dimensional detailed anatomical images [17]. It is often used for disease diagnosis, and treatment monitoring. It is based on sophisticated technology that excites and detects the change in the direction of the rotational axis of protons found in the water that makes up living tissues. MRI images employ powerful magnets which produce a strong magnetic field that forces protons in the body to align with that field [17]. When a radiofrequency current is then pulsed through the patient, the protons are stimulated, and spin out of equilibrium, straining against the pull of the magnetic field.

Preoperative breast MRI has become increasingly prevalent as its role in aiding diagnosis and treatment planning of breast cancer [18]. Adding MRI to mammogram increases screening sensitivity in women with BRCA1 and/or BRCA2 mutations (testing for inherited BRCA1 and BRCA2 variants may be done using a blood or saliva sample [19]) and is the recommended screening approach for BRCA mutation carriers and for women at substantially increased lifetime risk of breast cancer. Imaging evaluation also includes MRI for specific clinical indications, such as in women

for whom conventional imaging tests have been equivocal, inconclusive or discordant, for evaluating women with breast implants and for evaluating women with axillary nodal metastases but no detectable (occult) breast tumor [16]. Preoperative MRI is also selectively used for staging newly diagnosed disease, but this is a debated practice given the limited evidence on whether it enhances a patient's clinical outcomes. However, MRI is advised for preoperative assessment of newly diagnosed invasive lobular cancers.

Allying MRI with deep learning approaches, may be a promising area of investigation. MRI is a non-envasive and a high fidelity method that improves the quality of life of breast cancer patients. The use of CNNs applied to MRI images may allow for more accurate and efficient diagnosis of medical conditions. This can help radiologists diagnose more accurately and confidently. Moreover, it can also aid in the development of new treatments. A lot of studies have been done in this area, and will be reviewed in Chapter 3.

2.3 Summary

This chapter presented the clinical background necessary to understand this dissertation. Section 2.1 explains the biology of breast carcinoma, that will be fundamental for this study. Section 2.2 presented the type of image that will be used in this dissertation, why we chose to use this type of screening and why is so important for treatment and diagnosis of breast cancer.

Chapter 3

Literature Review

This chapter contains the literature review, regarding the topic and technologies that will be used for predicting the hormonal receptor status of breast cancer. In the following sections, the methodologies and technologies used regarding the use of artificial intelligence in the detection of breast cancer are explained. First a brief introduction is made to the technologies used, and then the studies that used the technology are presented. Finally, the main conclusions of the related works and also their biggest limitations are presented.

3.1 Machine Learning and Computer Vision

Machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment. They are considered the working horse in the new era of the so-called big data [20]. Techniques based on machine learning have been applied successfully in diverse fields ranging from pattern recognition, computer vision, spacecraft engineering, finance, entertainment, and computational biology to biomedical and medical applications. There are many ways that a computational algorithm can adapt itself in response to training. The input data can be selected and weighed to provide the most decisive outcomes. The algorithm can have varied numerical parameters that are adjusted through iterative optimization.

The ideal of machine learning is to emulate the way that human beings learn to process sensory signals in order to accomplish a goal. This goal could be a task in pattern recognition, in which the learner wants to distinguish apples from oranges. Every apple and orange is unique, but we are still able to tell one from the other. Rather than hard code a machine, can be programmed to learn to distinguish them through repeated experience with actual apples and oranges. This is a good example of **supervised learning**, in which each training example of input data (color, shape, odor, etc.) is paired with its known classification label (apple or orange). It allows the learner to deal with similarities and differences when the objects to be classified have many variable properties within their own classes but still have fundamental qualities that identify them [20].

Another type of machine learning is **unsupervised learning**. This type of machine learning involves algorithms that train on unlabeled data [21]. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined. Unsupervised learning algorithms are good for the following tasks: clustering; anomaly detection; association mining and dimensionality reduction. Two more types of ML are **semi-supervised learning** and **reinforcement learning**. The first approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set [21]. The last one is typically used to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task [21]. But for the most part, the algorithm decides on its own what steps to take along the way.

Computer vision can be leveraged in support of histological findings by inferring diagnosis from the structural heterogeneity observed within tumours on medical imaging. Furthermore, given the high frequency that medical imaging is performed in cancer management, medical imaging data can be used by computer vision applications to better characterize disease, predict progression at the time of diagnosis and monitor response to treatment [22].

In 2014, Maciej A. Mazurowski et al. [23] investigated associations between breast cancer molecular subtype and semi automatically extracted MRI features. Computer vision algorithms were applied to extract 23 imaging features from lesions indicated by a breast radiologist on MRI images from a public dataset in The Cancer Imaging Archive (TCIA) [24]. Morphological, textural, and dynamic features were extracted, and the molecular subtype was determined on the basis of genomic analysis. Associations between the imaging features and molecular subtype were evaluated by using logistic regression and likelihood ratio tests. They excluded the normal-like subtype from this analysis because only one subject belonged in this group. An association between the imaging features and the luminal B subtype was found but not between the imaging features and other subtypes. Because of this the study was focused on the analysis of the relationship between imaging features and the luminal B subtype. This work showed that computer-assisted extraction of image features can be used to help identify breast cancer molecular subtypes. Substitution of imaging features for genetic analysis may provide a prognostic benefit with no additional cost or delay in treatment planning. The main limitation for this study was the sample size of 48 patients, with only eight patients in the luminal B subtype.

A year later, in 2015, Lars J. Grimm et al. [25] used a private dataset to analysed routine clinical pre-operative breast MRI images from 275 breast cancer patients at a single institution. Computer vision algorithms were then used to extract 56 imaging features from the cancers including morphological, textural, and dynamic features. The models showed that the imaging features were able to predict luminal A and luminal B molecular subtype, but not HER2 or basal molecular subtype. The exploratory analysis of individual features showed that there were two imaging features that were each able to predict both luminal A and luminal B subtype. There were other

features that showed an association with individual subtypes but no other ones that were predictive of both luminal A and luminal B. The main limitation for this study is the lack of control over which patients underwent MRI.

Ashirbani Saha et al. [26] analysed in 2018 a set of 922 patients with invasive breast cancer and pre-operative MRI, from a public dataset available in TCIA. The MRI images were analysed by a computer algorithm to extract 529 features of the tumour and the surrounding tissue. Machine learning models based on the imaging features were trained using a portion of the data (461 patients) to predict the following characteristics: breast cancer hormonal receptors, ER, PR and HER2 factor status, as well as a tumour proliferation marker (Ki-67). Multivariate models were predictive of Luminal A subtype from other subtypes with AUC = 0.697 (95% Confidence Interval(CI): 0.647–0.746), triple negative breast cancer from the other subtypes with AUC = 0.654 (95% CI: 0.589–0.727), ER status with AUC = 0.649 (95% CI: 0.591–0.705), and PR status with AUC = 0.622 (95% CI: 0.569–0.674). The higher values of AUCs were obtained for discriminating HER2 molecular subtype versus others and Luminal A versus other subtypes. This study had some limitations: while the heterogeneous cohort used in this study allows for more generalized conclusions, it is likely that variability in imaging acquisition parameters were a significant source of noise in this analysis and stronger associations could be found in a cohort with uniform imaging parameters. Future analysis could be conducted to evaluate this issue when a larger number of patients scanned in a uniform manner are available.

3.2 Deep Learning

Deep learning (DL) is an important element of data science, which includes statistics and predictive modeling. It is extremely beneficial to data scientists who are tasked with collecting, analyzing and interpreting large amounts of data. Deep learning makes this process faster and easier. Various methods can be used to create deep learning models.

3.2.1 Convolutional Neural Networks

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other [27]. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex [27]. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlap to cover the entire visual area. An example, of a CNN is shown in Figure 3.1, where the CNN receives an input image of a number, and then attempts to guess the number written in the output.

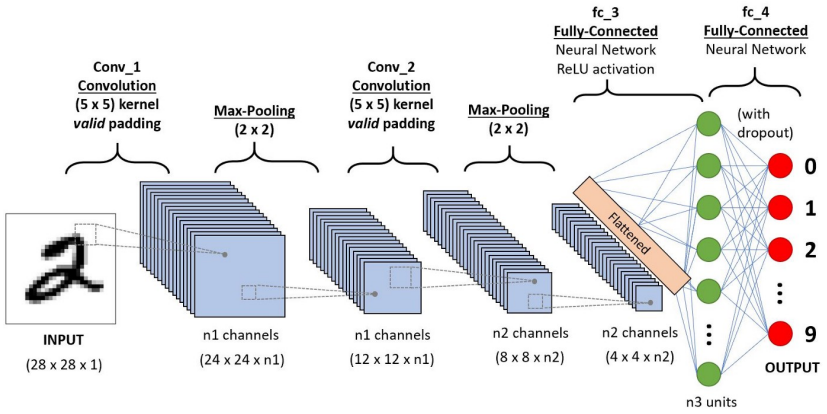


Figure 3.1: A CNN sequence to classify handwritten digits [27].

A CNN uses a system much like a multi layer perception that has been designed for reduced processing requirements. The layers of a CNN consist of an input layer, an output layer and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers and normalization layers [28].

There are various architectures of CNNs available which have been key in building algorithms. Some of these architectures are: LeNet, AlexNet, ResNet, VGGNet, GoogleNet and ZFNet.

Zhe Zhu et al. [29] studied the molecular subtype classification of breast cancer using deep learning applied to DCE-MRI, from a public dataset in TCIA. Three different deep learning approaches were investigated, and among them **off-the-shelf deep features** approach performed best. The highest AUC obtained was 0.65 using pre-trained GoogleNet as feature extractor and a polynomial kernel SVM trained on the extracted features. The highest AUC obtained for **training from scratch** was 0.58, which is only slightly higher than random guess. The performance of **transfer learning** is worse than off-the-shelf deep features but better than training from scratch. The most significant limitations in this study were the low number of cases in the context of deep learning, and considering only the distinction between the Luminal A subtype versus all other subtypes, as opposed to considering each molecular subtypes individually.

In 2018, Hong-Jun Yoon et al. [1] designed deep neural networks (DNN), which combined both radiomic and genomic features to predict pathological stage and molecular receptor status of invasive breast cancer patients. Utilizing imaging data from The Cancer Imaging Archive (TCIA) and gene expression data from The Cancer Genome Atlas (TCGA), they evaluated the predictive power of Convolutional Neural Networks. The focus was to predict pathological stage and molecular receptor status, including ER, PR, HER2; three very important factors of cancer prognosis and treatment response as explained in chapter 2.1. They investigated the potential of DL for the above predictive tasks. DL algorithms are well established as powerful feature learners with comparative and often superior performance compared to conventional machine learning algorithms, while requiring little or no feature engineering overhead. The main limitation for this study was

	Genomics	Radiomics	Radiogenomics
Stage	0.5617 ± 0.0311	0.6605 ± 0.0288	0.6966 ± 0.0262
ER	0.9017 ± 0.0242	0.8246 ± 0.0278	0.9766 ± 0.0109
PR	0.7811 ± 0.0302	0.7571 ± 0.0333	0.8992 ± 0.0226
HER2	0.8083 ± 0.0431	0.7226 ± 0.0275	0.8428 ± 0.0367

Table 3.1: Classification results of the deep neural networks for genomics, radiomics, and radiogenomics data. Performance measurement is the Area Under Receiver Operating Characteristics (AUC) curves. [1]

the limited number of cases available for training the proposed 3D CNN.

Richard Ha et al. [30] in 2019, tried to develop a CNN algorithm that can predict the molecular subtype of a breast cancer based on MRI features from a private dataset. First post-contrast MRI images were used for 3D segmentation using 3D slicer. A CNN architecture was designed with 14 layers. Extensive regularization was utilized including dropout, L2, feature map dropout, and transition layers. Seventy-four luminal A, 106 luminal B, 13 HER2+, and 23 basal breast tumors were evaluated. Testing set accuracy was measured at 70%. The class normalized macro area under receiver operating curve (ROC) was measured at 0.853. Non-normalized microaggregated AUC was measured at 0.871, representing improved discriminatory power for the highly represented Luminal A and Luminal B subtypes. One limitation in this study was that the patients in this study underwent MRI at different magnetic field strengths (1.5 or 3.0 T), potentially affecting the image quality, and another limitation is the small sample size used for the study.

In their study [1], Hong-Jun Yoon et al. performed experiments on MRI from 118 patients, provided by TCIA. It was utilized sagittal axis, T1-weighted images of both pre-contrast and post-contrast enhanced images, depending on availability. After this, implemented a CNN using Keras, and TensorFlow backend on Python environment, and ran the experiments on the Summitdev computers in Oak Ridge Leadership Computing Facility (OLCF). A hyper-parameter optimization technique based on random search strategy was utilized to determine the optimal number of convolution filters, number of fully-connected nodes, and number of nodes for each one of the predictive tasks.

To avoid intra-patient bias, patient-level 10-fold cross validation tests were performed. To increase the number of training samples, they augmented the training data by flipping images horizontally and/or vertically, and/or reversing the order of slides, resulting in 1,400 to 1,500 training samples for the 3-D CNN. Results of the classification tasks are listed in Table 3.1 taking as measure, the form of Area Under Receiver Operating Characteristics (AUROC) curves. Integrating both radiomic and genomic information boosted performance even further, compared to radiomic and genomic analysis, achieving the highest AUC scores for predicting pathological stage, ER, PR, and HER2 status.

Also in 2019, Richard Ha et al. [30] implemented a CNN, by a series of 3×3 convolutional kernels to maximize computational efficiency. Software code was written in Python using the

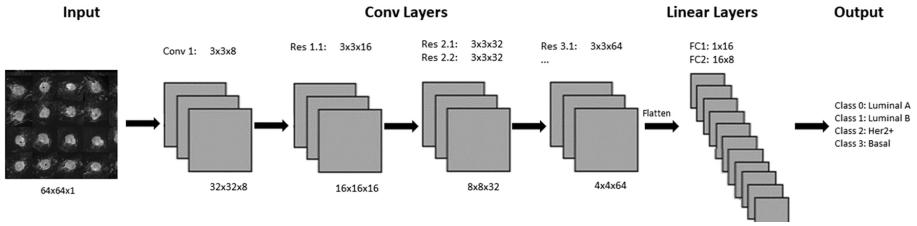


Figure 3.2: Network architecture: Dimensions of all of the intermediate layers of the convolutional neural network [30].

TensorFlow module on a Linux workstation. After an initial standard convolutional layer, a series of residual layers are utilized in the network. Downsampling of feature map size was implemented by means of a concatenated average and max pooling operation to decrease size by 75%. All nonlinear functions utilize the rectified linear unit (ReLU) which allows training of deep neural networks by stabilizing gradients on backpropagation. Additionally, batch normalization was used between the convolutional and ReLU layers to enhance network training by stabilizing the loss landscape. Figure 3.2 illustrates the network architecture used for the study.

Testing set accuracy was measured at 70%. The class normalized macro area under ROC was measured at 0.853. Non-normalized AUC was measured at 0.871, representing improved discriminatory power for the highly represented Luminal A and Luminal B subtypes. Aggregate sensitivity and specificity was measured at 0.603 and 0.958. One limitation for this work, was the reduced number of data, if using a larger dataset the potential to rival accuracy of standard genetic testing in determining breast cancer subtype would be possible. In conclusion for this study, MRI analysis of breast cancers utilizing a novel CNN can predict the molecular subtype of breast cancers.

Zhe Zhu et al. in 2019 [29] studied three methods (transfer learning, training from scratch and off-the shelf features) using three different architectures:

- **GoogleNet**:The main building block of GoogleNet is the Inception module that consists of three convolution layers with different kernel sizes. These kernels can be used to capture the features of patterns of different sizes;

- **VGGNet**:VGGNet is much larger than GoogleNet, which means it has more weights and needs more data to train. Thus it was only used reduced VGGNet due to the data insufficiency in the study;

- **CIFAR**:Are labeled subsets of the 80 million tiny images dataset and is a rather small network, compared to the other two. The input size set was $80 \times 80 \times 3$ and the output dimension of the last layer was set to 2. In training phase, the L2 regularization and dropout strategy were used to reduce overfitting.

Training from scratch is the most straightforward way to train convolutional neural network [29]. This technique is especially useful for new applications, as well as applications with a large

number of output categories. It is the most straightforward way to train convolutional neural networks. Current neural networks contain tens or even hundreds of layers, resulting in millions of free parameters to train. Building a large dataset requires large amount of time and labor, and in some situations such as in medical imaging the amount of available data is often insufficient. For GoogleNet were used both original GoogleNet and reduced GoogleNet while for VGGNet they just used the reduced VGGNet, resulting in four specific architectures in total. In this problem the lesions on MRI had different sizes which required the network to have multi-scale capability, that is why GoogleNet was chosen. Transfer learning is a two-step approach. The first step is to pre-train the network using a large dataset for a different task than the one at hand. The second step is fine-tuning the pre-trained network on the dataset representing the problem of interest. Usually the target dataset is much smaller than the dataset used for pre-training. The dataset used in the study [29] was natural image and the network pre-trained in this dataset are publicly available. GoogleNet and VGGNet pre-trained on ImageNet were the architectures chosen. Modifications were made to adapt the networks to the problem: the output dimension of the last fully connected layer was changed from 1000 to 2, and the weights in this layer were re-initialized randomly.

While training from scratch and transfer learning are both end-to-end approaches, off-the-shelf deep features approach explicitly separate feature extraction and classification [29]. Since this approach only uses pretrained networks as feature extractors and trains traditional classifiers using the extracted features, it can avoid the overfitting while taking advantage of the richness of deep features. Another benefit of using deep features approach is that the extracted features can be used for many different tasks.

For the study, GoogleNet and VGGNet pre-trained on ImageNet were chosen as feature extractors and trained Support vector machines (SVMs) using the extracted features. The task was to automatically determine whether the tumor is of the Luminal A subtype or of another subtype based on the MRI image patches representing the tumor. The best AUC performance for distinguishing molecular subtypes was 0.65 (95% CI:[0.57,0.71]) and was achieved by the off-the-shelf deep features approach. The highest AUC performance for training from scratch was 0.58 (95% CI:[0.51,0.64]) and the best AUC performance for transfer learning was 0.60 (95% CI: [0.52,0.65]) respectively. For the off-the-shelf approach, the features extracted from the fully connected layer performed the best

3.3 Summary

We can conclude with all these works, that it is possible to conjugate the MRI of breast cancer patients with the hormone receptors of breast cancer and try to predict the consequent molecular subtypes TNBC, Luminal A, Luminal B, and HER2 positive. Preoperative breast MRI has become increasingly prevalent as its role in diagnosis and treatment planning of breast cancer expands. With the increasing ubiquity of breast MRI, several studies were done, as we could see from this chapter.

Agner et al. [31] demarcated triple-negative cancers from other molecular subtypes on DCE-MRI using a computer-aided diagnosis (CAD) system. They retrospectively studied 76 breast lesions using quantitative feature extraction with a feed forward feature selection and linear discriminate analysis. Triple-negative tumors were found to be more heterogeneous in both texture and enhancement, with a higher degree of tumor tissue compactness. While this model improves on discriminating aggressive TNBC subtype, it relies on human MRI feature extraction. Tumor enhancement dynamics association with luminal type B molecular subtype was explored by Mazurowski et al. [23]. Twenty-three imaging features were extracted using computer vision algorithms after initial lesion delineation by a trained breast radiologist. Luminal B subtypes were found to have higher lesion enhancement. Similarly, computational MRI imaging features of luminal type A and B molecular subtypes were retrospectively evaluated using semi-automatically extracted imaging features by Grimm et al. [32]. Tumor tissue and the peak enhancement were shown to be main factors in discriminating subtype, with multivariate models able to predict luminal A and luminal B, but not HER2 positive or basal subtype.

MRI contrast enhancement patterns have shown promise in determining breast cancer subtype. Although the previously mentioned studies have shown promising results in breast MRI data analysis, these methods are dependent on feature engineering using semi-automated feature extraction. Feature engineering works by implementing domain knowledge to build feature extractors, which simplify the complex data and create more comprehensible patterns to be applied to algorithms. These methods are limited in their function, as they are dependent on accurate human extraction of crucial features.

In contrast, CNN algorithms are trained in a manner that allows automatic extraction of features from the input feed that are crucial to the defined problem domain. This process improves its ability to study the input features in an end-to-end manner, using complex, stacked layers to predict a desired output. To date, three studies utilized CNN to classify molecular subtypes using an MRI data set. Study by Zhu et al. [29] used VGGNet pre-trained on ImageNet. They used the feature maps of several convolution layers and fully connected layers, by training supervised machine learning algorithm on those feature maps. Their evaluation was limited to classifying one molecular subtype (Luminal A) versus the rest. Richard Ha et al. [30] evaluated four molecular subtype classification. Hong-Jun Yoon et al. [1] designed a DNN, which combined both radiomic and genomic features to predict pathological stage and molecular receptor status of invasive breast cancer patients.

With all the studies presented above, it was shown that is possible to determine features from MRI images of breast. This images can help to determine the type of hormonal receptor that is present or absentee in the patient, and therefore, determine the different types of molecular subtypes, that result for their expression. In our work we will evaluate two approaches: training from scratch and transfer learning, using the VGG16 model as proposed in the work [29].

Chapter 4

Methodology

In this chapter, the methodology of the proposed solution, is presented. In the first Section 4.1, a brief analysis of the used dataset is made, and Section 4.2 explains all the necessary processing of the dataset. Finally, Section 4.3, details the solution implementation.

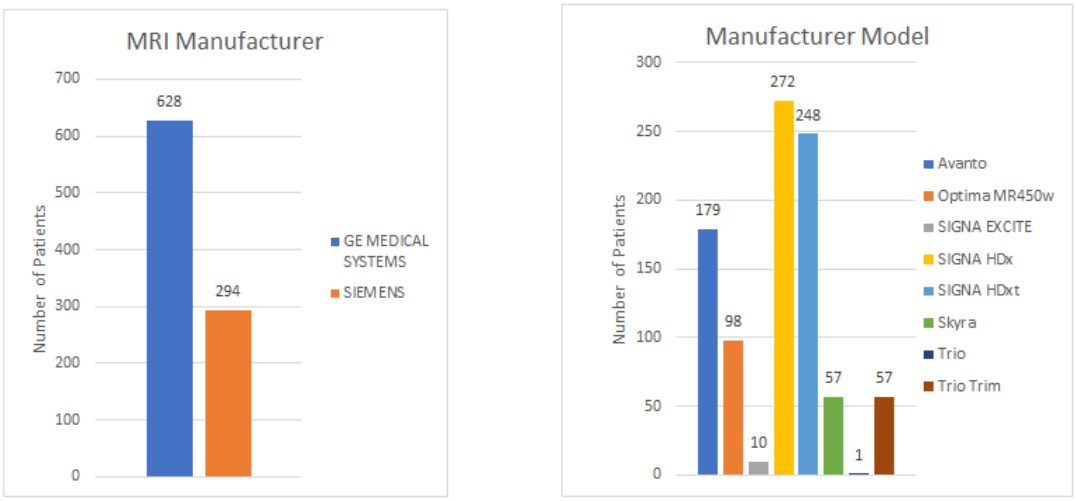
4.1 Dataset

The Dataset that is used for this work is a public dataset, available at the TCIA [26]. The breast MRI dataset contains 922 patients with invasive breast cancer and available pre-operative MRI, gathered from the Duke Hospital between the 1st of January 2000 and the 23rd of March 2014. These include axial breast MRI images acquired by 1.5T or 3T scanners in the prone positions. The MRI sequences available for each patient are: 1) Non-fat saturated T1-weighted sequence; 2) Fat-saturated gradient echo T1-weighted pre-contrast sequence; and 3) Mostly three to four post-contrast sequences. The other data available is:

- **Demographic, clinical, pathology, treatment, outcomes, and genomic data**: Collected from a variety of sources including clinical notes, radiology report, and pathology reports;
- **Locations of lesions in DCE-MRI**: Annotations on the DCE-MRI images by radiologists;
- **Imaging features from DCE-MRI**:A set of 529 computer-extracted imaging features by radiologists. These features represent a variety of imaging characteristics including size, shape, texture, and enhancement of both the tumor and the surrounding tissue.

The patients distribution was analysed according to the MRI manufacturer. To produce the MRI images, two different manufacturers were used, as shown in bar graph 5.8a, and each manufacturer had different models:

- **GE MEDICAL SYSTEMS**: Optima MR450w, SIGNA EXCITE, SIGNA HDx, SIGNA HDxt;
- **SIEMENS**: Avanto, Skyra, Trio, TrioTim.



(a) Manufacturer

(b) Manufacturer Model

Figure 4.1: Distribution of patients according to MRI manufacturer, on the left (a), and patients distribution according to the MRI manufacturer model name, on the right (b).

The distribution of patients according the MRI manufacturer and the corresponding model are in Figure 4.1

4.2 Data Preparation

The dataset images used the Digital Imaging and Communications in Medicine (DICOM) [33] format. DICOM is the international standard for medical images and related information. It defines the formats for medical images that can be exchanged with the data and quality necessary for clinical use. To be able to download the DICOM images from the dataset, the NBIA Data Retriever [34] open source software was used. The dynamic pre-contrast sequences were chosen, since these are used by medical doctors to evaluate the breast tumors [35]. To reduce unnecessary processing overhead, the unused images were discarded, resulting in a final dataset size of 10.3 GB, as opposed to the initial size of 368.4 GB. To be able to visualize the DICOM images, the RadiAnt DICOM Viewer [36] free software was used.

Patient ID	**Start Row**	**End Row**	**Start Column**	**End Column**	**Start Slice**	**End Slice**
Breast_MRI_001	234	271	308	341	89	112
Breast_MRI_002	251	294	108	136	59	72
Breast_MRI_003	351	412	82	139	96	108

Table 4.1: Annotation of lesions file from patient 1 to patient 3.

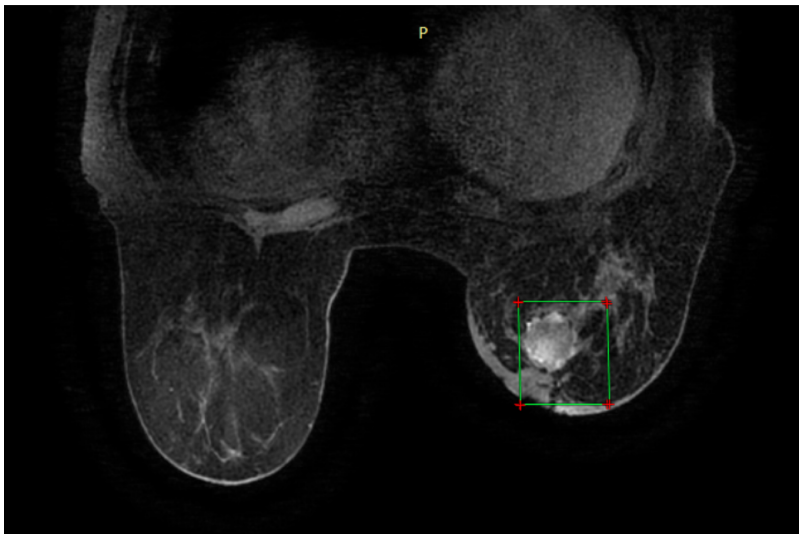


Figure 4.2: DICOM from MRI of breast cancer from patient with ID number 517 from dataset, with annotation of tumor.

4.2.1 DICOM images pre-processing

First, the tumor annotations were explored. The *Annotation_Boxes.xlsx* file, downloaded with the dataset, contained the annotations for every patient, it is notable, that the annotations were done as a box surrounding the tumor area, for all the slices where the tumor is visible. Table 4.1 exemplifies, with the first four patients how the annotations are provided, according to a box with **Row** x **Column** x **Slice** dimensions, where the tumor is visible. Figure 4.2 showcases an example of a slice from the dataset, with the respective Row x Column box to annotate the tumor location.

The use of neural networks required the definition of a unique image dimension for its input layer. However, the tumor sizes are different for every patient. Therefore, the annotation dimensions will also vary accordingly. To this end, it was necessary to define a crop size that would fit most of the annotations, always centering the tumor in the crop. Figure 4.3 shows the distribution of **Row** x **Column** annotations.

At first glance, a crop of 200x170 could be a possible solution to this problem, as it is the largest measurement in terms of row and column respectively. However, taking into account that there is a big percentage of patients with significantly lower annotation size, it could happen that in the presence of a very small tumor, most of the image area would be background instead of the tumor itself. Figure 4.4 exemplifies the aforementioned problem. For this reason, it was necessary to think of a different approach. Another approach was to define the crop dimension according to the mean value of the row and column range. For this, it could be possible to have a 50x50 crop in order to cover all the images in the mean range. If this size was chosen, 351 from the initial 922 patients, would be excluded. This would correspond to discarding approximately 38% of the dataset. In order to minimize the discarded images, while considering the best percentage of tumor in the image, a crop of 100x100 was defined for all slices in the dataset. Using this crop size, 88

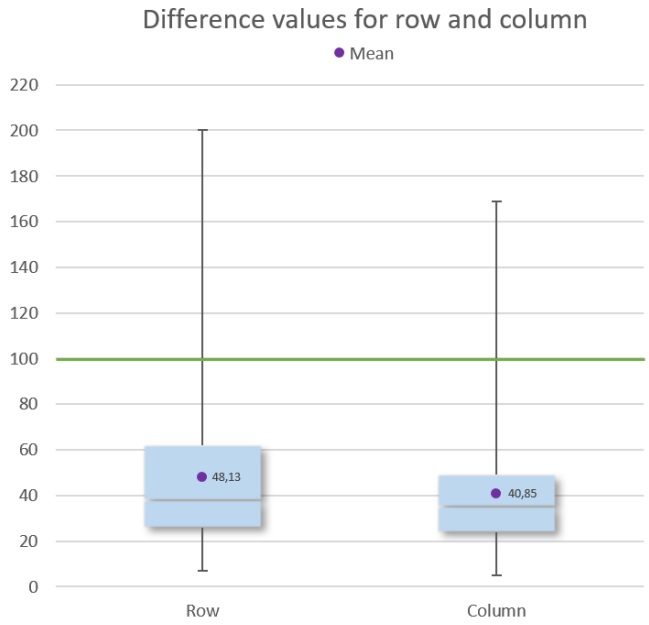


Figure 4.3: Box graph for values of row and column.

patients were excluded, – corresponding to 9.5% of the dataset – resulting in a total of 834 patients to evaluate. Figure 4.5 is an example of the 100x100 crop made in a MRI image, from the patient previously presented in Figure 4.2.

The dataset contains images from the whole breast MRI exam, but only the slices where the tumor is located are of interest, and this information is also available in the annotation file. Considering the Table 4.1 it is possible to calculate the number of relevant slices, for the first three patients, using the Start and End slices, resulting in 23, 13 and 12 respectively.

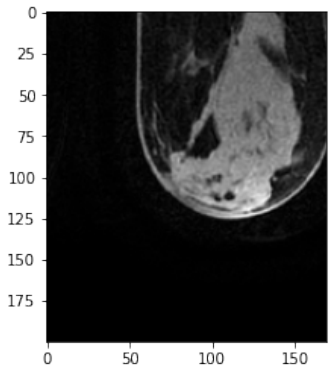


Figure 4.4: Crop of 200x170 from DICOM image.

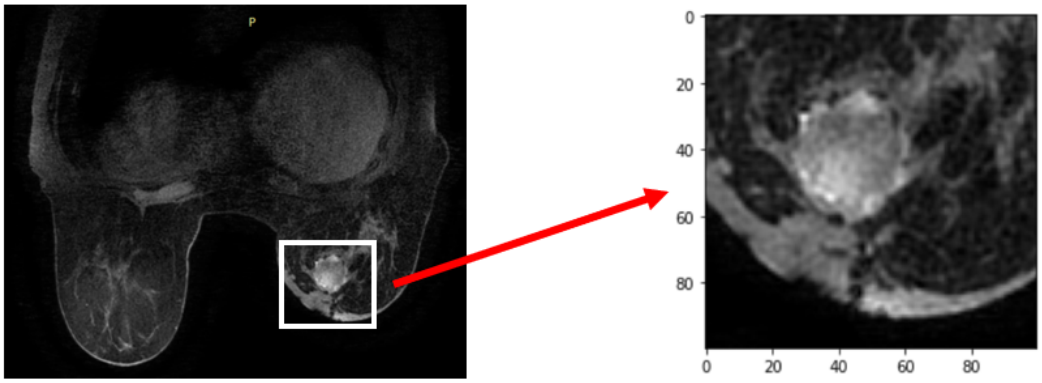


Figure 4.5: Crop 100x100 in MRI from a patient.

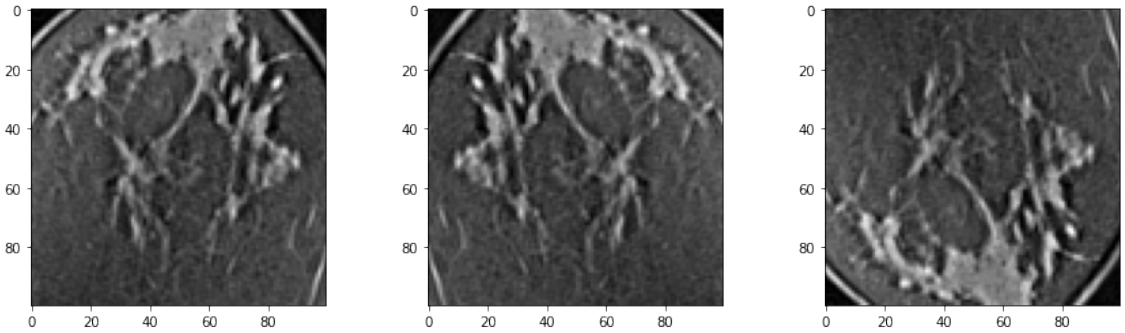
4.2.2 Data Augmentation

The information about the presence of each breast cancer hormonal receptors is available for each patient. This includes the ER, PR and HER2, and the respective molecular subtype according to the receptors, along with other genomic data. With this information it is possible to analyse the distribution of patients per hormonal receptor status in Table 4.2. It shows that the distribution of patients, according to the hormonal receptor presence/absence is not uniform.

Data augmentation are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. To augment the data a TensorFlow module, the *tf.image* [37] was used, which contains multiple functions for image processing. It was used the *flip left right*, to create a horizontal flip, in Figure 4.6b and the *flip up down*, to create a vertical flip, in Figure 4.6c, to create new images from the original ones (see Figure 4.6a). The data augmentation was performed only for the labels with least percentage of representation in the dataset. This means that for receptors ER and PR, data augmentation was made only for the images belonging to the label 0, and for receptor HER2, data augmentation was made for all the images belonging to the label 1. All the augmentation were made, after the crops, and the rational was to balance the number of images for each label, otherwise it could lead to bias in the classifier.

Hormonal Receptor	**Nº of patients for label1 (receptor present)**	**Nº of patients for label0 (receptor absent)**
ER	631	203
PR	548	286
HER2	143	691

Table 4.2: Distribution of patients according to breast cancer hormonal receptor status.



(a) Original Image (b) *Flip left right* augmentation (c) *Flip up down* augmentation

Figure 4.6: Example of data augmentation on image for patient with ID 132

4.2.3 Data Normalization

Data normalization is a set of techniques and rules used to improve the consistency of data, standardize it, and maintain its integrity. Moreover, it helps the model to calculate faster what are the optimal parameters for each input node. When normalizing data, all images, have their pixels scaled to a given range. Range examples include from $[-1,1]$ or $[0,1]$. For this work a $[0,1]$ range was decided, since it is the standard approach for data normalization. Equation 4.1 details the transformation applied to every image pixel.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4.1}$$

In a first approach it was decided to consider the x_{min} and x_{max} as the min and max values of the whole dataset. Considering the statistical study done in Section 4.1, regarding the different types of models and manufacturers that were used to acquire the MRI images, a second approach, was used for data normalization using the same mathematical expression, but taking into account the maximum pixel value of each image. Considering that all the models were used in a large number of patients, – with the exception of SIGNA EXCITE and Trio models – this could lead to differences in the way the images were acquired between patients.

4.3 Solution implementation

To evaluate our models, two metrics were used. Given the presence of unbalanced data and binary classification, the **balanced binary accuracy** was utilized. Additionally, the **AUC** was also utilized due to its prevalence in the literature review as a metric of performance. Another way to evaluate the performance of the model is with the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), when predicting the images in the test set:

- **True Positive** is an outcome when the model predicts correctly the positive label;
- **True Negative** is an outcome when the model predicts correctly the negative label;

- **False Positive** is an outcome when the model predicts that the label is positive, when actually it is negative;
- **False Negative** is an outcome when the model predicts that the label is negative, when actually it is positive.

With these values it is possible to calculate the **sensitivity** and the **specificity**, defined in Equations 4.2 and 4.3, respectively. The sensitivity measures the proportion of real positives that are correctly predicted out of all positive predictions that could be made by the model. The specificity measures the proportion of correctly identified negatives over the total negative predictions that could be made by the model. These two values are used to calculate the balanced binary accuracy, as shown in the equation 4.4

$$sensitivity = \frac{TP}{TP+FN} \tag{4.2}$$

$$specificity = \frac{TN}{TN+FP} \tag{4.3}$$

$$BalancedBinaryAccuracy = \frac{sensitivity + specificity}{2} \tag{4.4}$$

4.3.1 Environment

The development of this work, including all the presented algorithms, pre-processing of data and model training, was done, using Google Colab. [38]. Due to the large size of the dataset, the variables memory, time required to train the model, among other factors, the available RAM and GPU for regular Google Colab usage was insufficient. To this end subscription to Google Colab Pro was required. The GPU used was a NVIDIA A100-SXM4-40GB and the RAM capacity was 89.6GB. To train and create the models, the Keras API [39], was used. Keras is a deep learning API written in Python, running on top of the machine learning platform TensorFlow, which is an end-to-end, open-source machine learning platform.

The programming language used for this work was Python 3.8.10 [40], with the help of the pandas [41] module for the data processing methods. Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

4.3.2 Training from Scratch using VGG16

The Visual Geometry Group (VGG), belongs to the Department of Engineering Science of University of Oxford [42].The main goal when creating this Convolutional Neural Network (CNN), was to investigate the effect of the convolutional network depth on its accuracy in the large-scale image classification and recognition, using the ImageNet [43] dataset. This dataset was first released in 2009, and since then it has been a benchmark training and evaluate deep learning models. The

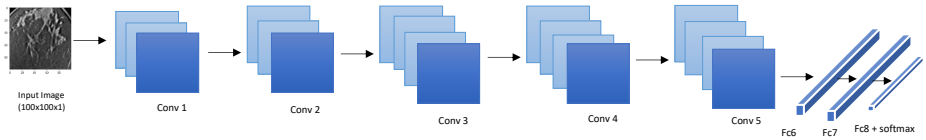


Figure 4.7: Vgg16 model with cropped image as input.

ImageNet contains more than 14 million images that are labeled according to their category, such as "cat", "dog", "bike", etc. For this reason it is mostly used for image classification tasks. The main contribution of this CNN is the use of very small convolutional filters (3 x 3), in order to increase the depth to 16–19 weight layers.

The VGG16 network has a RGB image of size 224 x 224 as input. VGG16 is composed of 16 layers in total, 13 of which are convolutional layers and 5 are max-pooling layers. The convolutional layers are responsible for extracting features from the input images, while the max-pooling layers are responsible for down-sampling the feature maps obtained from the convolutional layers. The final three layers are fully connected layers, the first two have 4096 channels and the last one has 1000 channels. This final layer is the soft-max layer. All the other layers have the activation function Relu (Rectified Linear Unit). Figure 4.7 represents the architecture of this network with the input image.

The cropped images, as depicted in Figure 4.5 are the input to a CNN which uses a VGGNet model. To be able to train the model, a split had to be done, where all the slices that belong from the same patient, also go to the same split (train, validation or test). This is important, because the potential presence of similar images in different subsets, may introduce bias in the model, due to model's exposure to similar images. The split was made, approximately in the following proportions: 80% of data to the train and the remaining 20% of the data were distribute equally for the validation and test. To do this, an algorithm that randomly selects patients ID from a dataframe, was used. All the images from the selected patients were then copied to the defined set (train, validation).

The data augmentation explained in the Subsection 4.2.2 was done for the images used to train the model. Both validation and test images are not augmented. The purpose of a validation set is to evaluate the performance of a model during the training process, while the test set is used to evaluate the final performance of the model. By augmenting the data on these sets, the model may become more accurate on these sets but will not generalize well to new unseen data and may lead to overfitting.

In Table 4.4 the distribution of images according to the label for each split is demonstrated, according to hormonal receptor. Two experiments were done with this distribution. The first one without augmentation, and the second one with augmentation. To evaluate the performance of the

Learning Rate	Batch Size
0.0001	128
0.0002	128
0.0002	256
0.005	256

Table 4.3: Hyper parameters tested during model training.

model, two hyperparameters were tested using multiple values, that are listed in Table 4.3.

4.3.3 Transfer Learning using VGG16

The second approach, uses a transfer learning method, also with VGG16, but previously trained with the ImageNet [43] dataset. A notable deviation in the current approach in comparison to the previous one pertains to the utilization of an input with a 100x100x3 shape, which is a result of incorporating a pre-trained model. The third dimension in the shape of the images, corresponds to the number of channels. A channel is simply an array of values, one per pixel, that together specify one aspect or dimension of the image. Previously, a grayscale image was being used, meaning only one channel is used. Herein, the requirement of three channels per image translates to the known RGB image format: red, green and blue. With the chosen dataset it is not possible to use RGB images, since MRI images, are grayscale. Therefore, the image transformation into 3 channels, was done by replicating the same channel 3 times, i.e., the 100x100 block of the image is replicated 3 times, thus obtaining a 100x100x3 image shape (see Figure 4.8).

For the transfer learning experiments, cross validation was used, with 5-fold train validation test splits. Cross validation is a technique that evaluates the performance of a machine learning model. It consists of dividing a dataset into a training set and a test set, training the model on the training set, and evaluating its performance on the test set. This process is typically repeated multiple times, with different subsets of the data used as the test set in each iteration. The goal of cross-validation is to estimate the performance of a model on unseen data by averaging its performance across multiple train-test splits. This can help to select the best model based on its generalization performance.

In this work three different sets of train, validation and test are used. Cross validation is performed by excluding $\simeq 10\%$ of the dataset, as the test set. The remaining 90% of the dataset is used for the validation and training set. For 5 different experiments, images were randomly chosen

Hormonal Receptor	Train		Validation		Test	
	Label0	Label1	Label0	Label1	Label0	Label1
HER2	11501	2547	1674	635	1282	499
PR	5024	10099	1045	415	998	502
ER	4002	10319	508	1429	492	1370

Table 4.4: Distribution of images according to label, for HER2, PR and ER.

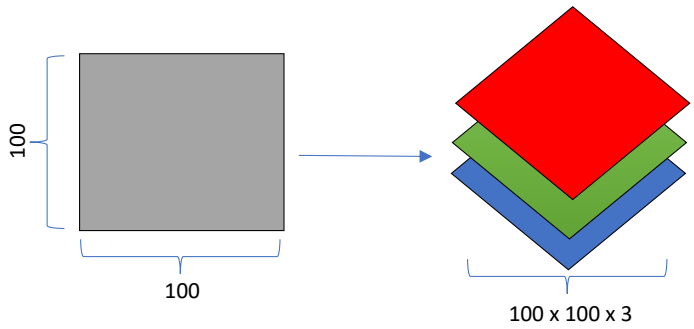


Figure 4.8: Exemplification of RGB image creation, from original image with shape 100x100x1.

for the training and validation set, taking into account that the train set should have $\simeq 80\%$ of the dataset. Finally, for all the 5 experiments, the model was tested with the test set and the results were evaluated.

In Tables 4.5, 4.6 and 4.7 the distribution of images according to the label for the 5 different experiments done for each hormonal receptor, HER2, PR, and ER, respectively. It is important to emphasize again, that the number of images containing tumor, differs from patient to patient, and it was not possible to have control on the number of images being allocated to each label. For this reason, the ratio between labels may differ slightly in some labels.

4.4 Summary

This chapter presented the methodology for this work. It is first presented the dataset chosen for this work. Before the solution implementation, the data had to be pre-processed. This pre-processing included discarding non relevant information and reducing the images data to the minimum relevant, in order to reduce the processing overhead during the implementation and training phase. This includes: cropping of the original images into an image of size 100x100 as well as data augmentation which was necessary, to balance each label percentage in the dataset. Finally, data normalization was performed in order to help with the model learning speed. Finally the environment used is detailed, and the methodology used for this work is also explained. When training from scratch, two hyperparameters were tested, and experiments with dropout and only one slice per patient were made. A transfer learning approach was made with a cross validation with 5 splits was made, and weight decay was applied to the model in order to find the mean performance for the model.

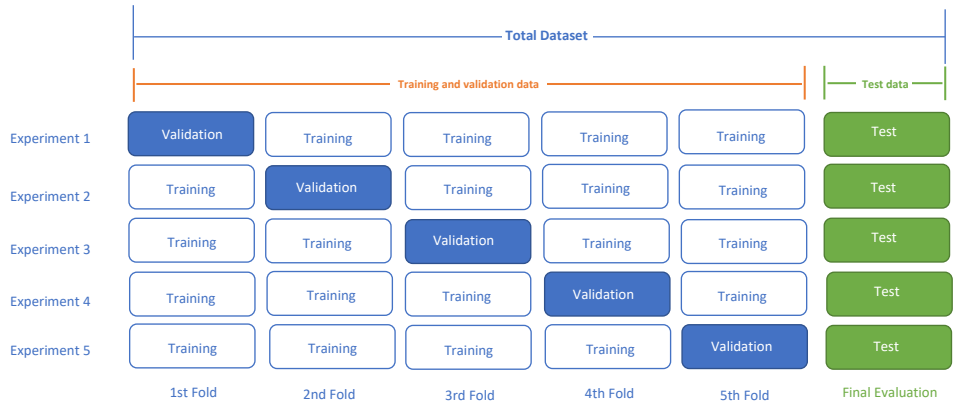


Figure 4.9: Cross validation used for the experiments with 5-fold, and training, validation and test split.

Split	Train		Validation		Test	
HER2	Label0	Label1	Label0	Label1	Label0	Label1
1	11615	2617	1542	565	1282	499
2	11751	2486	1406	696		
3	11662	2684	1495	498		
4	11572	2525	1585	657		
5	11648	2663	1509	519		

Table 4.5: Cross validation using 5-fold, with distribution of images for train, validation and test in label HER2.

Split	Train		Validation		Test	
PR	Label0	Label1	Label0	Label1	Label0	Label1
1	5178	8906	956	985	998	1097
2	5212	9009	922	822		
3	5087	9006	1047	885		
4	5182	8772	930	1119		
5	5242	8927	892	964		

Table 4.6: Cross validation using 5-fold, with distribution of images for train, validation and test in label PR.

Split	Train		Validation		Test	
ER	Label0	Label1	Label0	Label1	Label0	Label1
1	3922	10436	588	1312	492	1370
2	4119	10219	391	1529		
3	3962	10346	548	1402		
4	4084	10422	426	1326		
5	3999	10481	511	1267		

Table 4.7: Cross validation using 5-fold, with distribution of images for train, validation and test in label ER.

Chapter 5

Results and Discussion

In this chapter the results obtained from both, training from scratch and transfer learning are presented and discussed.

5.1 Training from Scratch

In this approach, several tests were made taking into account methods for model regularization. Taking into account the metrics explained in Chapter 4, the tests were made, and the results are discussed.

5.1.1 Data augmentation

Based on the distribution of images in Table 4.4 the model was trained, and the results analysed. Tables 5.1, 5.2 and 5.3 summarise the balanced binary accuracy values obtained in test for PR, HER2 and ER, respectively, when training the model with all images and with augmentation.

When comparing the experiment carried out with and without augmentation, the results show that the model performance is better when data augmentation is applied. Additionally, training with a batch size of 256 shows the best performance of all the tested combinations. There is a high correlation between the learning rate and the batch size. When using large batch sizes, the performance is better if the learning rate is also high [44]. Complementing these results,

All images			**All images and with label 0 augmentation**		
Batch Size	**Learning Rate**	**Balanced Binary Accuracy**	**Batch Size**	**Learning Rate**	**Balanced Binary Accuracy**
128	0.0001	$\simeq 0.50$	128	0.0001	$\simeq 0.55$
128	0.0002	$\simeq 0.50$	128	0.0002	$\simeq 0.50$
256	0.0002	$\simeq 0.50$	256	0.0002	$\simeq 0.56$
256	0.005	$\simeq 0.47$	256	0.005	$\simeq 0.48$

Table 5.1: Test set Balanced binary accuracy results for the PR hormonal receptor.

All images			All images and with label 0 augmentation		
Batch Size	**Learning Rate**	**Balanced Binary Accuracy**	**Batch Size**	**Learning Rate**	**Balanced Binary Accuracy**
128	0,0001	0.50	128	0.0001	$\simeq 0.55$
128	0.0002	0.50	128	0.0002	$\simeq 0.60$
256	0.0002	0.50	256	0.0002	$\simeq 0.62$
256	0.0005	0.50	256	0.005	$\simeq 0.51$

Table 5.2: Test set Balanced binary accuracy results for the HER2 hormonal receptor.

Figure 5.1 plots the train loss, balanced binary accuracy, and AUC. Taking the results of the PR hormonal receptor, it is noticeable, that although the balanced binary accuracy value in the test set is significantly lower when using a higher LR value, the training performance shows the opposite. The noise present in the loss value in the validation set when having a LR value of 0.0002 is significantly higher when compared to a LR value of 0.005, and also reaching much higher values.

The balanced binary accuracy for training reaches close to 100% accuracy. However, this is not reflected when the model is applied to new images during validation and testing. It is clear that there is a problem of overfitting being faced. Nevertheless, after this analysis, it was decided to set the values of 0.005 and 256 for the learning rate and batch size respectively for the following experiments.

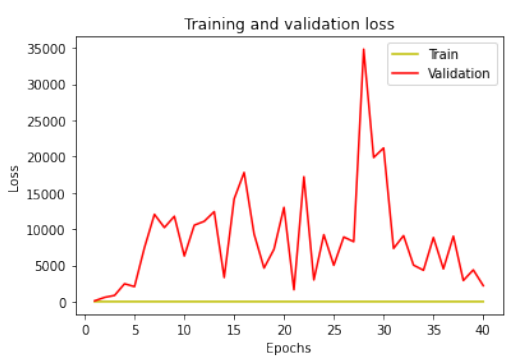
Overfitting is an undesirable behaviour that occurs when a model has learned the training data but can not generalize its learning for, unseen data. To prevent overfitting, regularization techniques are used, such as dropout and increasing the existing data with data augmentation.

5.1.2 Dropout with PR hormonal receptor

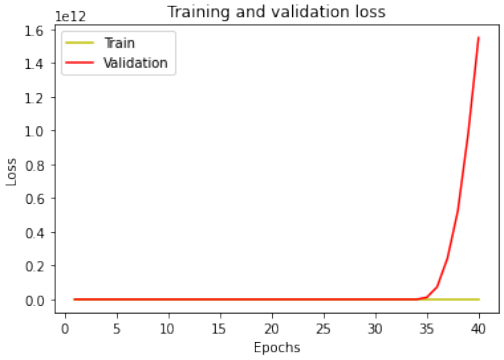
To avoid overfitting, a second experiment was done, using the same splits as before. Due to the time required to train multiple alternatives, the experiments for dropout were done only on the PR hormonal receptor. Dropout is a method that randomly drops out a certain percentage of neurons during training. This forces the model to learn multiple independent representations of the data, rather than relying too heavily on any other feature.

All images			All images and with label 1 augmentation		
Batch Size	**Learning Rate**	**Balanced Binary Accuracy**	**Batch Size**	**Learning Rate**	**Balanced Binary Accuracy**
128	0.0001	0.50	128	0.0001	$\simeq 0,52$
128	0.0002	0.50	128	0.0002	$\simeq 0,55$
256	0.0002	0.50	256	0.0002	$\simeq 0.56$
256	0.005	$\simeq 0.45$	256	0.005	$\simeq 0.46$

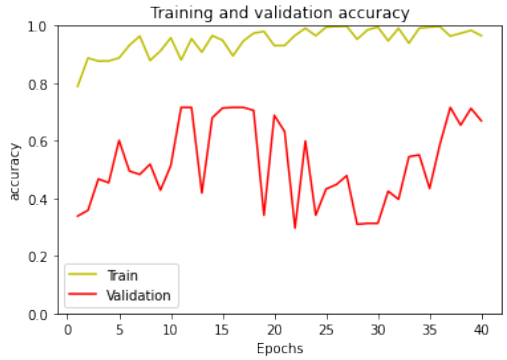
Table 5.3: Test set Balanced binary accuracy results for the ER hormonal receptor.



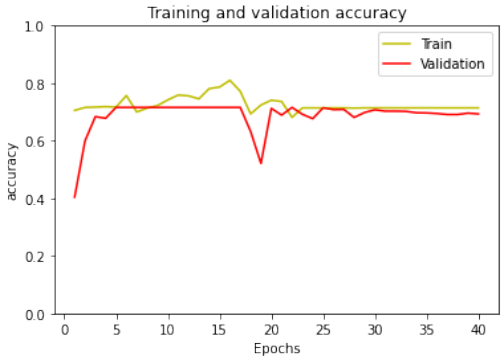
(a) Batch=256; LR=0.0002.



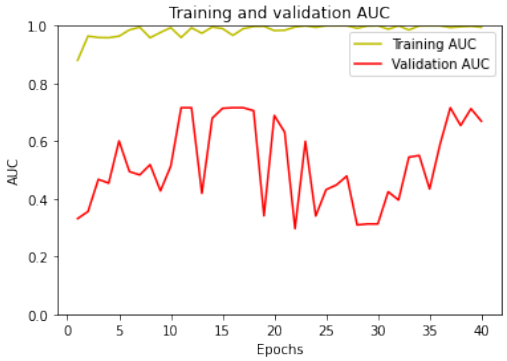
(b) Batch=256; LR=0.005.



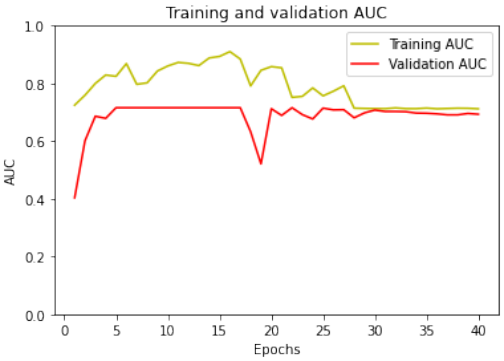
(c) Batch=256; LR=0.0002.



(d) Batch=256; LR=0.005.



(e) Batch=256; LR=0.0002.



(f) Batch=256; LR=0.005.

Figure 5.1: Results for loss a), balanced binary accuracy c) and AUC e) for the the experiment with augmentation and with batch size=256 and learning rate=0.0002 for the PR label. Results for loss b), accuracy d), and AUC f) for the same hormonal receptor, with augmentation and with batch size=256 and learning rate = 0.005.

Dropout	Balanced Binary Accuracy
1	$\simeq 0.56$
2	$\simeq 0.55$
3	$\simeq 0.62$

Table 5.4: Results for balanced binary after experiments with dropout in hormonal receptor PR.

The values of balanced binary accuracy on test set for experiments with one, two and three dropouts used in the model are available in Table 5.4, and the correspondent graphics for loss, balanced binary accuracy and AUC on train are on Figures 5.2, 5.3 and 5.4. The experiments with dropout also include data augmentation on label 0.

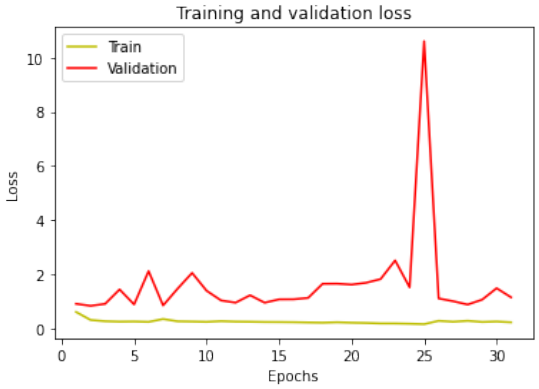
If the results obtained with the use of dropout are compared to the previous experiment without it, it can be seen that there is no improvement in the performance of the model. When using three dropouts it is notable that the results for validation loss are smaller, but comparing to the balanced binary accuracy, there was no improvement, on the contrary, it is noticeable lower accuracy values and more noise. Although the use of only one slice, decreases significantly the time for training, it does not show any notorious improvement, on the contrary, shows even worse performance.

5.1.3 Single patient slice with HER2 hormonal receptor

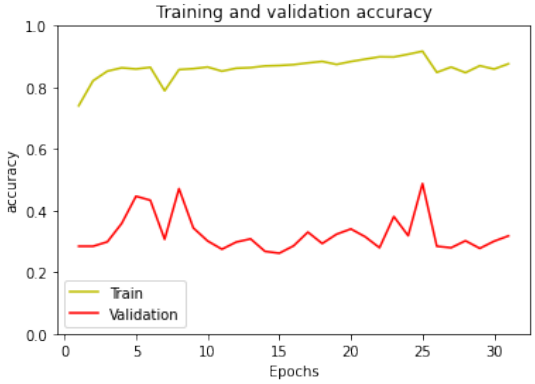
The last experiment was done using only one slice per patient. This means that instead of using all slices identified in the annotation, only one slice per patient is used to train, validate and test the model. The rationale for this experiment was to evaluate the model performance with a reduced number of images, when having fewer images. It may be possible that the tumor images belonging to the first and last slices do not contain much information, because they correspond to the beginning and end of the tumor. The middle slices are likely to have the most features. This is due to the tumor being more visible than when compared with the slices that are closer to the annotation limits. For this reason it was decided to provide the network only with the slice that corresponds to the median of the total number of slices per patient. Furthermore, this kind of analysis also made the training much faster. This experiment was made only for the HER2 hormonal receptor.

Two experiments were conducted, one without data augmentation, and the other with augmentation applied only to label 1. It was achieved a balanced binary accuracy of exactly 50% on test set, on the first experiment and 74% when doing augmentation. Despite the good accuracy values, when looking to the graphics 5.5 and 5.6 for loss and balanced binary accuracy on training and validation it is noticeable that the model did not behave as expected. The values for loss, keep reaching high values, and the accuracy on training is significantly lower than the expected and highly unstable.

From these experiments, the problem of overfitting still persists. There are multiple other regularization techniques that could be explored, but it was decided to proceed to a transfer learning approach.

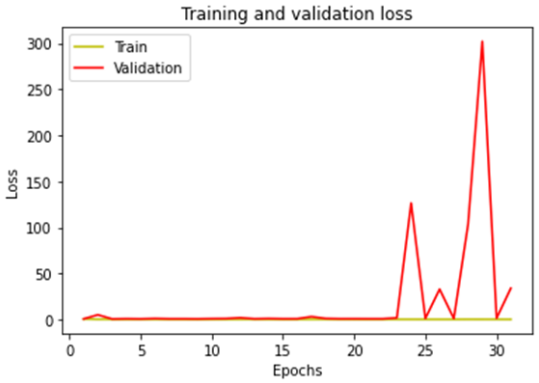


(a) Training and validation loss

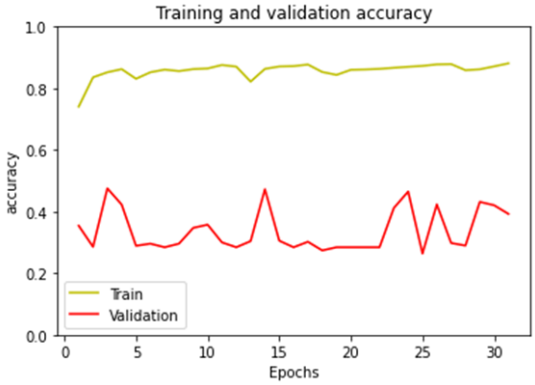


(b) Training and validation accuracy

Figure 5.2: Results for loss in training and validation a), and balanced binary accuracy in training and validation b), for experiment with augmentation and one dropout with PR hormonal receptor.

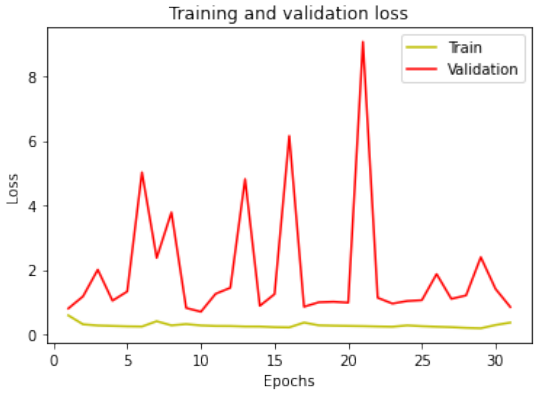


(a) Training and validation loss

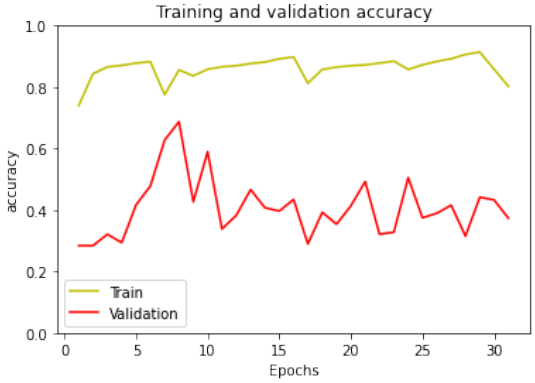


(b) Training and validation accuracy

Figure 5.3: Results for loss in training and validation a), and balanced binary accuracy in training and validation b), for experiment with augmentation and two dropouts with PR hormonal receptor.

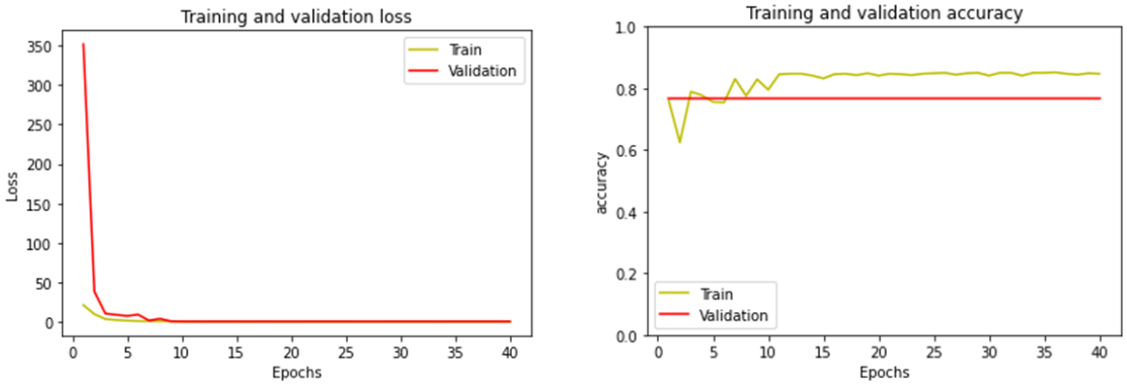


(a) Training and validation loss



(b) Training and validation accuracy

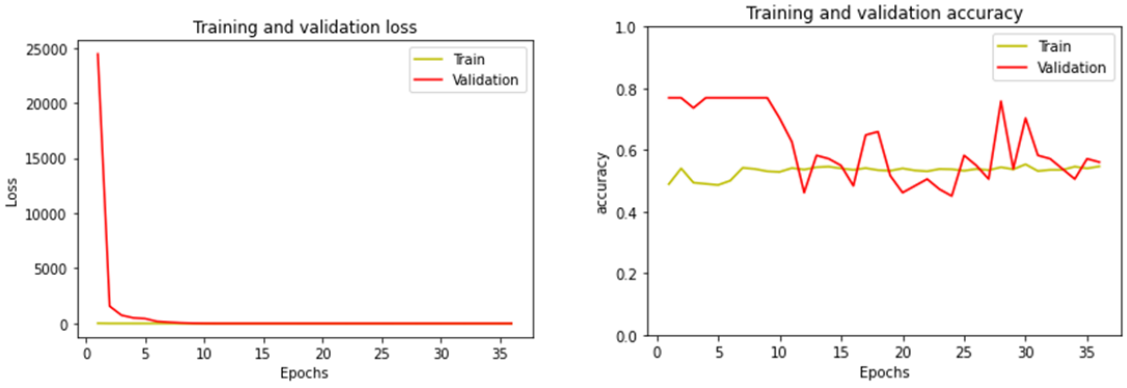
Figure 5.4: Results for loss in training and validation a), and balanced binary accuracy in training and validation b), for experiment with augmentation and three dropouts with PR hormonal receptor.



(a) Training and validation loss

(b) Training and validation accuracy

Figure 5.5: Results for loss in training and validation, on the left a), and balanced binary accuracy in training and validation on the right b), for experiment with the use of only one slice per patient on hormonal receptor HER2.



(a) Training and validation loss

(b) Training and validation accuracy

Figure 5.6: Results for loss in training and validation, on the left a), and balanced binary accuracy in training and validation on the right b), for experiment with augmentation and the use of only one slice per patient on hormonal receptor HER2.

Split PR	All Images		All images and with label 0 augmentation	
	Balanced Binary Accuracy with no decay	Balanced Binary Accuracy with decay	Balanced Binary Accuracy with no decay	Balanced Binary Accuracy with decay
1	$\simeq 0,48$	$\simeq 0.49$	$\simeq 0,52$	$\simeq 0.52$
2	$\simeq 0,51$	$\simeq 0.49$	$\simeq 0,50$	$\simeq 0.50$
3	$\simeq 0,51$	$\simeq 0.51$	$\simeq 0,52$	$\simeq 0.53$
4	$\simeq 0,50$	$\simeq 0.50$	$\simeq 0,49$	$\simeq 0.51$
5	$\simeq 0,50$	$\simeq 0.49$	$\simeq 0,51$	$\simeq 0.51$

Table 5.5: Balanced binary accuracy in five experiments for the PR hormonal receptor, using transfer learning.

5.2 Transfer Learning

Due to the poor performance exhibited by the first method considered, transfer learning is used as an alternative approach to address the problem, while also using the VGG16 architecture. In these experiments the second method for data normalization was used, which considers the higher pixel value value for each slice. The model which was pre-trained in ImageNet, needed to be fine-tuned with our training. This can be done by loading the pre-trained weights into the model and then training the model on our dataset while keeping the pre-trained weights fixed. In other words, the hidden layers are freezed and only the classifiers are trained.

As explained in Subsection 4.3.3, cross validation with five experiments was done for all hormonal receptors. Two experiments were made, with and without augmentation. In addition to this, in all splits and both experiments, the weight decay regularizer was considered. The weight decay penalizes large weights, thus encouraging the model to have smaller weights. The rationale of weight decay is that large weights can lead to overfitting, so by adding a penalty term to the loss function, the overfitting is prevented.

5.2.1 Using weight decay in hormonal receptor HER2 and PR

The module *regularizes* from keras was used, and a L2 regularization to the model was applied. It is a form of weight decay, which adds a penalty term to the loss function equal to the sum of the squares of the weights. This term is multiplied by a scalar hyperparameter called the weight decay rate, the l2, which controls the strength of the regularization. The effect of weight decay is that it causes the optimizer to update the weights in such a way that the sum of the squares of the weights is minimized, which leads to smaller weights, as explained in the equation 5.1 to compute the penalty.

$$loss = l2 \times reduce\_sum(square(x)) \tag{5.1}$$

Split HER2	All Images		All images and with label 0 augmentation	
	Balanced Binary Accuracy with no decay	**Balanced Binary Accuracy with decay**	**Balanced Binary Accuracy with no decay**	**Balanced Binary Accuracy with decay**
1	$\simeq 0.480$	$\simeq 0.49$	$\simeq 0.58$	$\simeq 0.56$
2	$\simeq 0,47$	$\simeq 0.50$	$\simeq 0.55$	$\simeq 0.53$
3	$\simeq 0,48$	$\simeq 0.49$	$\simeq 0.58$	$\simeq 0.58$
4	$\simeq 0,50$	$\simeq 0.50$	$\simeq 0.57$	$\simeq 0.54$
5	$\simeq 0,47$	$\simeq 0.48$	$\simeq 0.52$	$\simeq 0.56$

Table 5.6: Balanced binary accuracy in five experiments for the HER2 hormonal receptor, using transfer learning.

If the weight decay rate is set too high, it can prevent the model from learning at all. Therefore, it is important to tune the weight decay rate to find the optimal balance between preventing overfitting and allowing the model to learn effectively. It was set up a value, after some experiments to 0,001.

Tables 5.5 and 5.6 present the results of balanced binary accuracy of the test set for all the five experiments done considering the HER2 and PR hormonal receptors. Even though the image splitting for the ER hormonal receptor has been done, the experiments for this hormonal receptor with transfer learning were not performed due to time limitations. Analysing the obtained results, shows that when weight decay is used, an improvement in performance is observed in experiments without augmentation. When observing the results for the experiments with augmentation, it is not clear whether the use of weight decay can improve the performance of the model or not.

The results of mean values for the balanced binary accuracy results for the two hormonal receptor were calculated, and the results presented on Table 5.7. It is noticeable that without augmentation the hormonal receptor PR achieves a better performance, but when considering data augmentation, hormonal receptor HER2, has a better performance. For this reason, the behavior of the model in training and validation will be analysed in more detail. For the case of the PR hormonal receptor, this will be done without augmentation, while for the HER2 hormonal receptor, the analysis will be done with data augmentation.

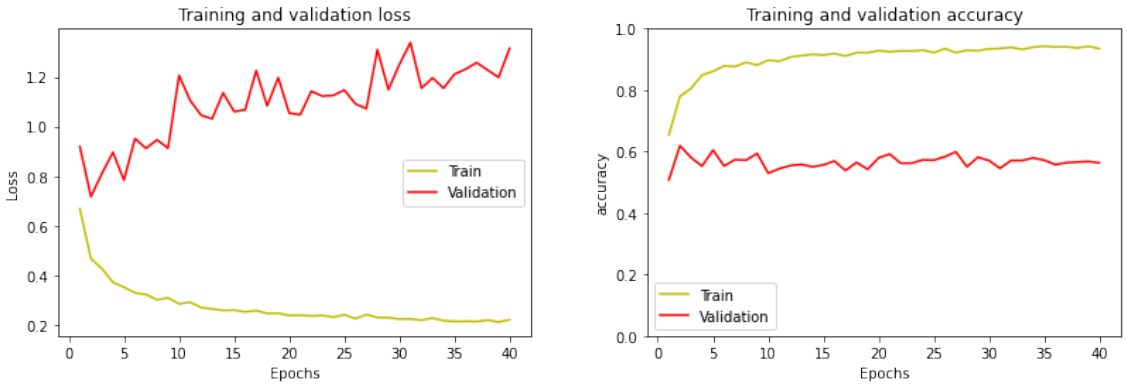
When comparing this results from the previous one, it is noticeable that the use of the new normalization poses a big difference to the model when training, since an accuracy of almost 100% is obtained, that was not possible before (see figures 5.7 and 5.8). If the values obtained during training and validation for loss are analysed, it is noticeable that the values are much lower now, when compared with training from scratch. There is still some noise present during training and validation, but even though, in general the model behaves better when training scratch, reaching a higher performance when using transfer learning.

Hormonal Receptor	All images		All images and augmentation	
	No decay	Decay	No decay	Decay
PR	0.5	$\simeq 0.50$	$\simeq 0.51$	$\simeq 0.51$
HER2	$\simeq 0.48$	0.49	$\simeq 0.56$	$\simeq 0.55$

Table 5.7: Mean performance on all the experiments done, with PR and HER2 hormonal receptors.

5.3 Summary

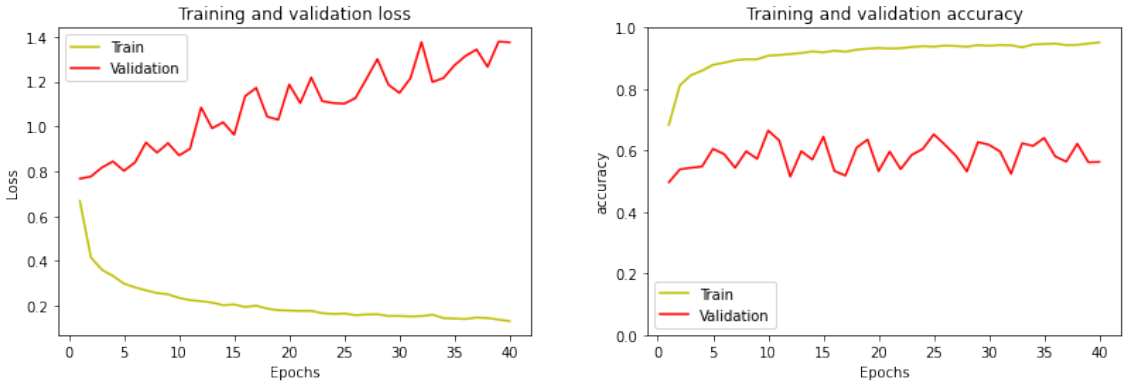
This chapter provided the results for different experiments done throughout the work, and the main conclusions are herein summarised. The use of transfer learning brings a higher performance to the model. The use of data normalization, taking into account the biggest pixel value from each image, as opposed to the biggest pixel value of the whole dataset, leads to a better performance of the model. Weight decay may improve the model, in the case where data augmentation was not applied. The experiments with different hormonal receptors introduce different performances and behaviour to the models.



(a) Training and validation loss

(b) Training and validation accuracy

Figure 5.7: Results for loss in training and validation, on the left a), and balanced binary accuracy in training and validation on the right b), for experiment with transfer learning, with decay on split2 without augmentation on hormonal receptor PR.



(a) Training and validation loss

(b) Training and validation accuracy

Figure 5.8: Results for loss in training and validation, on the left a), and balanced binary accuracy in training and validation on the right b), for experiment with transfer learning, without decay on split3 with augmentation on hormonal receptor HER2.

Chapter 6

Conclusions and Future Work

In 2020 breast cancer was one of the most common cases of cancer worldwide. Nowadays the diagnosis of breast cancer is made with various exams, such as mammogram and confirmed by taking a biopsy of the concerning tissue, that is later analysed, which is an invasive intervention and has some limitations. During and after biopsy, the patient may have pain, bruising, and bleeding at the biopsy site.

To overcome this, the diagnosis of breast cancer may be possible through MRI image analysis, with the use of DL techniques. MRI is a noninvasive procedure that does not use radiation, unlike the mammogram. The high quality of tissue can make it possible to extract features, that have the possibility to describe tumor characteristics and therefore help distinguishing between different hormonal receptors.

This study proposes the use of a CNN, using the VGG16 architecture to predict the hormonal receptor status with MRI images. First a training from scratch approach was proposed, however the model performance was poor. Then a transfer learning approach was proposed, also using the VGG16 architecture, but pre trained with ImageNet. Using this approach, the model had a better performance. The highest balanced binary accuracy on test obtained was 58% when predicting hormonal receptor HER2.

We had some limitations in our work. Discarding of 88 patients result of the initial steps of our work to define the best crop, may have led to worse model performance. Additionally, the unbalanced images in the dataset results in a big difference in terms of distribution of patients per hormonal receptor status. Lastly, the use of ImageNet to pre train model may not be suited for this classification task, since the images from this dataset, do not have any similarities with medical images, such as MRI or CT.

As future work, the use of medical images to pre train the models may be promising in the detection of the breast cancer hormonal receptors. The use of other state of the art models, with other regularization methods that were not explored in this work, could also improve the overall performance. Finally associating the hormonal receptor status with the breast cancer molecular subtypes, in order to differentiate the different types of breast cancer tumors, could lead to interesting results.

References

[1] Hong-Jun Yoon et al. Deep radiogenomics for predicting clinical phenotypes in invasive breast cancer. *Proc. SPIE 10718, 14th International Workshop on Breast Imaging (IWBI 2018)*, July 2018. doi:10.1117/12.2318508.

[2] Cancer. [Acessed 26-Jan-2023]. URL: https://www.who.int/news-room/fact-sheets/detail/cancer.

[3] Global cancer observatory. [Acessed 26-Jan-2023]. URL: https://gco.iarc.fr/today/online-analysis-pie?v=2020&mode=cancer&mode_population=continents&population=900&populations=900&key=total&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=7&group_cancer=1&include_nmsc=1&include_nmsc_other=1&half_pie=0&donut=0.

[4] Estimated number of new cases of breast cancer from 2020 to 2030, females, all ages. [Acessed 26-Jan-2023]. URL: https://gco.iarc.fr/tomorrow/en/dataviz/isotype?types=0&sexes=2&mode=population&group_populations=1&multiple_populations=1&multiple_cancers=0&cancers=20&populations=903_904_905_908_909_935&single_unit=100000&years=2030.

[5] Cancro da mama. [Acessed 26-Jan-2023]. URL: https://www.ligacontracancro.pt/cancro-da-mama/.

[6] Deteção. [Acessed 26-Jan-2023]. URL: https://www.ligacontracancro.pt/cancro-da-mama-deteccao/.

[7] Programa de rastreio de cancro da mama. [Acessed 26-Jan-2023]. URL: https://www.ligacontracancro.pt/servicos/detalhe/url/programa-de-rastreio-de-cancro-da-mama/.

[8] Avice O'Connell and Daniel Kawakyu-O'Connor. Dedicated cone-beam breast computed tomography and diagnostic mammography: Comparison of radiation dose, patient comfort, and qualitative review of imaging findings in bi-rads 4 and 5 lesions. *Journal of clinical imaging science*, 2:7, 02 2012. doi:10.4103/2156-7514.93274.

[9] Breast biopsy. [Acessed 26-Jan-2023]. URL: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy.html.

[10] Tycho Bismeijer PhD et al. Radiogenomic analysis of breast cancer by linking mri phenotypes with tumor gene expression. *Radiology*, 2020. doi:https://doi.org/10.1148/radiol.2020191453.

[11] Ella F. Jones, Deep K. Hathi, Rita Freimanis, Rita A. Mukhtar, A. Jo Chien, Laura J. Esserman, Laura J. van't Veer, Bonnie N. Joe, and Nola M. Hylton. Current landscape of breast cancer imaging and potential quantitative imaging markers of response in er-positive breast cancers treated with neoadjuvant therapy. *Cancers*, 12(6), 2020. URL: https://www.mdpi.com/2072-6694/12/6/1511, doi:10.3390/cancers12061511.

[12] P. C Nowell. The clonal evolution of tumor cell populations. science (80-), 1976 Oct 1. doi:10.1126/science.959840.

[13] Joana Mourato Ribeiro. Biologia do carcinoma de mama. In *MANUAL DE ONCOLOGIA SPO Abordagem e tratamento do cancro da mama*, pages 39,40. Sociedade Portuguesa de Oncologia, 2020.

[14] Pervolarakis N. Blake K. et al. Nguyen, Q.H. Profiling human breast epithelial cells using single cell rna sequencing identifies cell diversity. *Nat Commum*, 9, 2018. doi:https://doi.org/10.1038/s41467-018-04334-1.

[15] Cancro da mama. [Acessed 26-Jan-2023]. URL: https://www.cuf.pt/saude-a-z/cancro-da-mama.

[16] Harbeck N. et al. Breast cancer. *Nat Rev Dis Primers 5*, 2019. doi:https://doi.org/10.1038/s41572-019-0111-2.

[17] Magnetic resonance imaging. [Acessed 26-Jan-2023]. URL: https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri.

[18] Elizabeth A.Morris. Diagnostic breast mr imaging: Current status and future directions. *Radiologic Clinics of North America*, page 18:57–74, 2010. doi:https://doi.org/10.1016/j.rcl.2007.07.002.

[19] Brca gene mutations: Cancer risk and genetic testing. URL: https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet#:~:text=Testing%20for%20inherited%20BRCA1%20and,the%20BRCA1%20and%20BRCA2%20genes.

[20] Issam El Naqa and Martin J. Murphy. What is machine learning? *Machine Learning in Radiation Oncology: Theory and Applications*, 2015. doi:https://doi.org/10.1007/978-3-319-18305-3_1.

[21] Machine learning. [Acessed 26-Jan-2023]. URL: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML.

[22] Andre Esteva et al. A guide to deep learning in healthcare. *Nat Med 25*, page 24–29, 2019. doi:https://doi.org/10.1038/s41591-018-0316-z.

[23] Maciej A. Mazurowski et al. Radiogenomic analysis of breast cancer: Luminal b molecular subtype is associated with enhancement dynamics at mr imaging. *Radiology: Volume 273: Number 2*, pages 365–72, Nov 2014. doi:10.1148/radiol.14132641.

[24] Imaging cancer archive. [Acessed 26-Jan-2023]. URL: https://www.cancerimagingarchive.net/.

[25] Lars J. Grimm et al. Computational approach to radiogenomics of breast cancer: Luminal a and luminal b molecular subtypes are associated with imaging features on routine breast mri extracted using computer vision algorithms. *Journal of Magnetic Resonance Imaging 42*, pages 902–907, 2015. doi:10.1002/jmri.24879.

[26] Ashirbani Saha et al. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. *British Journal of Cancer*, July 2018. doi:https://doi.org/10.1038/s41416-018-0185-8.

[27] A comprehensive guide to convolutional neural networks. [Acessed 26-Jan-2023]. URL: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b11

[28] convolutional neural network. [Acessed 26-Jan-2023]. URL: https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network.

[29] Zhe Zhu et al. Deep learning for identifying radiogenomic associations in breast cancer. *Computers in Biology and Medicine*, 2019. doi:https://doi.org/10.1016/j.compbiomed.2019.04.018.

[30] Richard Ha et al. Predicting breast cancer molecular subtype with mri dataset utilizing convolutional neural network algorithm. *Journal of Digital Imaging (2019)*, page 32:276–282, July 2018. doi:https://doi.org/10.1007/s10278-019-00179-2.

[31] Shannon C. Agner et al. computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced mr images: A feasibility study. *Radiology, volume 272*, page 272:91–99, July 2014. doi:https://doi.org/10.1148/radiol.14121031.

[32] Lars J. Grimm and Maciej A. Mazurowsk. Breast cancer radiogenomics: Current status and future directions. *The Association of University Radiologists*, 2019. doi:https://doi.org/10.1016/j.acra.2019.09.012.

[33] Digital imaging and communications in medicine. [Acessed 26-Jan-2023]. URL: https://www.dicomstandard.org/.

[34] Installing the nbia data retriever. [Acessed 26-Jan-2023]. URL: https://github.com/CBIIT/NBIA-TCIA.

[35] R M et al. Mann. Breast mri: guidelines from the european society of breast imaging. *European radiology*, 18,7:1307–18, 2008. doi:10.1007/s00330-008-0863-7.

[36] Radiant dicom viewer. [Acessed 26-Jan-2023]. URL: https://www.radiantviewer.com/.

[37] Module: tf.image. [Acessed 26-Jan-2023]. URL: https://www.tensorflow.org/api_docs/python/tf/image.

[38] Welcome to colaboratoy. [Acessed 26-Jan-2023]. URL: https://colab.research.google.com/.

[39] Keras. [Acessed 26-Jan-2023]. URL: https://keras.io/about/.

[40] Python 3.8.10. [Acessed 26-Jan-2023]. URL: https://www.python.org/downloads/release/python-3810/.

[41] pandas. [Acessed 26-Jan-2023]. URL: https://pandas.pydata.org/.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. URL: https://arxiv.org/abs/1409.1556, doi:10.48550/ARXIV.1409.1556.

[43] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

[44] Ibrahem Kandel and Mauro Castelli. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, 6(4):312–315, 2020. doi:https://doi.org/10.1016/j.icte.2020.04.010.