POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

Deep Learning Techniques for Breast Cancer Characterization in Magnetic Resonance Images

Supervisors

Prof. Fabrizio LAMBERTI

Dr. Lia MORRA

Candidate

Angelo LAUDANI

April 2021

Abstract

Background: The aim of this thesis is to explore the solutions that Deep Learning techniques can offer in the field of Medical Imaging, in particular for breast cancer characterisation in magnetic resonance images. The thesis proposes the development of a Deep Learning architecture for a concrete problem such as the evaluation of pathological complete response (pCR) to neoadjuvant chemotherapy in breast cancer.

Methods: The mpMRI dataset analysed includes 37 patients, each of whom underwent two studies: before and after 2 cycles of NAC. An index slice was extracted from each available sequence by an experienced radiologist. Pathological results were used as ground truth. The proposed architecture seeks to make the most of the multi-parametric nature of the dataset, extracting features separately from each of the available image modalities (DCE, DWI and T2). The resulting sub-sequences are used as input for a multi-task ensemble learning model that takes into account the different type of information represented by each of them, as well as the time dimension due to the two studies per patient. The use of a specific branch for each sub-sequence combined with the use of Grad-CAM aims to provide an additional level of interpretability to a model that starts directly from full slices.

Results: Using 4-fold cross-validation, with each training set consisting of 28 patients and each validation set of 9, the mean area under the receiver operating characteristic (ROC) curve (AUC) of the model was 0.90, with a positive predictive value of 86.3% using all the available sub-sequences. Experiments with different configurations have shown that the combined use of all sub-sequences and both studies (pre-NAC and post-NAC) available per patient results in a model capable of better performance and generalisation.

Conclusion: The work conducted in this thesis demonstrates the great potential of Deep Learning applied in this specific medical field, proposing a solution that achieves significant results in the use of an mpMRI dataset for early prediction of pCR to NAC, an area that is still little explored in the available literature and that could provide valuable information in a crucial task such as treatment prediction.

Acknowledgements

Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino.

Table of Contents

List of Tables

List of Figures

Acronyms

AI

Artificial Intelligence

ADC

Apparent Diffusion Coefficient

AUC

Area Under the Curve

CAD

Computer-Aided Detection and Diagnosis

CNN

Convolutional Neural Network

DL

Deep Learning

DCE

Dynamic Contrast Enhanced

DSC

Dice Similarity Coefficient

DWI

Diffusion Weighted Imaging

GAN

Generative Adversarial Network

LDA

Linear Discriminant Analysis

LSTM

Long Short-term Memory

MCT

Motion Correction Technique

MIP

Maximum Intensity Projection

MRI

Magnetic Resonance Imaging

NAC

Neoadjuvant Chemotherapy

pCR

Pathological Complete Response

ROC

Receiver Operating Characteristic

SVM

Support Vector Machine

Chapter 1

Introduction

In a world that is becoming more data-driven with each passing day, it should come as no surprise that the availability of large-scale data is slowly but surely shaping every aspect of our lives. With an average of about 1.7 MB of data per second generated by each person, as estimated by the International Data Corporation (IDC), steadily shifting towards individuals over companies, the challenge of making the most of the great potential hidden in this growing mass of digitized information permeates every sector of research, be it industrial or academic.

Information is a valuable resource that can be integrated into any activity, in order to improve it from an organizational and productive point of view. It can help in understanding an event, predicting others, or finding seemingly invisible correlations among them. But analysing and finding meaning in large amounts of often unstructured data is not a straightforward task. The three "Vs" of "big data", as first theorized by Douglas Laney [1], can give a more precise idea of the scope of this task: volume, variety and velocity describe in fact the main characteristics of this field. With an estimated 44 ZB of stored data (as of 2020), this volume is growing at an exponential rate, contributing to the speed with which data are collected and available, in an ever-increasing variety of formats and types.

In this scenario, the development and increased affordability of hardware and software with fast and cost-efficient computational power has been a driving force for the development of solutions that not only allow efficient analysis and storage of this large amount of data, but also use them as a starting point for training models capable of autonomously extracting complex representations.

This is the case of machine learning, whose purpose is to learn the representation of data used as input, extracting generalized patterns from it, in order to then make predictions on new data never seen before; that is, learning from data, to perform tasks without being explicitly programmed to do so.

Deep Learning can be seen as a complex and, to some extent, natural evolution of machine learning algorithms, an evolution made possible by the growing amount

of data that can be used as input and processed with increasing power. Designing a hierarchical, layered architecture capable of extracting highly abstract features from data and emulating the human brain learning process, can also be applied to the most fundamental processing element of these architectures, the neuron. These stacked non-linear features extractor have achieved impressive results by taking full advantage of the information contained in large-scale data, without the need for a domain expert to reduce the complexity of the data.

Quoting Andrew Ng: *"The analogy to Deep Learning is that the rocket engine is the Deep Learning models and the fuel is the huge amounts of data we can feed to these algorithms."* [2].

It is therefore not surprising that Deep Learning is being used in a variety of increasingly essential and innovative applications, from natural language processing to computer vision. In this last field, among the many applications, Medical Image Analysis represents one of the most difficult and stimulating challenges, with results in continuous evolution that make its contribution no longer negligible in a crucial field such as medicine.

The aim of this thesis is to realize a Deep Learning architecture able to help in the resolution of a concrete problem, the prediction of pathological complete response in neoadjuvant treatment of breast cancer. Neoadjuvant chemotherapy (NAC) is widely used to treat locally advanced breast tumors before surgery, aiming to reduce tumor size and enabling breast-conservation surgery instead of mastectomy, among other advantages. Pathological complete response (pCR), the absence of residual invasive disease in the breast or lymph nodes, is associated with a significantly improved disease-free and overall survival. However, a pCR is achieved in just around the 30% of the patients after the completion of NAC, making the prediction of treatment response a crucial early step for identifying patients who do not benefit from NAC, allowing them to immediately undergo a different treatment. Speeding up and helping the decision-making process of a radiologist can therefore be extremely important in such a context, and one possible way to achieve this goal is to intensively exploit the information contained in the MRI studies performed on the patient, with particular focus on a study acquired before the start of NAC therapy and one after completing the first two cycles.

The architecture proposed in this thesis aims to make the most of a dataset that, in addition to the aforementioned temporal component consisting of the two studies per patient, is composed of mpMRI studies, i.e. multi-parametric, allowing access to different sequences using different acquisition protocols, each containing peculiar information about the lesion. A multi-branch ensemble learning architecture is therefore used to process all this heterogeneous information as best as possible, trying to produce a result that is not only reliable, but also enjoys a certain level of interpretability in the decision-making process. This is done by trying to maintain the highest possible level of automation, reducing the need for domain knowledge

and manual intervention by qualified personnel on the source data to a minimum.

The use of all the latest state-of-the-art techniques in the field of Deep Neural Networks is only part of the construction of such a specific architecture, which uses as a starting point the studies available in the literature on breast MRI analysis with Deep Learning, addressing all the challenges typical of the Medical Imaging field, from the scarcity to the structure of the source data, to specific and often necessary architectural choices, up to the integration with a traditional medical pipeline.

Chapter 2

Medical Image Analysis

Medical Imaging defines the techniques and processes used to generate images of the human body for various clinical purposes such as medical procedures and diagnosis or medical science including the study of normal anatomy and function [3]. Extending its definition to that of discipline, it can be considered part of biological imaging and incorporates radiology.

The technological evolution of this field, which has reached various stages over the years, from the first applications of X-rays to the most recent MRI modalities, now poses new challenges and possibilities in the era of big data and digitization. The role of images, which are becoming more numerous and richer in information, is increasingly key in the decision-making process of medical experts, from diagnosis to prognosis and treatment of a disease.

But the interpretation of these complex images can be tedious and prone to misinterpretation, requiring a great deal of effort from qualified personnel. The introduction of tools to support the quantification and classification of these data is an important step towards new standards of quality, service efficiency, and costs of healthcare along with the reduction of medical errors [4].

2.1 Computer-Aided Diagnosis

CAD systems, standing for *Computer-Aided Detection/Diagnosis*, have over time become an integral part of routine clinical work.

The idea and enthusiasm for computers capable of totally replacing humans in medical tasks dates back to the first examples of digitization of medical images in the 1960s. But, following a common trend for this type of technology in those years, excessive enthusiasm soon gave way to disinterest in the face of poor results due to the lack of adequately powerful technologies at the time. In the 1990s, however, the concept of CAD as we understand it today re-emerged, this time as a tool to

be used alongside a radiologist rather than replacing him completely [5].

From a technical point of view, these systems use pattern recognition and classification techniques to help either in the detection of an abnormality or in its classification. This makes them particularly suitable for application in the field of detection and diagnosis of cancerous lesions, a potentially repetitive task for a radiologist but one with well-defined rules to be learned by a machine.

Supporting evidence from existing medical literature, together with the rapid development of hardware and software capable of sustaining such applications and scaling better on the huge amount of data available, as well as the now achieved integration and standardization by regulatory bodies such as the FDA, make CAD systems central to many applications such as breast cancer detection.

2.2 Breast Cancer and NAC

According to the World Health Organization, impacting around 2.1 million women each year, breast cancer is the main cause of cancer-related death among women and the second most common tumor overall. Outcomes for breast cancer are very different depending on factors such as the cancer type, the person's age, and the extent of disease. The five-year survival rates in developed countries is between 80 and 90%, sinking to 40% for developing countries [6].

Therefore, early detection and an appropriate diagnosis leading to a correct treatment are critical for improving the survival rate to a certain extent.

NAC, *Neoadjuvant Chemotherapy*, refers to the systemic treatment of breast cancer prior to surgery, and is an important option in patients with early-stage breast lesions. This therapy can convert an unresectable, inoperable tumor to an operable one, as well as increasing the chances of breast-conserving surgery in patients with operable breast cancers. Another advantage is the possibility of observing direct and early stages of response to treatment, allowing fast modifications of the treatment plan in the event of poor response [7]. This kind of observation offers the chance for the evaluation of treatment response with complete pathologic response acting as a surrogate marker of survival, as well as representing an excellent model to determine the predictive role of tumor characteristics thanks to its correlation with clinical outcome [8]. Establishing the response to therapy is therefore a crucial step, and can be done through clinical and pathologic examination, and using breast imaging studies. Since neoadjuvant chemotherapy induces changes in the morphology of the tumor, the distribution of residual tumor may change in studies acquired at different time points during the therapy process. For this reason, the choice of the imaging modality should be taken according to its ability to clearly demonstrate the extent of disease at presentation, making MRI the most informative one for neoadjuvant treatment.

2.3 Magnetic Resonance Imaging (MRI)

One of the most important discoveries in the field of medical imaging, MRI, or *Magnetic Resonance Imaging*, uses the natural properties of the body's magnetic field to produce extremely detailed images of any part of the body. It is particularly suitable for disease detection, diagnosis, and treatment monitoring (especially in non-bony parts or soft tissues of the body), being a non-invasive technology that produces detailed, three dimensional anatomical images.

Using a powerful magnet, the protons in the body are forced to align with a strong generated magnetic field. In particular, the protons' axes, normally aligned in a random way, will be all aligned creating a magnetic vector oriented along the axis of the scanner. The strength of this filed is usually between 0.5 and 1.5 Tesla [9]. An additional energy, a radio-frequency current, is then used to stimulate the protons, which will now spin out of equilibrium, deflecting the magnetic vector. When the radio-frequency field is turned off, the realignment of the protons to the magnetic field causes a radio wave to be emitted. Receiver coils are used to detect this signal, plotting its intensity on a new generated grey scale, cross sectional image. The environment and the chemical nature of the molecules, determining specific tissues or abnormalities, produce different relaxation times (the time needed to realign after the radio-frequency is switched off). These different magnetic properties make it possible to distinguish between these tissues, such as water and fat. Different pulse sequences can emphasize different aspects, as in the case of "fat suppression", which will remove the signal from fat leaving only the abnormalities laying in it. Finally, a further distinction can be made on the basis of the proton relaxation time considered: T1 relaxation represents the time required for the magnetic vector to return to the state of rest, while T2 relaxation considers the return to the state of rest of the axial spin.

2.4 The Future of Medicine?

In a context where the exponential growth of available and processable data, coupled with astonishing results achieved by machine learning (Deep Learning in particular), places the emphasis on the possibility of automating human labour, specific considerations must be made for the medical field. Geoffrey Hinton, a charismatic figure in the artificial neural network field, stated that *"it is quite obvious that we should stop training radiologists"* and Andrew Ng affirmed how *"a highly-trained and specialized radiologist may now be in greater danger of being replaced by a machine than his own executive assistant"* [10].

While the available literature suggests how a considerable progress has been made in recent years, both in terms of accuracy and efficiency, in the use of CAD

systems (and more specifically in all the main steps of breast tumor detection and classification), many challenges still remain and many questions need to be answered regarding the integration of this breakthrough technology to the medical workflow. An investigation paper carried out by the European Society of Radiology (ESR) [11] has concisely listed most of the issues that machine learning has to face in the medical field: from the need for large, unbiased and annotated datasets, to the regulating issues (often inadequate to handle the rapidity with which algorithms evolve by using the same validation standards applied, as example, to new drugs). Furthermore, medico-legal responsibility should not be underestimated, with the question "who is responsible for the diagnosis" more relevant than ever due to the introduction of outputs from "black-box" algorithms to the clinical workflow.

It should become more and more clear how the figure of a radiologist can't be totally replaced by an algorithm, however the impact of machine learning is undeniable and potentially decisive into shifting the expertise of trained radiologist from more time-consuming and mechanical tasks to more sophisticated and useful ones. The only radiologists at risk of losing their job are the ones who refuse to work with AI [12].

Chapter 3

Deep Learning applied for Breast DCE-MRI

The complex nature of MRI sequences, which constitute a source of information in four dimensions (three spatial dimensions plus the temporal one), provides a valuable source of information in the medical field, while introducing an element of difficulty for their correct interpretation, and not only by a neural network.

Reading and analyzing such information-rich sequences, leading to a diagnostic result, can be a very error-prone and tedious task for a human. The introduction of Computer-aided Detection and Diagnosis systems (CAD) in the clinical workflow can potentially reduce the radiologists' workload, reduce the inter- and intra-observer variability and provide a second reading, by automating some of the diagnostic tasks.

A typical CAD system is composed by two main phases: the analysis one and the diagnostic one. During the former, after a pre-processing stage, anatomical structure of the images and their features are extracted. This can be achieved by delineating a region of interest (ROI) and acquiring its features (CADe). These features are then used in the diagnostic phase, by integrating them in a classification procedure (CADx). The basic components of a CAD system are, therefore: pre-processing, segmentation, feature extraction and selection, and classification [13].

The aim of the following analysis is to explore the available literature about these "modules", retracing all the steps from image acquisition to tumor classification and treatment response prediction in the field of breast DCE-MRI, focusing on the state-of-the-art approaches, most of them exploiting a Deep Learning solution.

3.1 Applications in the Medical Field

We can define radiomics as the conversion of medical images into high dimensional mineable data [14]. Traditional machine learning methods require hand-engineered features, both quantitative and qualitative, composing a "radiomic signature" used to learn patterns in data. While attaining promising results even in the specific field of the prediction of pathologic response to neoadjuvant therapy in breast cancer, as Cain et al. [15] demonstrated by training two multivariate machine learning models (logistic regression and SVM) only on features of pre-treatment MRIs, or in still not fully explored contexts such as multi-parametric MRIs, where Tahmassebi et al. [16] achieved high accuracy and stable performances testing eight different classifiers, the trend in medical imaging analysis is rapidly shifting towards Deep Learning.

Even if the use of Deep Learning in the era of big data analytic in the context of medical imaging is far from fully documented, the ability provided by Deep Learning algorithms to remove human knowledge in parsing any substantial information is a crucial step [17]. Being able to learn features as data representatives, directly from raw medial images, without additional effort or prior knowledge is indeed a key characteristic of Deep Learning algorithms, that if combined with the increasingly available computational power explains the growing popularity of this approach in the medical imaging field, and also in the specific breast DCE-MRI one. But, although several overview papers such as Lo Gullo et al. [18] and Debelee et al. [19] have stated the impressive performance of Deep Learning, especially Convolutional Neural Networks, compared to traditional machine learning, several new challenges need to be tackled.

In spite of the automated feature extraction advantages, data centrality poses additional challenges that can be especially hard to solve in the medical field, one of them representing the need of large datasets to provide sufficient training samples. On one hand, the performance of CNNs in this field increases when more training data is available, contrary to radiomics algorithms [20], on the other hand Hamidinekoo et al. [21] schematically indicated in their overview paper that more standardized procedures, both in technical phases like image acquisition by scanners and annotation operations for an appropriate ground truth, as well as a better interpretability of information provided by the layers of the model are still required for building suitable datasets. Finally, it is worth noticing how several studies have assessed that an optimal result is achieved when CNN are used for fusion classification together with human-engineered radiomic features, suggesting that prior knowledge, when available can still make a substantial difference [22].

All these advantages and future challenges will become clearer while analyzing the implementation choices of state-of-the-art Deep Learning approaches used in the CAD pipeline for breast DCE-MRI, here taken in consideration.

3.2 Pre-processing

Each step of the MRI-based radiomics workflow lacks standardization in the published literature. Granzier et al. [23] in their systematic review, using a Radiomics Quality Score (RQS) applied to different studies, stated that the overall promising state-of-the-art results are difficult to compare, due to the great methodological differences especially in segmentation, feature selection and model development. In particular, considering the information-rich data provided by DCE-MRI, very different choices can be made to exploit them: focusing on the 3D spatial aspect, or on different slices extracted by them rather than on the temporal dimension, using images acquired at different time-points. These choices, that represent a crucial aspect of the state-of-the-art approaches, will be further explored when the specific studied solutions will be described, both in lesion detection and in classification.

However, preliminary low-level operations on the image aimed to reduce noise and improve quality represent a crucial pre-processing step described in all available literature references. These operations also include a first segmentation step, involving the whole breast, which by exploiting anatomical features extracts a mask only representing the breast parenchyma, thus removing other tissues, the pectoral muscle and the chest wall.

3.2.1 Image Normalization and Denoising

Since Deep Learning algorithms are very sensitive to these factors, essential image pre-processing operations are image resizing, as well as normalization (by excluding extreme values, transforming the value range, subtracting the mean and dividing by the variance) and bias-correction to eliminate the heterogeneity of light distribution.

It must be noted that the peculiar characteristics of a DCE-MRI dataset usually involve additional steps even in this phase, combining the traditional normalization used for Deep Learning networks with more domain specific applications. The introduction of a contrast agent, implicating a considerable variation along the temporal dimension, generates a different intensity distribution in the post-contrast frames, compared to pre-contrast ones. A subtraction between the former and the latter is often performed, aiming to emphasize the contrast enhancement while suppressing the constant background.

Applying a denoising deep neural network like DnCNN on the normalized images could also be a viable solution [24].

3.2.2 Breast Volume Segmentation

The process of breast volume segmentation in MRI images is not so straightforward, and many automated or semi-automated solutions have followed across the years,

from conventional image processing ones, to conventional machine learning all the way to Deep Learning.

The removal of unwanted voxels (hence reducing the following computational effort) was initially achieved with pixel-based approaches, like the one based on Otsu's thresholding and morphological refinements post-processing proposed by Viganti et al. [25]. Other geometrical and atlas-based solutions are now outperformed by Deep Learning algorithms, in particular the ones exploiting the semantic segmentation provided by U-Nets [26].

Piantadosi et al. [27] automatically segmented the breast parenchyma from air and other tissues by applying a 2D U-Net to the 3D volume MRI data, feeding the architecture with a composition of different slices from different projection planes. The use of a 2D network instead of a 3D one allowed the saving of around 66% of trainable parameters, while achieving state-of-the-art result measured in accuracy, sensitivity and *Dice Similarity Coefficient* (DSC). The performance was also comparable on all projection planes. In a follow-up study [28] this solution was further expanded by segmenting the three projection planes with different U-Net networks, using a multi-planar approach by merging these three outputs with a combination rule (the *Weighted Majority Voting* strategy performed best).

Xu et al. [29] also assessed the potential of U-Net breast segmentation, specifically on transverse fat-suppressed DCE-MRI, enhancing the performance with appropriate post-processing, exploiting the segmented breast candidates' volumes (in the choice making step the smaller one was deemed as a scar and discarded).

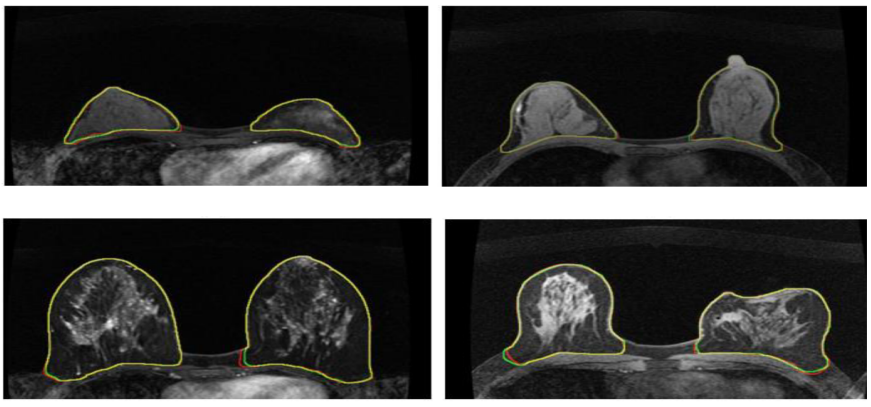


Figure 3.1: Segmentation results. Red line: the segmentation of proposed method; green line: manual segmentation; yellow line: the overlap of green line and red line (extracted from the original paper by Xu et al. [29]).

Zheng et al. [30] helped the 2D U-Net in the still difficult task of identifying the boundary between breast and pectoral muscle on MRI images by adding, along with the pre-processed MRI slice, two spatial coordinates indicating the breast

position. The middle point of the breast-air boundary was selected as the origin of coordinates, allowing the so created coordinate-guided U-Net to perform better than a traditional one, especially by avoiding false positives.

The advantages of U-Net were also extensively reported in [31], where no significant benefits emerged in the use of 3D U-Nets over 2D ones and in the use of pre-contrast T1 weighted images with or without fat suppression.

3.2.3 Motion Correction

Once the *volume of interest* (VOI) is efficiently cropped, the next pre-processing step requires a *motion correction technique* (MCT) to perform image registration: to reduce motion artefacts that may occur in the image acquisition stage, the volumetric images acquired at different time-points need to be aligned, in order to better compare their respective information. This can be a crucial operation especially when tumor volumes are evaluated before and after chemotherapy, or when a subtraction between a post and a pre-contrast image needs to be done.

In most of the literature examined, motion correction operations were not specified, or simply achieved with an affine 3D transformation. Antonio Galli et al. [32] assessed the impact of motion correction on Deep Learning approaches used on breast MRI-DCE, suggesting that a simple MCT can still increase the performance of lesion segmentation tasks, while having an almost irrelevant influence on lesion classification.

The use of Deep Learning for the task of image registration is still new and barely covered in literature.

3.2.4 Data Augmentation

In a context where the lack of training data is a major issue and large annotated dataset are often unavailable, data augmentation can represent an essential solution. Through transformations like rotation, translation and flipping (which are particularly suitable for MRI images), new training images can be created, improving the performance of Deep Learning networks while also avoiding overfitting. This can also help in the task of balancing the often-unbalanced classes within medical datasets. Almost the totality of the current literature used this pre-processing technique.

An innovative approach could be represented by the use of *Generative Adversarial Networks* (GANs), able to perform photo-realistic image synthesis. Guan and Loew [33] explored the application of GANs both for image augmentation and transfer learning to improve the performance of a CNN classifier, using a breast mammogram dataset: the generated GAN ROIs helped the training process, avoiding overfitting, and the image augmentation provided by GANs was shown

as necessary to train a CNN from scratch. However, they did not succeed into training a CNN with transfer learning only on generated ROIs.

The application of this image synthesis approach on MRI image data was tested by Shin et al. [34], by generating abnormal brain tumor MRI images with a GAN based on pix2pix. Improvements in segmentation performance were measured, as well as allowing the training on a completely anonymized dataset.

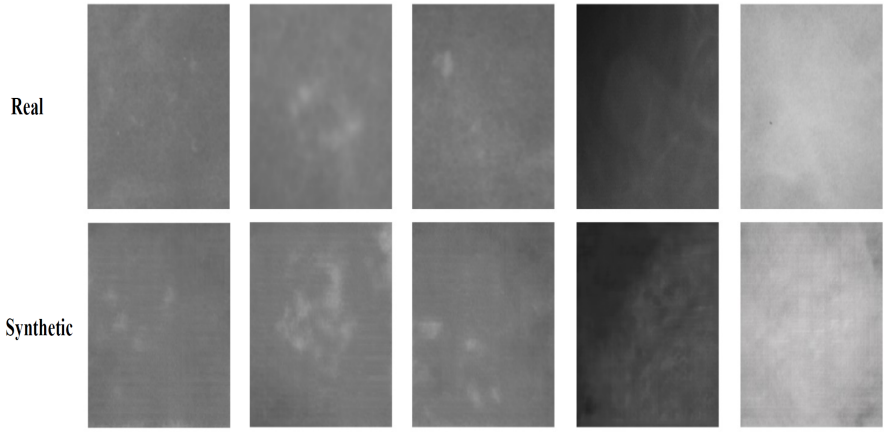


Figure 3.2: Real and synthetic abnormal ROIs generated from GAN (extracted from the original paper by Guan and Loew [33]).

3.3 Lesion Detection

An accurate breast lesion detection through segmentation is a decisive step in a CAD workflow, representing an essential tool in providing a meaningful input to the classification module, in extracting a ROI to perform quantitative radiomics analysis or into evaluating tumor region changes. This challenging task, made difficult by the complexity of the background breast tissue and by the tumor shape irregularity, is now approached with Deep Learning methods, thus avoiding manual segmentation and classical machine learning methods dependent on handcrafted features.

As in the case of breast segmentation, the U-Net architecture seems to be the most common choice in the latest literature, being able to perform an efficient tumor segmentation using MRI slices, both in a 2D and 3D U-Net fashion [35]. However, very different implementation choices are available within this same architecture, based on how to exploit the 4D data provided by DCE-MRI and the different clinical screening settings. Dalmış et al. [36] focused only on the early-phase scans, extracting spatial information from them: given a MRI axial slice, a U-Net outputs

a likelihood map for each voxel, combining those maps provides a likelihood volume used to select a candidate. Among those candidates, false positives are further reduced through the use of a 3D CNN, able to exploit the 3D morphology of the candidate regions.

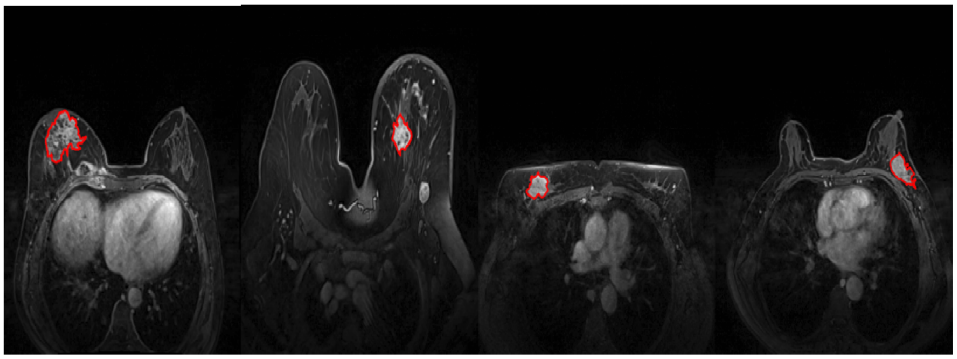


Figure 3.3: Sample slices showing the great variation inherent in the task of lesion detection (extracted from the original paper by Zhang et al. [35]).

Lu et al. [37] instead took advantage of four different image modes from breast MRIs (T1W, T2W, DWI, and SYN), building a four-mode linkage backbone with a CNN (DenseNet, with transfer learning and data augmentation performed best) used for feature extraction. The features extracted from the four different image modes are then concatenated and used to refine the tumor segmentation task by initializing a U-Net with a modified upsampling process, which uses a sub-pixel method.

The temporal dimension of breast MRI plays a key role in [38], where the tumor segmentation is still achieved with a U-Net, but with a training that uses three different, well-defined temporal acquisitions of the same MRI slice as channels of the input image. This approach obtained a good compromise between results in terms of DSC and efficiency provided by the simple architecture (no additional nets are required to train temporal features).

An attempt of using both sides of the spatio-temporal information was made by Chen et al. [39], using a convolutional long short-term memory network for feature extraction. This architecture is particularly suitable for processing sequence data with spatial correlation, and the use of three parallel ConvLSTM pathways, each given respectively the current, preceding and subsequent sequence of the same slice, allows the extraction of 3D information around the lesion. After a feature fusion stage, a U-Net with only two pooling operators is trained for lesion segmentation. Besides a promising result provided by the dice coefficient, this work also assessed how a dice coefficient loss is more suitable to solve the unbalance problem in breast lesion segmentation.

A radically different approach is taken in consideration by Maicas et al. [40], proposing a deep reinforcement based method for accurate lesion detection with less inference time. A deep Q-network is used to decide the next segmentation action, evaluating the features extracted from a bounding box by a ResNet network "agent". The DQN network can decide to translate/scale the bounding box, or to end the search for the lesion. State-of-the-art performance is achieved.

3.4 Lesion Classification

The final step of a CADx system is the classification of the lesion, meaning the binary classification of a lesion as benign or malignant. An automated approach that provides the radiologist with a valuable aid can be particularly useful considering the complex nature of a breast DCE-MRI scan, where even the slightest change in the lesion's temporal enhancement pattern can be decisive in determining the nature of the lesion itself. The literature trend of the last years shows, as predictable, a strong trend towards Deep Learning approaches, exploiting automatic feature extraction from the segmented lesion ROIs. However, some studies still value the additional support provided by classical hand-engineered features.

The first implementation of a CNN for this task can be tracked back to Antropova et al. [41], which already adopts one of the still most common architectural choices in this field, in relation to one of the main problems of medical imaging, namely the lack of training data: a pre-trained CNN is used. An AlexNet architecture, pre-trained on the ImageNet database, is employed as a feature extractor from ROIs, features that are then used as input for a *Support Vector Machine* (SVM), classifying the lesions. The obtained area under the receiver operating characteristic curve (AUROC), useful as a metric because independent of cancer prevalence, was 0.85, showing promising potential. The use of Deep Learning architectures in this task rapidly increased, showing again different variations based on how to 4D data was used.

In [42], a mixture ensemble of CNN has been employed to efficiently process the high-dimensional data, with a fixed number of "expert" networks focusing each on a region of the input space and a "gating" network used to properly fuse their outputs. All the CNNs are trained together in an end-to-end optimization approach, achieving good performance and fast execution time.

The temporal dimension plays a key role in [43], where three different time-points of the same ROI are used as input in the color channels of a pre-trained VGG19 network, operating as feature extractor. Specifically, features are extracted from the five max-pooling layers, and then average-pooled along the spatial dimension, before as being used as input for an SVM. It is also worth noticing how the best results were achieved when traditional hand-crafted CADx features were fused

together with the CNN extracted ones.

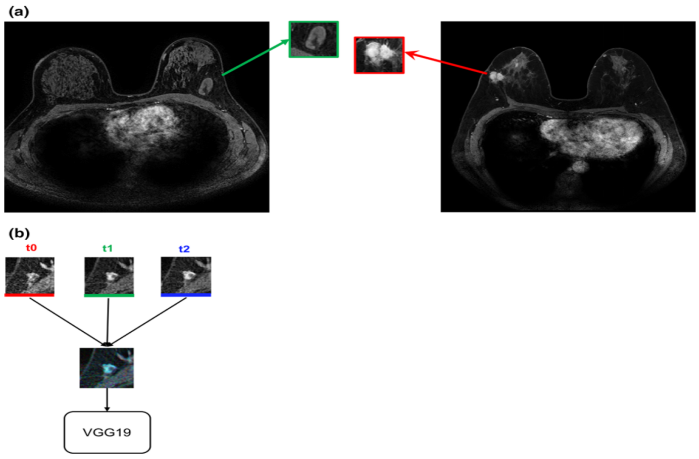


Figure 3.4: (a) Examples of slices with the corresponding ROIs extracted. Benign case on the left, malignant one on the right. (b) ROIs, extracted from the pre-contrast time-point (t0) and the first two post-contrast time-points (t1, t2), input into the three color channels of VGG19 (extracted from the original paper by Antropova et al. [43]).

In a follow-up study, Antropova et al. [44] took another step forward, improving the accuracy, by also including the spatial dimension along the temporal one, exploiting *Maximum Intensity Projections* (MIP): the idea is to use the same architecture as before, but with MIP images as input for the VGG19. They are obtained by a subtracted MR image collapsed by selecting the voxel having the maximum intensity along the projection through all transverse slices, forming a 2D image from a 4D DCE-MRI, therefore retaining information about enhancement changes throughout the whole lesion volume. Hu et al. [45] obtained even better results by max-pooling CNN features from all slices of a given lesion, reducing the image at feature level on the axial dimension instead of image level, as in the MIP case.

Long Short-term Memory (LSTM) networks were also tested in the task of lesion classification, allowing morphological and temporal information to be captured by extracting features from different time-points of the same MRI ROI using pretrained CNNs, one for each. The extracted features are then used to train the LSTM network, outperforming in this way a fine-tuned VGGNet [46].

Besides the focus on Deep Learning, knowledge provided by the radiomics field has regained great consideration in recent studies, but this time integrated along with the data extracted features. Gravina et al. [47] followed the radiomics methodology of the "three time points", using well defined time-points of a single slice as channels for the input image to a fine-tuned CNN, combining all the outputs

from all the slices of a single lesion with a specified rule, to classify each one as benign or malignant. The results are independent of the image acquisition protocol used and of the selected pre-trained CNN, as long as three different temporal acquisitions are available and the network adopts a three-channel input layer the approach remains valid. In [48] instead, the inclusion of some peritumor tissue, adjacent to the ROI containing the tumor, has been considered: this approach further demonstrated how Deep Learning methods have the potential to outperform ROI or radiomics based ones in diagnostic accuracy, and by evaluating different bounding boxes showed how the inclusion of a relatively small amount of peritumor tissue along with the tumor itself as input ROI for a pre-trained CNN can provide useful information, boosting the accuracy.

Feng et al. [49] included some domain knowledge along the data-driven features, to better relate those otherwise "black box" features to a clinically relevant phenomena. The main idea is to divide the DCE-MRI sequence in different sub-sequences, each one specifically focusing on a different feature type (exploiting pre-processing techniques), constraining in this way those features to semantic characteristics of the lesion. After this sequence division module, an adaptive weighting one is used for feature integration. The different sub-sequences are individually processed by different Deep Learning architectures, chosen accordingly to the specific highlighted characteristics of the lesion. The weighting will then integrate all those features and has also the ability to determine which sub-sequence contributed the most to the classification choice. This approach achieved state-of-the-art results on the used dataset in terms of sensitivity, specificity and accuracy, showing the importance of implementing domain knowledge into Deep Learning.

Further attention to feature extraction is given in [50], where a CMSL-driven deep network is used to learn more separable inter-class features and more compact intra-class features, tackling the heterogeneity problem of tumors. CMSL, standing for "cosine margin sigmoid loss", embeds the deep feature vectors onto a hypersphere and learns a decision margin between classes in the angular feature space. A 3D ResNet, trained with CMSL, outperformed other architectures, learning more underlying feature patterns in data.

3.5 NAC Response Prediction

Finally, particular attention was paid to the recent literature dealing with the prediction of NAC treatment response. In this scenario prediction of pathologic complete response is crucial, and other considerations besides the binary classification of the segmented lesion need to be considered, such as taking into account and comparing the follow-up DCE-MRI scans with the pre-treatment ones, rather than evaluating only the initial scans to perform the prediction process.

A first evaluation on how to exploit the temporal dimension of DCE-MRI scans in relation to the prediction of response to neoadjuvant chemotherapy was done by Huynh et al. [51] by feeding a pre-trained VGGNet with different subsets of MRI data (acquired before and after treatment), separated in pre-contrast, first and second post-contrast time-points (the different combinations, as well as the integration in a RGB channel where also evaluated). The performance of this features was tested by an *Linear Discriminant Analysis* (LDA) classifier, showing how ROIs from the pre-contrast time-point subset are particularly useful.

Ha et al. [52] explored the capability of a CNN to predict NAC treatment response by only using breast DCE-MRI scans obtained prior the initiation of the therapy, achieving an accuracy of 88%. Ravichandran et al. [53] performed a similar attempt on pre-treatment MRIs with comparable results when considering both the pre and post-contrast phase, but further increased performance by adding clinical HER2 status to the CNN's consensus.

The use of two breast tumor MRI slices acquired before and after the first round of chemotherapy was analysed by El Adoui et al. [54] with the use of a CNN employing two VGG-like branches. Using slices from the appropriately pre-processed and aligned VOIs from both the DCE-MRI scans notably improved the classification accuracy.

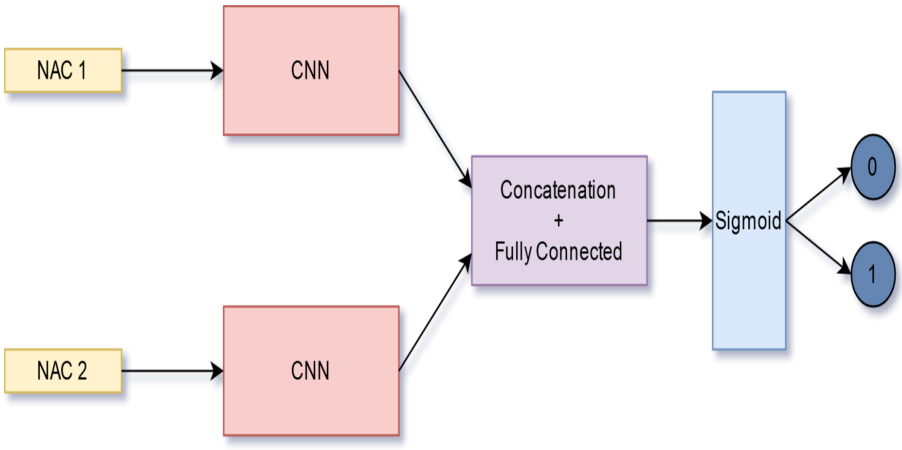


Figure 3.5: Overview of the architecture proposed by El Adoui et al. [54]. The branches take two 120 x 120 breast tumor ROIs. Each branch uses four blocks of 2D convolutuion (32 kernels for the first two, 64 for the remaining), with ReLu as activation function and Max Pooling.

In a follow-up study [55] the 3D nature of the data was exploited, modifying the CNN architecture: the two branches, fed with pre and post chemotherapy slices, are now repeated three times, processing respectively the axial, transversal and coronal slice of the DCE-MRI data. The proposed 3D approach achieved an AUC

value of 0.92, delivering a very good performance and demonstrating again the additional value of MRI data acquired after the first round of chemo.

The potential of Deep Learning architectures for this specific task, as well as that of the combined use of data from pre- and post-NAC studies is further confirmed by Byra et al. [56], showing how a Siamese network using data after the second cycle of NAC can improve prediction compared to a neural network based solely on studies acquired before the start of therapy. The best AUC value reported was 0.84, and regardless of the data considered, the performance of automated Deep Learning architectures was better than models using manually extracted morphological features. These results are further extolled by Choi et al. [57], using a simple architecture based on AlexNet to obtain good results on a dataset consisting not only of MRI but also PET/CT images. Several quantitative features are extracted and used as a baseline for predicting response to therapy, demonstrating the potential of CNNs and once again the usefulness of integrating studies performed after the first cycles of NAC.

Qu et al. [58] made another step towards performance improvement by building a Deep Learning model able to predict pCR to NAC from pre and post NAC MRI data: a multi-path CNN was inputted by six distinct enhancement phases from pre-NAC MRI and by other six from post-NAC. The individual feature extraction process for each of the 12 channels involved a concatenation of max-pooling layers and cropping operations. A molecular sub-type index was also added as an additional input channel in the feature concatenation stage, which did however not provide any benefits in terms of accuracy but could still carry valuable information for clinical practice. The combined model, using both pre and post NAC data, produced impressive state-of-the-art results in terms of AUC (0.97) and positive predictive value (100%), exploiting in the best possible way the changes inside the tumor, both in volume after the therapy and during the enhancement phases, dealing in the most efficient way such a difficult and crucial task.

The idea of a multi-path architecture is also explored by El Adoui et al. [59], developing two parallel CNN sub-networks, each analysing pre- and post-NAC studies separately. The extracted features are concatenated and used for pCR prediction, post dropout and a fully connected layer. The inputs taken into account are VOIs extracted from lesions shown in DCE-MRI images. An interesting experiment is done giving in input to the network these volumes both segmented (i.e. extracting only the tumour) and not (i.e. including besides the tumour also the surrounding tissue) analysing then the different performances. It is also the combined use of both pre- and post-NAC studies that records the best AUC (0.91). Interestingly, the non-segmented input obtains the best results, a result that is also confirmed by visual inspection made possible by Grad-CAM. In particular, the area surrounding the tumour seems highly relevant for non-pCR patients, while the tumour itself is more focused for pCR patients. The importance of the

region surrounding the lesion is emphasised not only by the latter but also by several studies analysed, both in the case of lesion classification and NAC treatment response, as it can provide information related to lymphocytic infiltration.

3.5.1 The NAC Baseline

The studies analyzed show how the use of Deep Learning applied to NAC response prediction can drastically improve the whole clinical workflow when used in conjunction with the knowledge of radiologists, which can be crucial especially for NAC where an early feedback provided to a radiologist can really make the difference in the care of a patient. The solutions differ greatly in terms of the input data used and the architecture employed, as well as the performances and the type of focus chosen for the output.

Starting with these results as a baseline, this thesis aims to build an architecture that exploits a mpMRI dataset as input (a case still not fully explored in literature), using only full slices, for assessing treatment response to NAC.

Network	Dataset (MRI Patients)	MRI Sequences Used	Input Type	Architecture	Performance
Tahmassebi et al. [60]	37	DCE, DWI, T2	Qualitative/quantitative features	XGBoost	AUC: 0.86
Huynh et al. [51]	64	DCE (3-Time-Points)	ROI crops	VGGNet	AUC: 0.85
Ha et al. [52]	141	DCE	Voxel crops	VGGNet-like	Acc: 88%
Ravichandran et al. [53]	166	DCE (pre-treatment only)	Patches from index slice	CNN with 6 blocks	AUC: 0.77
El Adoui et al. [54]	42	DCE	ROI crops	VGGNet-like	AUC: 0.96
El Adoui et al. [55]	42	DCE	ROI crops	VGGNet-like	AUC: 0.92
Qu et al. [58]	302	DCE	ROI crops	Custom Multi-path CNN	AUC: 0.97
El Adoui et al. [59]	42 + 14 external for validation	DCE, DWI	VOI crops	Custom Multi-input CNN	AUC: 0.91

Table 3.1: Characteristics and classification performance of other NAC studies.

Chapter 4

The Dataset

The analyzed dataset consists of 37 patients who were diagnosed with histopathologically established breast cancer and underwent NAC treatment between April 2008 and April 2013.

Some precise inclusion criteria were followed: each patient was over 18 years old, not pregnant, not lactating and had not undergone nor given any signs of contraindications for MRI or MRI contrast agents.

All patients underwent mpMRI; in particular, the sequences considered refer to a scan performed two weeks before the start of the first NAC cycle, and to one performed after two cycles of the neoadjuvant therapy.

The assessment of pathological treatment response was achieved through the residual cancer burden (RCB) score, with RCB 0 standing for pCR.

The RCB is an index that combines pathology measurements of the primary tumor and nodal metastases and has been shown to be effective for the prediction of disease recurrence and survival across all breast cancer subtypes [61].

Originally defined by Symmans et al. [62] as a continuous variable calculated on the bi-dimensional diameters of the primary tumour bed, on its proportion that contains invasive carcinoma, on the number of axillary lymph nodes containing metastatic carcinoma and on the diameter of the largest metastasis, it is a reliable indicator for assessing pathological treatment response, especially after NAC treatment [63].

Out of the 37 patients 9 achieved a pCR with no evidence of residual disease, 28 did not.

4.1 MRI Modalities

The mpMRI scans of the breast have been acquired at 3 Tesla in the prone position (Trio Tim; Siemens Medical Solutions, Erlangen, Germany) using a dedicated

4-channel breast coil (In Vivo, Orlando, FL), providing high-resolution and high-quality imaging of both breasts.

The following protocol was used before and during NAC:

- **DCE-MRI**: a hybrid DCE-MRI protocol was used with the following sequences: until December 2011 a T1-weighted volume-interpolated breath-hold examination sequences (TR/TE, 3.62/1.4 milliseconds; FOV, 320 mm; 72 slices; 1.7 mm isotropic; matrix, 192 x 192; one average; TA, 13.2 seconds per volume; 37 measurements) and T1-weighted turbo fast low-angle shot 3-dimensional sequences with selective water excitation (TR/TE, 877/3.82 milliseconds; FOV, 320 mm; 96 slices; 1 mm isotropic; matrix, 320 x 134; one average; TA 2 minutes) with a total time of acquisition of 9:20 minutes [64]. From January 2012, a transversal T1-weighted time-resolved angiography with stochastic trajectories was acquired (water excitation fat saturation; TR/TE, 6.23/2.95 milliseconds; flip angle, 15 degrees; FOV, 196 x 330 mm2; 144 slices; spatial resolution, 0.9 x 0.9 x 1 mm; temporal interpolation factor 2; temporal resolution, 14 seconds; matrix, 384 x 384; one average; center k-space region with a resampling rate of 23%; re-acquisition density of peripheral k-space of 20%; and TA, 6:49 minutes).

- **DWI**: a double-refocused, single-shot echo-planar imaging with inversion recovery fat suppression: TR/TE/time of inversion, 13700/83/220milliseconds; FOV, 340 117mm; 40 slices at 3.5mm; matrix, 192 x 64 (50% oversampling); 2 averages; b-values, 50 and 850 s/min2; and TA, 3 minutes 19 seconds.

- **T2**: a T2-weighted turbo spin echo sequence with fat suppression: time of repetition (TR)/time of echo (TE), 4800/59 milliseconds; field of view (FOV), 340 mm; 44 slices at 4 mm; flip angle, 120 degrees; matrix, 384 x 512; and acquisition time (TA), 2:35 minutes.

The total MRI examination time for each study lasted from 10 to 12 minutes. A standard dose of gadoterate meglumine (Gd-DOTA; Dotarem; Guerbet, France) of 0.1 mmol/kg body weight was injected intravenously as a bolus at 4 mL/s followed by a saline flush.

The great heterogeneity of the dataset, which differs profoundly in the protocols used in addition to the obvious and inevitable variations inherent in the acquisitions made on the various patients, represents at the same time the greatest challenge to overcome and the greatest asset to build an architecture capable of making predictions on new data.

Being mpMRI studies, that is multi-parametric, for each patient a potentially large amount of information is available to be extracted from different sequences (DCE-MRI, DWI and T2), each of which could provide significant indicators to predict the response to NAC therapy.

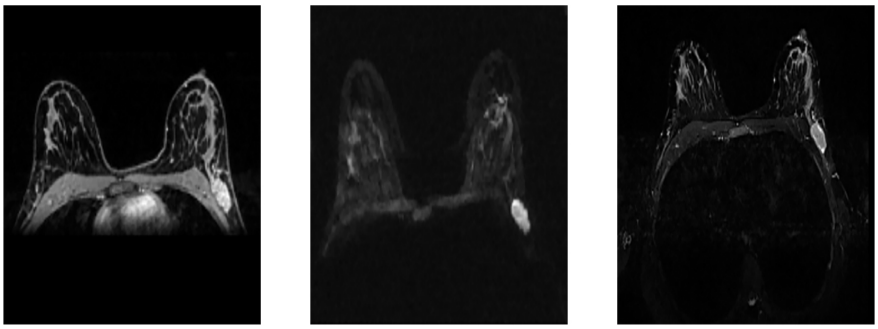


Figure 4.1: Example of slices extracted from sequences belonging to the same pre-NAC study. Starting from left: DCE-MRI, DWI, T2.

Dynamic Contrast Enhanced-Magnetic Resonance Imaging (DCE-MRI) has been shown as a powerful solution in screening different tumor tissues, gaining an increasing popularity in the field of breast cancer early detection, due to the high sensitivity and high resolution in dense breast tissues as well as the characteristic of retaining high 3D resolution and dynamic information. A 3D volume of the breast is acquired at different times, before and after of the intravenous injection of a contrast agent, thus resulting in a 4D volume (three spatial dimensions and a temporal one).

This non-invasive acquisition technique allows the visualization of the extent of disease and its angiogenic properties, as well of visualization of lesion heterogeneity, detection of changes in angiogenic properties before morphological alterations, and can carry very useful information for the prediction of overall response either before the start of therapy or early during treatment, as extensively documented by Turnbull [65].

DWI, standing for *Diffusion-Weighted Imaging*, can instead provide insights into tissue micro-structure by visualization and quantification of water diffusivity. It allows the evaluation of the *Apparent Diffusion Coefficient* (ADC), which has been proven helpful for cancer sub-typing in breast cancer patients and in prediction of response to neoadjuvant chemotherapy. Spick et al. [66] showed that DWI of breast lesions can provide consistent characteristics to support its use as a potential QIB (Quantitative Imaging Biomarker), being an objective characteristic derived from an in vivo image that can be measured and used to indicate a biological process, disease process, or drug response.

Finally, the T2-weighted sequences can provide a valuable aid in false-positive rate reduction, distinguishing well-circumscribed breast carcinomas from common

benign breast masses. In addition, edema, mucus, hemorrhage and cystic fluid within a lesion are clearly depicted on T2-weighted sequences [67].

Adding a further level of available information, as well as complexity, to the dataset is the temporal nature of a prediction based on response to NAC therapy: for each patient two studies are available, carried out at different times (two weeks before the first cycle of NAC and after two cycles) using the same acquisition modalities.

4.2 Machine Learning with mpMR for Early Prediction of Response to Neoadjuvant Chemotherapy

In 2018, Tahmassebi et al. [60] conducted a joint study by Florida State University and the Medical University of Vienna, stating the potential of machine learning to perform early prediction of pathological complete response to NAC on the same dataset used for this thesis.

Since the objective of the latter study is the same as that of the architecture proposed in this thesis, but using classical machine learning algorithms instead of Deep Learning, it is important to deepen the results already achieved, inspecting them taking into account the various substantial differences.

The study used 8 machine learning algorithms (logistic regression, support vector machine, linear discriminant analysis, decision tree, random forest, stochastic gradient descent, adaptive boosting and extreme gradient boosting "XGBoost") to predict the pCR based on the RCB class on the mpMRI data.

In order to properly train and use these algorithms, intensive feature extraction work is required to facilitate the learning process from the input data. These hand-crafted features inevitably add a semi-automatic component to the pipeline, requiring the prior intervention of experienced personnel, which can be very time-consuming. In particular, two experienced radiologists extracted both qualitative and quantitative features from the imaging data (23 features for each lesion). As features are measurable properties of the data, in the case of an mpMRI dataset they will include parameters such as lesion size, signal intensity, presence or absence of edema, shape, margins, as well as mean plasma flow, volume distribution and the mean, minimum, and maximum ADC coefficient. By using recursive feature elimination an optimum ranking of the features has been computed, reflecting in this way their importance in the model and showing how both quantitative and qualitative features, containing information found in all the available image modalities (DCE, DWI, T2), are necessary.

Regarding the classification process, the best results were yielded by XGBoost. XGBoost is a machine learning algorithm developed by Tianqi Chen and Carlos

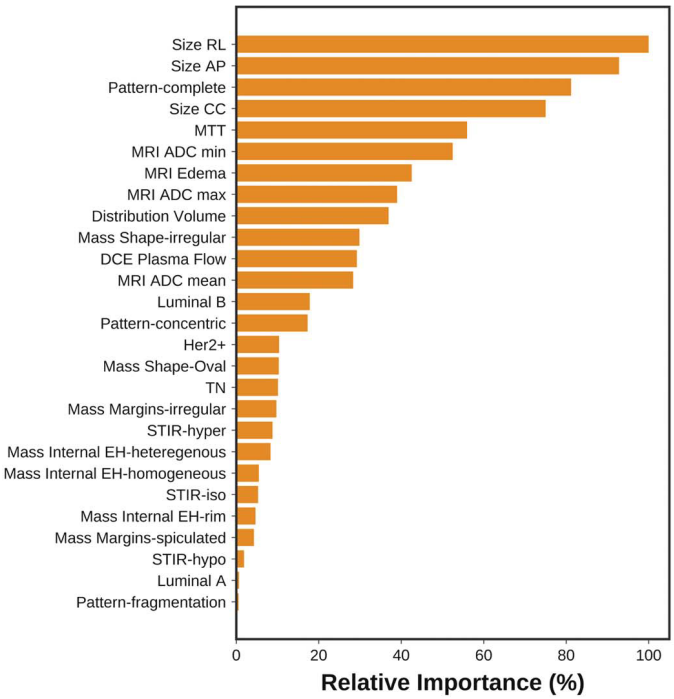


Figure 4.2: Feature importance of mpMRI model in prediction of RCB class (extracted from the original paper by Tahmassebi et al. [60]).

Guestrin [68] which gained popularity for prediction problems on small-to-medium structured data, exploiting its decision-tree-ensemble with gradient boosting nature.

Classifier	RCB
XGBoost	0.9430 (0.8577)
AdaBoost	0.8523 (0.8112)
Linear SVM	0.8767 (0.8450)
LDA	0.7544 (0.6608)
LR	0.8684 (0.8207)
SGD	0.8303 (0.7086)
Decision tree	0.8113 (0.7729)
RF	0.8857 (0.8364)

Figure 4.3: Best AUC (mean AUC) reported for each classifier (extracted from the original paper by Tahmassebi et al. [60]).

On the mpMRI NAC dataset, taking full advantage of the hand-crafted features, the model achieved a good performance on pCR prediction, with a mean AUC value of 0.86 based on 4-fold cross-validation.

These results constitute an important reference point, not only to highlight the

potential of machine learning applied to a multi-parameter MRI dataset, but also to build a new Deep Learning architecture that extracts features from the same dataset in a totally different way.

Patient	DCE (MRI1)	DWI (MRI1)	T2 (MRI1)	DCE (MRI2)	DWI (MRI2)	T2 (MRI2)
1	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
2	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn_fs	Resolve_0_850_FS+	tirm_tra
3	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
4	t1_dyn 2min	ep2d_diff_stir_10b	tirm_cor	t1_dyn 2min	ep2d_diff_stir_2b_sl3.5	tirm_cor
5	t1_dyn 2min	ep2d_diff_stir_10b	tirm_cor	t1_dyn 2min	ep2d_diff_stir_10b	tirm_cor
6	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
7	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
8	t1_dyn 2min	ep2d_diff_stir_2b_sl3.5	tirm_cor	t1_dyn 2min	ep2d_diff_stir_2b_sl3.5	tirm_cor
9	t1_dyn 2min	ep2d_diff_stir_2b_sl3.5	tirm_cor	t1_dyn 2min	ep2d_diff_stir_2b_sl3.5	tirm_cor
10	t1_dyn 2min	ep2d_tra_2b_spair	tirm_sag	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
11	TWIST_tra_dyn	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
12	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
13	t1_dyn 2min	ep2d_diff_stir_10b	tirm_cor	t1_dyn 2min	ep2d_diff_stir_10b	tirm_cor
14	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
15	t1_dyn 2min	ep2d_diff_stir_2b_sl3.5	tirm_tra	t1_dyn 2min	ep2d_diff_stir_2b_sl3.5	tirm_tra
16	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
17	t1_dyn_fs	Resolve_0_850_FS+	tirm_tra	t1_dyn_fs	Resolve_0_850_FS+	tirm_tra
18	TWIST_tra_dyn	Resolve_0_850_FS+	tirm_tra	TWIST_tra_dyn	Resolve_0_850_FS+	tirm_tra
19	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
20	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
21	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
22	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
23	TWIST_tra_dyn	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
24	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
25	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
26	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
27	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
28	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
29	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_sag
30	t1_dyn 2min	ep2d_diff_stir_2b_sl3.5	tirm_tra	t1_dyn 2min	ep2d_diff_PA_SPAIR	tirm_sag
31	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
32	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
33	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	TWIST_tra_dyn	Resolve_0_850_IR+	tirm_tra
34	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra	t1_dyn 2min	Resolve_0_850_IR+	tirm_tra
35	TWIST_tra_dyn	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
36	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra
37	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra	t1_dyn 2min	Resolve_0_850_FS+	tirm_tra

Table 4.1: Sequences available for each patient. Three for the pre-NAC MRI study, three for the post-NAC MRI study.

Chapter 5

Deep Learning Architecture

The proposed architecture aims to optimally utilize the abundant information found in mpMRI scans in order to correctly classify full slices containing a lesion, predicting whether or not pCR is achieved. The idea is to identify and extract sub-sequences for each patient, and then appropriately include the time dimension represented by the available second mpMRI scan, performed after two cycles of NAC.

5.1 Input Data

The multi-parametric scanning protocol used, allows access to different types of sequences, each with its own properties to be exploited in order to extract as much information as possible from the lesion. Specifically, Dynamic Contrast- Enhanced MRI (DCE-MRI) provides a complex sequence of rich lesion information that can simultaneously reflect morphological and hemodynamic related characteristics, while the Diffusion-Weighting Imaging (DWI) can reveal the free water molecules diffuse movement information of lesions, and the T2-weighting images have an intense signal for edematous lesions.

Let the dataset be represented by $D = \{P_i^w, y_i\}_{i=1^D}$ with P_i^w a patient and $y \in Y = \{0,1\}$ the label for the lesion, having value $y_i = 0$ when pCR is achieved and $y_i = 1$ otherwise. Since for each patient P_i two mpMRI studies are available, the index $w \in W = \{1,2\}$, represents the selected study, pre-NAC or post-NAC. For each of these studies we identify and extract S^k sub-sequences, with k index of a sequence. Finally, each sub-sequence is represented by X_j slices. The selected sub-sequences will thus be:

- Lesion slices extracted from DWI sequences.

- Lesion slices extracted from T2 sequences.
- Lesion slices extracted from the peak-enhancement phase of DCE-MRI sequences.
- Lesion slices extracted from DCE-MRI sequences according to the 3 Time-Points paradigm and combined into the RGB channel of a single image. No image registration is performed.

Therefore, we can represent the input data as $D = \{S_i^{w,k}, y_i\}_{i=1^D}$ with $w \in W = \{1,2\}$ ($w = 1$: pre-NAC study, $w = 2$: post-NAC study) and $k \in K = \{1,2,3,4\}$ ($k = 1$: DWI, $k = 2$: T2, $k = 3$: DCE peak, $k = 4$: DCE 3TP). Moreover, $S_i^{w,k} = \{X_{i,j}^{w,k}\}$.

$X_{i,j}^{w,k}$ thus denotes the *j-th* slice of the *k-th* sub-sequence in the *w-th* study of patient *i*.

Figure 5.1 shows an example of input, consisting of four slices extracted from the sub-sequences defined for both the pre-NAC and post-NAC study.

It is important to note that it is obviously not possible to use all the slices available for each scan as input, since this choice would inevitably introduce excessive noise in the learning phase, at the expense of valuable information.

To overcome this problem while limiting manual intervention on the data to the bare minimum, an index slices is extracted by an experienced radiologist (K.P. 14 years of experience) from each available sequence.

An index slice is nothing more than the slice within the sequence where the lesion is most visible. It can be identified with relative ease and efficiency, allowing to recognize the ideal input data with respect to the amount of information carried.

5.2 Architecture Structure

To optimally extract all lesion features, emphasizing the morphological, hemodynamic and water molecular diffusion information according to the specific sub-sequence, as well as the pre- and post-NAC studies for each patient, the architecture is designed to execute this process independently for each sub-sequence.

This should not only improve the performance of the architecture, but hopefully also provide an additional level of interpretability, indicating how much each of the sub-sequences affects the final classification choice with respect to pCR.

According to recent literature and given the size of the dataset, using a pretrained network for feature extraction should be the most feasible solution: ResNet50 pretrained on ImageNet is the most suitable choice.

Being probably the most disruptive innovation in the field of computer vision since the introduction of AlexNet [69], the ResNets theorized by He et al. [70]

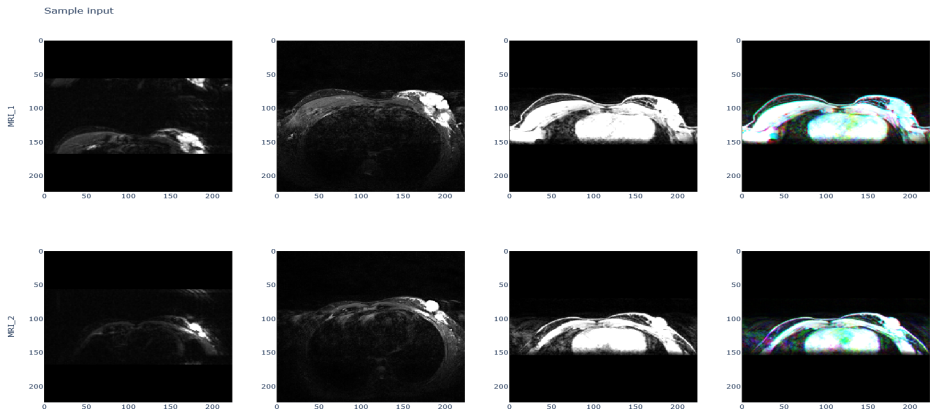


Figure 5.1: Example of index slices used as input, from left to right: DWI, T2, DCE-MRI peak and DCE-MRI 3-Time-Points. The top and bottom rows show the pre-NAC and post-NAC study of the same patient respectively. In the case of DCE-MRI 3-Time-Points (last column), the image is obtained by inserting slices from different and precise time instants into the three RGB channels. No image registration was performed.

break the taboo of an extremely deep but still performing and easy-to-train neural network.

The goal of increasing the depth of a network has always been countered by the problem of vanishing gradient during back-propagation. The repeated multiplications on many levels can in fact make the gradient infinitesimally small, saturating or even worsening the performances of the network with great rapidity.

The solution to the problem has been found moving the focus of attention from the stacked layers, considered at first as the possible culprits, to the mapping instead. The theory allows in fact the construction of a deep network with the same identical performances of a shallower one, using only identity mappings in the additional layers for which they differ. Working on the mapping instead it is possible to obtain great results of optimization to a practically null cost. The hypothesis is that letting the stacked layers fit a residual mapping is easier than letting them directly fit the desired underlying mapping. In a certain sense, the blocks in this way "simply" fine-tune the output of their previous block, instead of having to learn the output from scratch.

Networks employing these residual blocks with skipped connections are less vulnerable to perturbations that may cause to leave the manifold, thus avoiding the need for additional data to recover.

All these elements make a pretrained ResNet particularly appropriate for this architecture, having to extract in the most efficient way possible spatial features from

very complex and heterogeneous input data, such as MRI sequences of different protocols, constituting among other things a very small dataset and therefore prohibitive to analyze without the aid of transfer learning.

The Deep Learning architecture will be in this way composed by four ResNet50, one for each sub-sequence, from which features are extracted at the last convolutional layer, with dimension 2048. These features are then concatenated and, following some dropout and a fully connected layer, are used for the prediction of pCR.

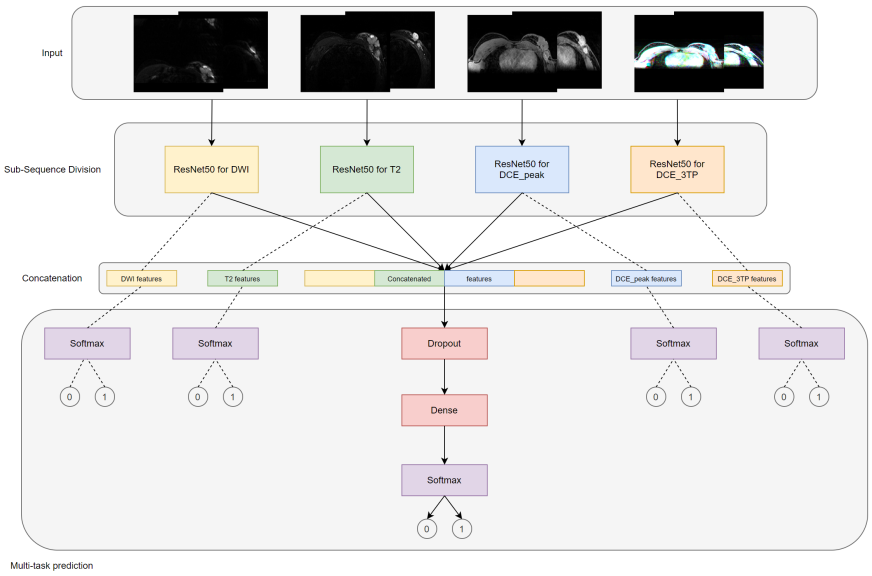


Figure 5.2: Overview of the proposed architecture. Four ResNet50 process sub-sequences, pre- and post-NAC. Each individual ResNet outputs a classification result for its sequence. The final classification regarding pCR is performed following the concatenation of features extracted from the ResNets, after some Dropout and a Fully Connected Layer.

It is important to note that the same sub-sequences, but from the two different studies (post and pre-NAC) available for the same patient, are analyzed by the same ResNet50, using weight-sharing. The assumption at the base of this choice is that two sequences coming from two studies of the same patient carried out in different temporal instants, but afferent to the same acquisition protocol, should produce about the same effect on the parameters of a ResNet. Weight-sharing should therefore ensure an almost identical result but with a great computational advantage given the large number of parameters to be optimized in the entire architecture.

5.3 Multi-Task Learning

The classification for pCR performed using the vector composed of all concatenated features is not the only one done by the architecture: each ResNet50 is also linked to a fully connected layer with two neurons, which will predict the result for the single sub-sequence. This sort of multi-task learning aims not only to improve the performance of the architecture in general, but also to provide an additional layer of interpretability.

The multi-task approach allows to extrapolate useful information contained in multiple related tasks, improving in this way the generalization capability of the entire network. Overviews on Multi-Task Learning for Deep Neural Networks have stated how this technique can, among other advantages, help the network's attention focusing (as more task will more likely provide additional evidence for the relevance or irrelevance of the features) and also introduce a sort of representation bias (since the model will prefer representations that other tasks also prefer) [71].

Again, having a dataset so heterogeneous in terms of acquisition mode and difference between the various full slices used as input, focusing on different tasks may prove to be a fundamental and necessary solution. For a given MRI study of a patient, it is in fact possible that a sub-sequence may introduce an unforeseen amount of noise: focusing on each sub-sequence as a single task can help the architecture to determine more precisely which of these sub-inputs to consider as more reliable to make predictions about the patient.

This could not only help to improve the training and the performance of the network, but could reduce the "black box"-effect generated at the output, showing the relative performance on each acquisition mode present in the mpMRI dataset, suggesting which of these provided the most significant features to make the final prediction.

5.4 Hand-Crafted Features Auxiliary Task

In the literature reviewed and in the computer-automated processing medical imaging field, the joint utilization of raw data and features engineered from these data based on specific domain knowledge, leads very often to optimal and easily interpretable results.

While the goal of this Deep Learning architecture is to perform the prediction in the most rapid and automated way possible starting only from the scans, it is possible to add an additional branch that uses as input also the hand-crafted features extracted by two expert radiologists from the same dataset, and then evaluate its overall contribution in terms of performance.

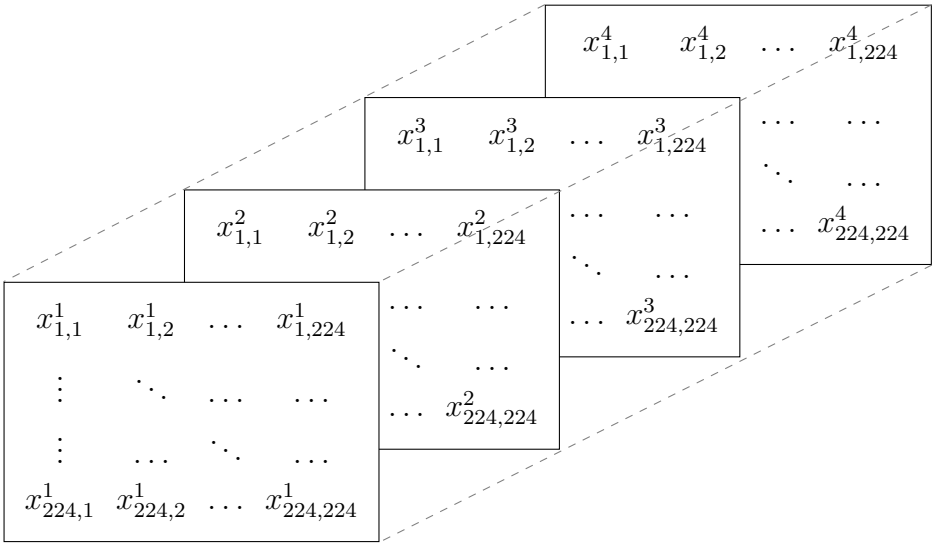
We used the same data as in the study done by Tahmassebi et al. [60], and in

particular the choice fell on the first 7 features ordered according to their importance in the machine learning algorithms used (see figure 4.2).

Each slice, whether it comes from the pre- or post-NAC study and regardless of the sub-sequence to which it belongs, in addition to the information relating to the image data will also carry as input the following numerical/categorical features:

- MRI Mean-Transit-Time.
- MRI ADC (min), MRI ADC (max).
- MRI size anterior-posterior, MRI size cranial-caudal, MRI size left-right.
- MRI edema.

Each input will thus have a shape of:



With x^k and $k \in K = \{1,2,3,4\}$ ($k = 1$: DWI, $k = 2$: T2, $k = 3$: DCE peak, $k = 4$: DCE 3TP) representing the values of each of the 4 sub-sequences input slice used, of shape 224 x 224 x 3 (representing respectively rows, columns and channels of the resized image data) and the additional feature vector:

$$\begin{array}{|cccc|}\hline f_1 & f_2 & \dots & f_7 \\ \hline \end{array}$$

If the four slices are each processed by a ResNet, the feature vector will be used as input for a simple MLP (multilayer perceptron) which will perform a simple binary classification with respect to the pCR prefix. Also in this case the

features extracted from this branch are concatenated with the others, before the aggregated output of all the branches. A graph of this additional branch can be seen in Figure 5.3, the features are concatenated in the same level where those extracted from the other branches using ResNets are concatenated.

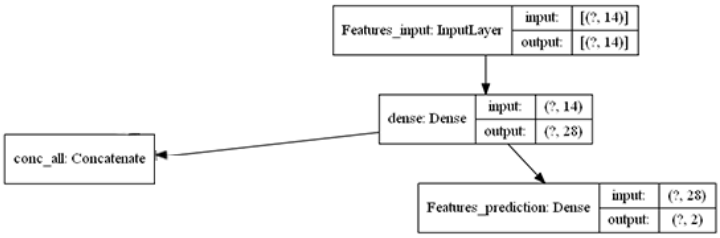


Figure 5.3: The axuiliary Features branch plot. The last fully connected layer is used for the task-specific prediction. The features extracted by the previous fully connceted layer are concatenated with the ones extracted by the 4 ResNets.

5.5 Grad-CAM

Finally, Grad-CAM is used to aid understanding of the choices made by the architecture during learning. Grad-CAM, developed by RR Selvaraju et al. [72], produces a gradient-weighted localization map, highlighting the most important regions of the image, decisive for the final choice in the classification task performed. This can be especially useful given the use of full slices. In this way, it is possible to test whether the architecture can independently focus its attention on the area containing the lesion, without any bounding box previously selected by a radiologist.

5.6 Experimental Settings

5.6.1 Axial Resampling

The first preliminary step was, starting from the index slices identified, to resample the entire dataset composed of MRI slices (in *.dicom* format) into the axial plane. The original sequences are in fact presented with different spatial orientations (axial, coronal, sagittal) and with the aim of having input data as uniform as possible, the choice fell on the axial plane, often considered as the easiest to use in the training phase of a neural network.

The operation was performed by working on the starting *.dicom* files and exploiting the attributes contained in them, in particular the fields *ImagePositionPatient*, *ImageOrientationPatient*, and *PixelSpacing* were suitably modified to perform a resampling in the axial plane that did not lose any information.

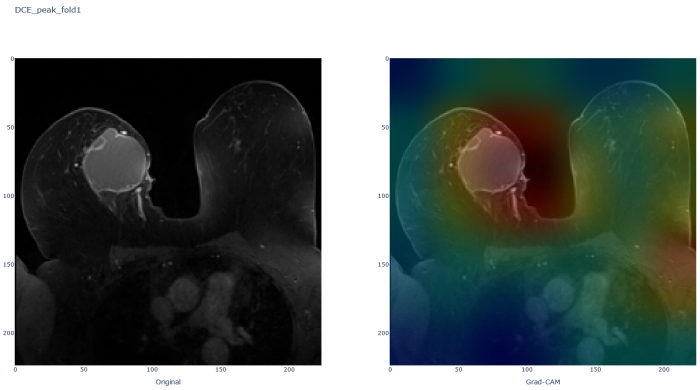


Figure 5.4: The Grad-CAM activation map for a slice of the DCE peak subsequence. The architecture seems to focus on the area of the lesion to make its decision

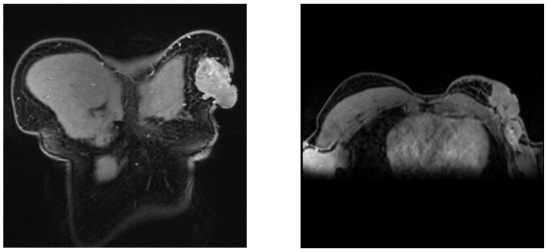


Figure 5.5: Example of coronal (left) to axial (right) resampling, computed for the same DCE-MRI index slice.

5.6.2 Resizing and Normalization

To achieve maximum compatibility with the pretrained ResNet50, all slices are resized to a fixed size of 224 x 224 x 3 with the gray-scale image repeated three times in the RGB channel for all sub-sequences, except in the case of the DCE-MRI 3-Time-Points. In this specific case three slices were selected from DCE-MRI sequences of the same study, but in three precise time instants according to the "Three Time Points" approach: t_0 pre-contrast, t_1 approximately two minutes after CA injection and t_2 approximately six minutes after CA injection. Each of these slices constitutes an RGB channel of the input image.

Given the heterogeneity of the input, with different sequences acquired by scans using different parameters, some signal normalization is necessary. As a pre-processing step, Z-Score Normalization is applied sample-wise. The distribution mean and standard deviation is calculated for each feature. Then, the mean is

subtracted from each feature and this value is divided by the standard deviation:

$$Z = \frac{X-\mu}{\sigma}$$

5.6.3 Data-augmentation

Considering the limitations of available data, data augmentation plays a key role in increasing training data.

By processing at an image level, data augmentation can be performed with elastic transformations as well as intensity variations. Specifically, given the symmetry of the breast lesion, a flip on the vertical axis is performed. Changes in signal intensity, and affine transformations are also executed.

In detail, the values used by the Data Generator used to perform the data augmentation are:

- Vertical flip = True
- Zoom in/out from the original image with a random value of max 0.3
- Brightness range (to change signal intensity) with a random value between 0.3 and 5
- Shear range (to perform an affine transformation) with a random value of max 0.3

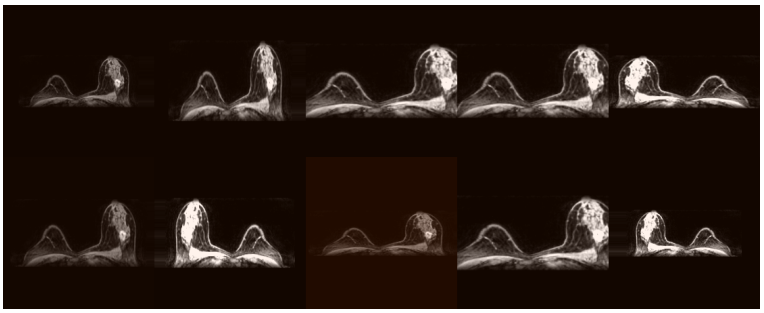


Figure 5.6: Data augmentation examples generated from a DCE-MRI peak slice

5.6.4 Input Slices

A very important parameter that can be set concerns the number of slices considered for each input scan. In fact, the index slice is always included but there is the possibility of adding more slices starting from it. For example, by setting the parameter $slices = 3$, in addition to the index slice the three immediately following

slices will be included as well as the three immediately preceding ones, bringing the total number to 7. The optimal value recorded is around 3: a lower number reduces excessively the information available to the model in the learning phase, and a higher number introduces an excessive number of slices where the lesion is not sufficiently visible, thus adding only noise.

5.6.5 Hyperparameters

The optimal number for batch size found in the experiments was 12, the learning rate 0.0001.

An adaptive learning rate strategy is also applied, with the learning rate being reduced by an order of magnitude when the loss on the validation set is in a plateaus. This evaluation is computed every 5 epochs.

The initial number of epochs is set at 100, however during the experiments the ideal number of epochs to use for effective model training without over-fitting was noted to be 15. A checkpoint containing model weights is saved after each batch if improvements are measured.

Alternatively, an "early stopping" strategy can be used, where the training is stopped earlier when no more improvements are found in the loss term of the validation set. In this case, at the end of training the weights of the model that reported the best results are loaded before the final prediction. However, in reporting performance, the choice made was to consider experiments using a fixed number of epochs so that the results are not overly based on validation folds.

To further counteract over-fitting, a very likely risk given the large number of parameters present in the architecture and the nature of the dataset, L2 regularization with $\alpha = 0.0001$ is applied to the convolutional layers of ResNet50, as well as dropout of 0.5 applied to the vector of concatenated features. Before the output layer for pCR prediction, a fully connected layer with 128 neurons is added, making use of *relu* as an activation function.

A summary of the hyper-parameters tested and the optimal values chosen can be seen in Table 5.1.

5.6.6 Loss Function

The model is trained in a end-to-end manner using *binary cross-entropy* as a loss function. Since the model produces multiple outputs, one for the pCR and one for each of the included sub-sequences (plus the optional hand-crafted features branch), the overall loss function takes into account all of these instances, each specifically weighted with a γ index. The main goal is obviously to predict the pCR, whose

weight will therefore be 1; the optimal measured γ value for sub-sequences is 0.2.

$$Loss = Loss_{pCR} + \sum_{k=1}^{K} \gamma Loss_k$$

A further loss function tested, given its affinity with the type of dataset used, is the *focal loss*. Originally proposed by Lin et al. [73] for *Dense Object Detection*, the focal loss can be considered as an extension of the cross-entropy loss function that would down-weight easy examples and focus on training on hard negatives. It is a dynamically scaled cross entropy loss, where the scaling factor decays to zero as confidence in the correct class increases. This can address the issue of the class imbalance problem, represented in the dataset by the larger number of slices belonging to patients who did not achieve pCR.

$$L_{focal} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

with $\gamma > 0$ tunable focusing parameter and

$$\alpha_t = \begin{cases} \alpha, & \text{if } y = 1 \\ 1 - \alpha, & \text{otherwise} \end{cases} \tag{5.1}$$

and α value between 0 and 1 used to balance the positive labeled samples and negative labeled samples.

5.6.7 Cross-validation

The dataset is split at the patient level in a 8/2 manner for training and validation. To overcome over-fitting, 4-fold cross-validation was used. In each split, the ratio of patients who achieved pCR to those who did not is kept as homogeneous as possible with the original distribution of the dataset. All lesions from the same patient were kept together in the same fold in order to eliminate the impact of using correlated lesions for both training and testing.

5.6.8 Evaluation Metrics

The evaluation metrics used are *binary accuracy* and *AUC*. In particular, the choice to use the *Area Under Curve of the Receiver Operating Characteristic* is due to the fact that the ROC Curve summarizes the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds. This is particularly useful for a dataset that presents a class unbalance like the one analyzed, in fact the chance of a high number of false negatives raises a problem that is not sufficiently emphasized by a metric such as accuracy. These metrics are computed for each fold, and an average one is used as a final result.

The prediction is done for each slice, then a majority voting scheme is applied to obtain the result per patient, based on the diagnosis made for the slices of its sub-sequences.

Parameter	**Tested values**	**Optimal values**
Slices	0, 1, 2, 3, 4, 5	3
Batch size	6, 8, 12	12
Learning rate	0.1, 0.001, 0.0001, 0.0005	0.0001
L2 regularization	0.1, 0.001, 0.0001	0.0001
Dropout rate	0.4, 0.5, 0.6, 0.7, 0.8	0.5
Momentum rate	0.8, 0.9, 0.99	0.9

Table 5.1: Tested and optimal values of the used parameters.

Chapter 6

Experimental Results

In order to fully understand the results obtained from the proposed architecture and to make a correct comparison with other solutions, it is important to remember the previously exposed characteristics of the dataset. In particular, the use as input of index slices (*i.e. complete slices extracted from the single MRI slices*) and not of ROI (*i.e. an annotated area containing only the lesion*) has a decisive impact both on the performance of the network and on the architectural choices made on it.

Different configurations of the architecture were tested, taking advantage of its multi-branch nature, and thus trying to underline its strengths in the use of multi-task learning to exploit the multi-parametric nature of the MRI dataset.

6.1 Settings

6.1.1 Hardware and Software

The models were run on a Linux Shared-Memory Cluster using a Intel Xeon E5-2680 v3 2.50 GHz with 12 cores as CPU and a nVidia Tesla K40 (12 GB) with 2880 cuda cores as GPU.

The architecture was constructed using Python (V3.7.6) based on Keras (V2.3.1) with TensorFlow (V2.1.0).

6.1.2 Performance Measures and Experimental Method

For the binary classification problem regarding treatment response to NAC therapy, with label 0 assigned to slices from patients who achieved pCR and label 1 assigned to slices from patients who did not achieve pCR (i.e. with RCB score other than 0), the following basic measures were assessed:

- True Positives (TP): number of samples in the no-pCR class that are correctly classified.
- True Negatives (TN): number of samples in the pCR class that are correctly classified.
- False Positives (FP): number of samples in the pCR class that are incorrectly classified.
- False Negatives (FN): number of samples in the no-pCR class that are incorrectly classified.

From these, the following measures considered important for assessing classification performance on this dataset were calculated:

- Accuracy: measures the percentage of correctly classified instances.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

- Sensitivity: also known as *True Positive Rate*, measures the proportion of positives (no-pCR) that are correctly identified.

$$Se = \frac{TP}{TP+FN}$$

- Specificity: also known as *True Negative Rate*, measures the proportion of negatives (pCR) that are correctly identified.

$$Spe = \frac{TN}{TN+FP}$$

The operating point for the calculation of these parameters is left at 0.5

Finally, the main indicator used to assess performance is calculated from the *Receiver Operating Characteristic* (ROC) curve, using the *Area Under the ROC Curve* (AUC).

6.2 Experiments

The model used below as a baseline has the following architectural features: all four branches are included, thus processing the DWI, T2, DCE_peak and DCE_3TP sub-sequences, each of which is extracted from both the pre-NAC and post-NAC studies.

The *slices* parameter is set to 3, i.e. a total of 6 slices plus the index slice for each sequence. As each of the four folds consists of 28 patients for training and 9 for validation, the total number of slices used as input for each of the branches will

be 196 for training (repeated twice, the first time using pre-NAC slices and the second time using post-NAC slices) and 63 for validation.

The epochs are set to 15, batch size is 12, learning rate is equal to 0.0001, as is the L2 regularization, and a dropout value of 0.5 is applied to the concatenated features vector before the last FC layer with 128 neurons. The loss function used is the *binary cross-entropy*, with a weight of 1 assigned to the task specific loss regarding pCR prediction using the features extracted from all the branches and a weight of 0.2 to each loss regarding the pCR prediction task using only one specific sub-sequence. A class weight argument was also used in the training phase, giving a higher weight to the class in the minority in the dataset, i.e. pCR, in order to decrease the number of false positives.

Branch	**Input Slices (Training)**	**Input Slices (Validation)**
DWI	(196, 224, 224, 3) (196, 224, 224, 3)	(63, 224, 224, 3) (63, 224, 224, 3)
T2	(196, 224, 224, 3) (196, 224, 224, 3)	(63, 224, 224, 3) (63, 224, 224, 3)
DCE_peak	(196, 224, 224, 3) (196, 224, 224, 3)	(63, 224, 224, 3) (63, 224, 224, 3)
DCE_3TP	(196, 224, 224, 3) (196, 224, 224, 3)	(63, 224, 224, 3) (63, 224, 224, 3)

Table 6.1: Branches used with their inputs. Each branch will have as input two arrays (one pre-NAC and one post-NAC) composed of 196 or 63 slices (for training or validation) resized to 224x224x3.

The results obtained with this configuration achieve a mean AUC value (calculated on the 4 folds) of 0.90 for the classification of the single slice, and of 0.91 for the classification at patient level.

Sub-sequences	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
DWI, DCE_peak T2, DCE_3TP	0.90	85.3	86.3	82.7

Table 6.2: Mean 4-fold classification, performance for pCR at slice level.

Sub-sequences	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
DWI, DCE_peak T2, DCE_3TP	0.91	86.1	89.3	79.2

Table 6.3: Mean 4-fold classification, performance for pCR at patient level.

It is interesting to note that the performance for the 'final' classification task, using features extracted from all sub-sequences, is always clearly better than that of the single sub-sequence tasks, also and above all due to the greater weight given to its loss in the network training phase, the weights of which will be updated accordingly. It remains however a first suggestion towards the better performance of an ensamble model with respect to the use of single sub-sequences.

Sub-sequence	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
DWI	0.69	55.6	53.3	66.7
T2	0.60	54.0	53.1	60.1
DCE_peak	0.65	62.3	63.2	56.5
DCE_3TP	0.78	71.8	87.6	29.2

Table 6.4: Mean 4-fold classification, performance for pCR at slice level for single sub-sequence tasks.

The full model also records, as one would expect, varying performances on individual folds. Each fold is in fact made up of several patients, which already introduces a certain diversity, and is in addition composed of full slices that are often profoundly different depending on the protocol used and the location of the breast or lesion.

Fold	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
Fold 1	1	96.8	95.2	100
Fold 2	0.96	88.9	92.9	81.0
Fold 3	0.97	81.0	75.5	100
Fold 4	0.69	74.6	81.6	50.0

Table 6.5: Classification performance for pCR at slice level, for single fold.

Figure 6.1: Plot of loss and AUC for each epoch, Fold 1 and Fold 2. The value of the loss functions, both for training and validation, is equal to the weighted sum of the individual loss terms for each task. It can be seen that training and validation loss, in blue and green respectively, decrease towards optimal values as the epochs pass, with the training loss being much faster in this descent which risks leading to overfitting. A trend inversely proportional to the latter occurs instead in the AUC value recorded, with the value recorded for training soon stable on 1 and that of validation more oscillating.

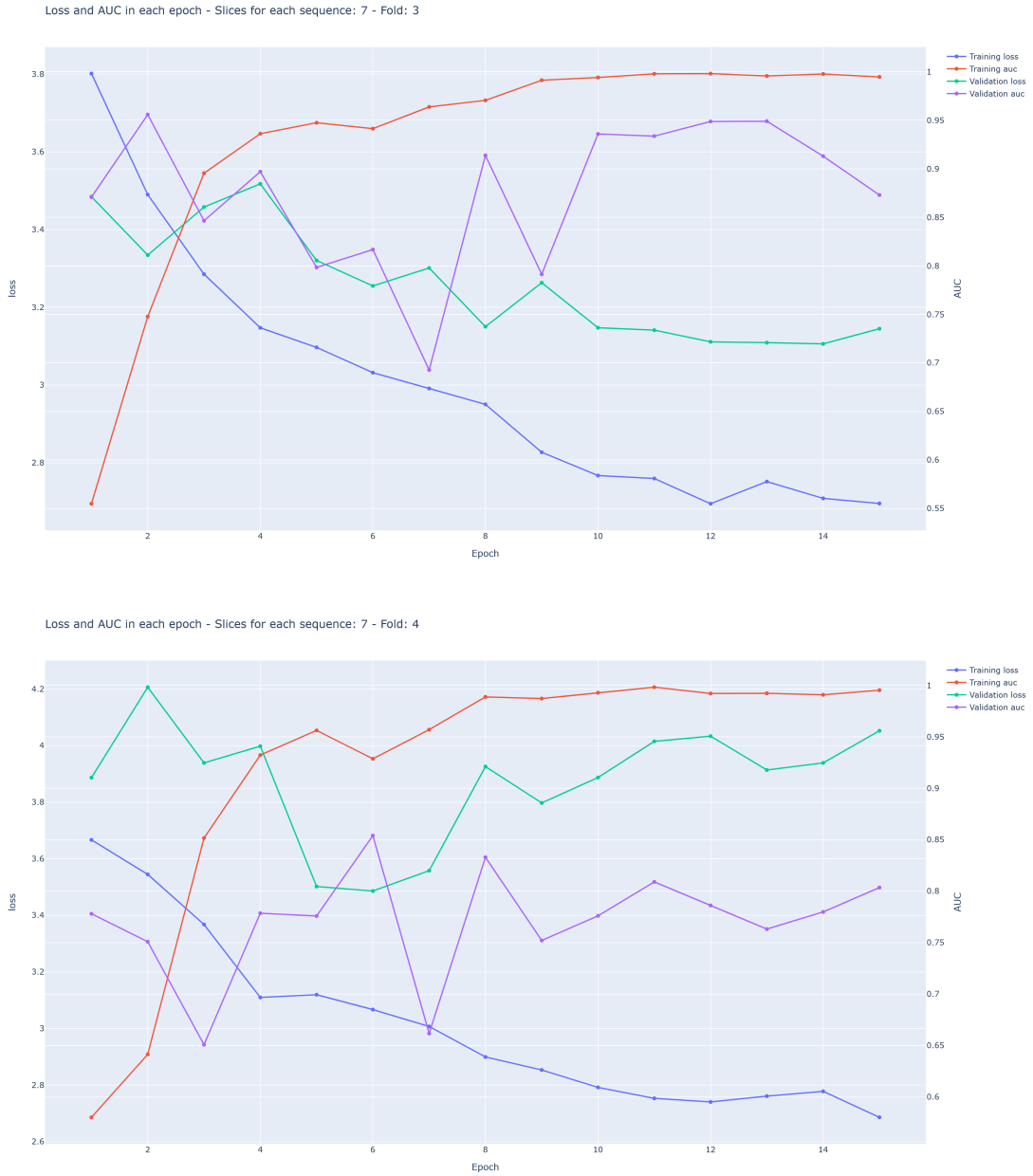


Figure 6.2: Plot of loss and AUC for each epoch, Fold 3 and Fold 4. It can be seen that in Fold 4, which performed worst in the reported experiment, the validation loss decreases in the first few epochs and then increases steadily, thus compromising the AUC value recorded on this fold.

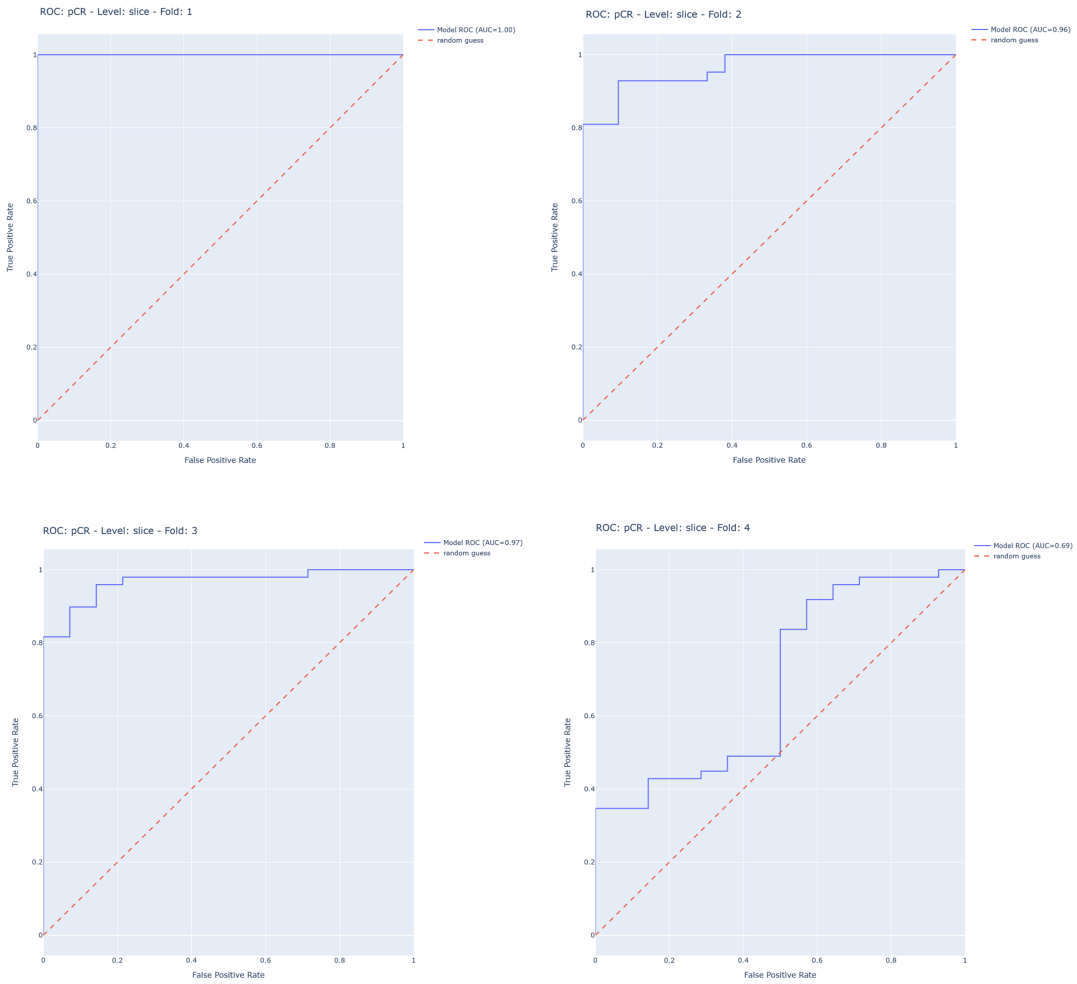


Figure 6.3: ROC plot for pCR at slice level, for each single fold. The architecture's excellent performance is confirmed in 3 out of 4 folds, with AUC values of 1, 0.96, 0.97 and 0.69 respectively.

The graphs summarise the performance of the model using all sub-sequences for each individual fold, showing the loss and AUC trends for training and validation, the ROC curve and the Grad-CAM activation map respectively.

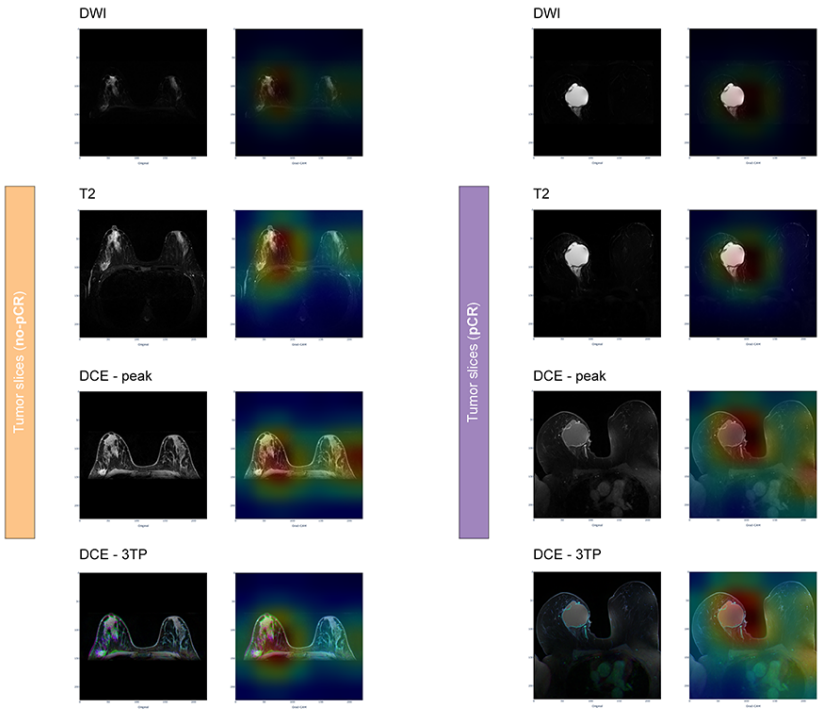


Figure 6.4: Grad-CAM activation map focusing on the results obtained for both a patient who has achieved pCR and a patient who has not. Even though the architecture is based on full slices without any manually drawn bounding boxes, it is able to detect the tumour.

The results obtained, which can be visualised and more easily interpreted using Grad-CAM, show that the architecture is able to focus its attention on the lesion, while using full slices as input without any supporting prior knowledge. It is interesting to note that the area of greatest importance for decision making does not only include the tumour, but also a large portion of the surrounding tissue, suggesting that this may be of some importance in correctly performing treatment response prediction.

This result is in line with both the purely medical literature, which emphasises that the area surrounding the lesion contains important information especially about lymph nodes, and the recent literature using Deep Learning for this task, as in the case of the already mentioned El Adoui et al. [59] study.

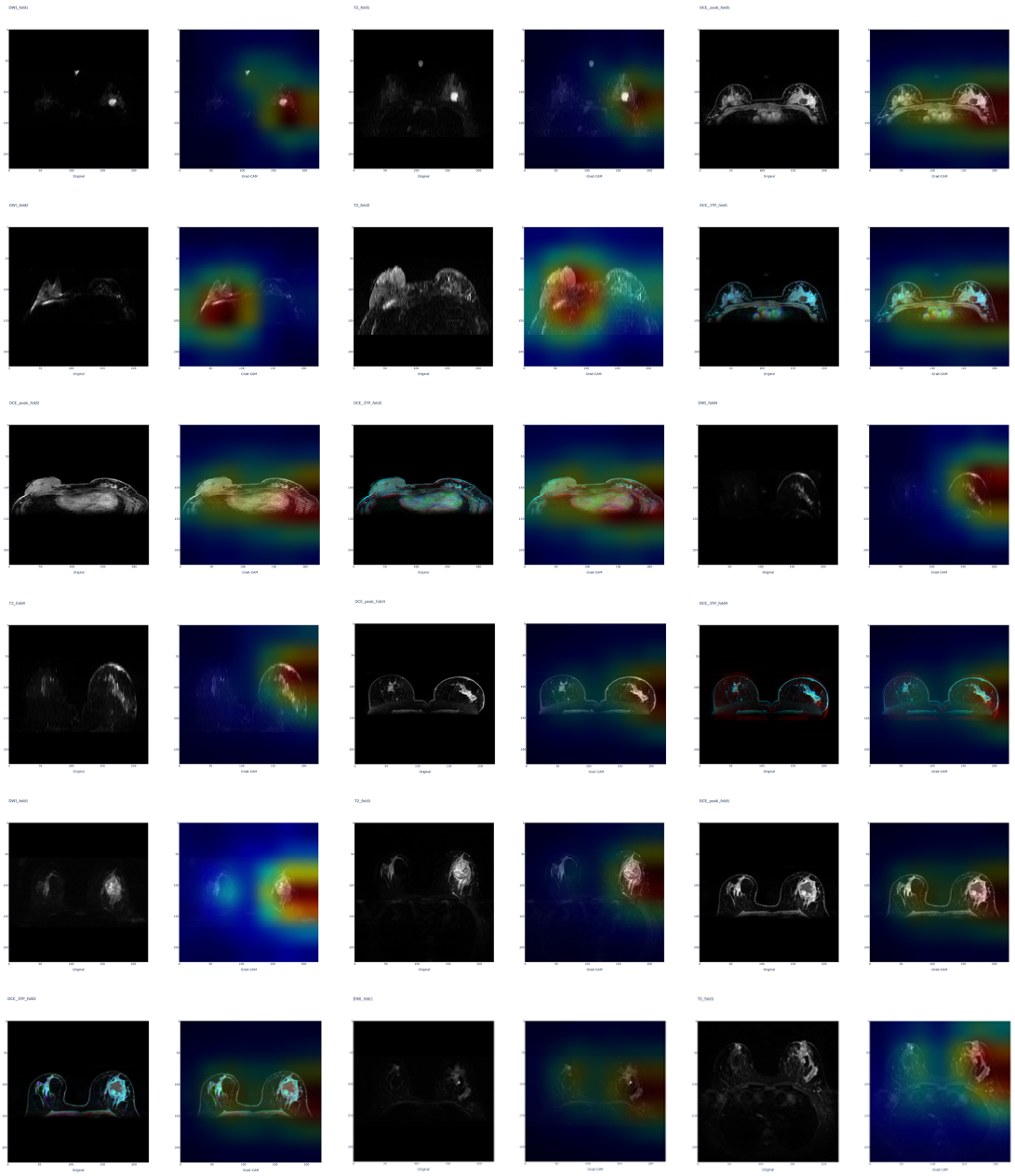


Figure 6.5: Grad-CAM activation map for different patients, highlighting the great heterogeneity of the input full slices and the ability of the network to focus on the lesion and surrounding tissue.

6.2.1 Experiment with Features-Branch

The architecture, with the same parameters, was tested with the addition of the extra branch using hand-crafted features. The vector composed of the 14 extracted features, 7 for the pre-NAC and 7 for the post-NAC studies, used as input for the branch using the MLP to perform a further classification task on the pCR, however, worsened the overall results of the architecture. The reduced number of available features, extracted at patient level and repeated for each respective slice, and the insignificant number of optimisable parameters in the relative branch compared to those of the ResNet used at image level, are probably the cause of the negative influence of this sub-task.

The mean AUC over the four folds is 0.87 at the slice level and 0.88 at the patient level.

Sub-sequences	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
DWI, DCE_peak T2, DCE_3TP, Features-Branch	0.87	75.0	84.7	56.0

Table 6.6: Mean 4-fold classification with features-branch, performance for pCR at slice level.

Sub-sequences	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
DWI, DCE_peak T2, DCE_3TP, Features-Branch	0.88	77.8	85.7	62.5

Table 6.7: Mean 4-fold classification with features-branch, performance for pCR at patient level.

6.2.2 Experiment without Class-Weigh and with Focal Loss

Correctly training the architecture by appropriately addressing the imbalance of data available for the two classes is one of the crucial steps. To this end, an experiment was performed without the "class weight" parameter in order to verify whether its use actually improves performance, and finally an experiment using a loss function that intrinsically deals with this imbalance: the focal loss.

In both cases, performance did not improve with respect to the baseline model described above.

In the case of the experiment without class weight, it can be seen that the specificity drops dramatically, indicating that the model tends to incorrectly classify negative (pCR) slices as positive (no-pCR achieved) as the latter are in the majority for training, thus negatively affecting the correct setting of the weights. This suggests that the class weight parameter does bring benefits.

Sub-sequences	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
DWI, DCE_peak T2, DCE_3TP	0.89	86.5	93.1	41.7

Table 6.8: Mean 4-fold classification without class-weight, performance for pCR at slice level.

The focal loss instead, while useful, is not entirely suitable for this dataset (it was designed for datasets with much greater imbalances of the order of 1:1000).

Sub-sequences	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
DWI, DCE_peak T2, DCE_3TP	0.82	75.0	85.8	69.8

Table 6.9: Mean 4-fold classification with focal loss, performance for pCR at slice level.

6.2.3 Experiment with different Slices

The number of slices to be included as input obviously has a great impact on the capacity of the network. As already reported, the ideal number has been identified as 3, i.e. the inclusion of three slices before and three slices after the index slice. A lower number incorporates too little information, a higher number uses slices where the lesion begins to be too little visible, thus adding noise.

Sub-sequences	**Slices**	**AUC**	**Acc (%)**	**Se (%)**	**Spe (%)**
DWI, DCE_peak T2, DCE_3TP	2	0.88	81.1	81.7	80.0
DWI, DCE_peak T2, DCE_3TP	4	0.87	67.3	60.0	82.4

Table 6.10: Mean 4-fold classification with different slice parameter, performance for pCR at slice level.

6.2.4 Experiments with different Branches

The fundamental concept of the proposed architecture is obviously its use of different branches, creating a multi-input ensemble learning model. Experiments were therefore carried out using only certain branches, in order to verify the variation of performance as they changed.

pre-NAC or post-NAC Only

A first attempt was made to test the network using as input only slices from sub-sequences belonging either to pre-NAC or post-NAC studies, and not both simultaneously as in the baseline model.

Each of the four ResNets will then be used exclusively to extract features from a single input array, instead of two (with weight sharing). The results are in line with what reported in the literature, with both models, pre-NAC only and post-NAC only, having considerably poorer performance than the model using both. In particular, performance was extremely poor in the pre-NAC only model, suggesting that information from full slices belonging to MRI studies prior to the start of neoadjuvant therapy is insufficient to predict treatment response.

Sub-sequences	MRI Study	AUC	Acc (%)	Se (%)	Spe (%)
DWI, DCE_peak T2, DCE_3TP	pre-NAC	0.77	63.5	72.9	47.0
DWI, DCE_peak T2, DCE_3TP	post-NAC	0.86	78.2	89.1	51.8

Table 6.11: Mean 4-fold classification with different MRI studies, performance for pCR at slice level.

Two Branches and Three Branches

The architecture was tested by trying out the various possible pairs of sub-sequences, thus having as tasks the two predictions related to the specific sub-sequences used and the prediction of the pCR carried out by exploiting the features extracted from both branches. It is important to note that the number of slices used as input for the single ResNet is the same as the baseline model, it is only the number of ResNet itself that varies, and therefore the length of the feature vector used to make the final prediction.

Sub-sequences	AUC	Acc (%)	Se (%)	Spe (%)
DWI - DCE peak	0.85	75.4	79.9	68.5
DWI - T2	0.76	65.0	63.5	72.6
DWI - DCE 3TP	0.84	76.2	91.8	39.3
T2 - DCE peak	0.80	74.6	79.8	62.5
T2 - DCE 3TP	0.77	75.8	94.8	26.8
DCE peak - DCE 3TP	0.83	77.4	86.7	56.0

Table 6.12: Mean 4-fold classification using only two sub-sequences, performance for pCR at slice level.

It is interesting to note that in all cases the performance is worse than in the model using all sub-sequences simultaneously. In particular, AUC and specificity are particularly affected by the use of only two branches. This result suggests that features extracted from all four sub-sequences are necessary to perform a correct classification, with the information contained in one of them possibly being key for the patients contained in a given fold.

The use of all the information available in the mpMRI dataset is therefore key to the construction and training of a network capable of effectively generalising on such a small ad heterogeneous dataset composed of full slices.

The same experiment was carried out with only three branches active at the same time:

Sub-sequences	AUC	Acc (%)	Se (%)	Spe (%)
DWI - DCE peak - T2	0.84	74.6	76.0	73.2
DWI - DCE peak - DCE 3TP	0.83	75.8	88.8	45.8
DCE peak - DCE 3TP - T2	0.82	76.9	86.8	51.2
DCE peak - DCE 3TP - DWI	0.84	77.0	91.3	42.9

Table 6.13: Mean 4-fold classification using only three sub-sequences, performance for pCR at slice level.

The addition of a third branch makes the results more consistent in terms of AUC and specificity, but they are still lower than in the model using all four sub-sequences.

Chapter 7

Discussion and Future Developments

The results obtained from the proposed architecture are in line with those of other state-of-the-art studies dealing with treatment response prediction for NAC therapy, and even with some discrepancies due to the different datasets used and different architectural choices seem to confirm several assumptions made in the recent literature.

7.1 Discussion and Comparison

It is useful to recall once again the starting point from which these results were obtained, i.e. the dataset used, which is often one of the obstacles to making an objective comparison in terms of pure numbers with other studies on the same subject. The dataset used does not have a large number of patients and has a large number of different and heterogeneous sequences. However, this latter element embodied in the multi-parametric nature of the MRI studies was used as the major source of information and innovation in the implemented architecture. The substantial difference with the other Deep Learning studies considered is represented by the use of full slices as input, i.e. the whole acquired image of the breast area and not ROIs previously extracted from radiologists or already labelled data. Having to learn from such raw and diverse data, not very large in number, represents the biggest challenge of the model implemented.

In the light of these considerations, the first and perhaps most natural comparison to be made in terms of results is with the study by Tahmassebi et al. [60], carried out on the same dataset but without directly using the images as input, but rather quantitative and qualitative features extracted from them and used for traditional machine learning methods. Using XGBoost as a classifier, the AUC value recorded

was 0.86, using features extracted from all the different sequences available in the dataset, some of which were obtained by comparing the post-NAC study with the pre-NAC study. The architecture proposed in this thesis, with the values considered as baseline (*obtained from the best configuration*) of AUC: 0.90, Accuracy: 81.7, Sensitivity: 81.4, Specificity: 82.7, shows a better performance. This confirms an assumption that is almost taken for granted in the literature, namely that Deep Learning models have the ability to outperform traditional machine learning algorithms. In addition, Thambassemi et al. state that hand-crafted features extracted from all types of sequences (DCE, DWI, T2) were necessary to obtain good results. This is also confirmed by the proposed architecture, but in terms of input images, showing better results using all four identified sub-sequences and thus underlining the potential of an mpMRI dataset.

Network	Dataset (MRI Patients)	MRI Sequences Used	Input Type	Architecture	Performance
Tahmassebi et al. [60]	37	DCE, DWI, T2	Qualitative/quantitative features	XGBoost	AUC: 0.86
Proposed Architecture	37	DCE, DWI, T2	Full slices	Multi-task ensemble learning model	AUC: 0.90

Table 7.1: Comparison of the proposed architecture with the baseline study.

A first study using an end-to-end neural network to perform feature extraction that can be used as a comparison is that of Ravichandran et al. [53]. Although the dataset has a large number of 166 patients, the input consists of simple patches extracted from DCE-MRI index slices. The only study that is considered at all is the pre-NAC one, obtaining with a fairly simple architecture an AUC value: 0.77. The architecture proposed in this thesis obtains the same AUC using only pre-NAC studies, a value that increases considerably in the case of post-NAC studies and further when both are used simultaneously.

Network	Dataset (MRI Patients)	MRI Sequences Used	Input Type	Architecture	Performance
Ravichandran et al. [53]	166	DCE (pre-treatment only)	Patches from index slice	CNN with 6 blocks	AUC: 0.77
Proposed Architecture (pre-NAC)	37	DCE, DWI, T2 (pre-treatment only)	Full slices	Multi-task ensemble learning model	AUC: 0.77
Proposed Architecture	37	DCE, DWI, T2	Full slices	Multi-task ensemble learning model	AUC: 0.90

Table 7.2: Comparison of the proposed architecture with the pre-NAC only study.

The importance of both studies is widely shared in the most recent literature although considerable results were also obtained only from images taken before the start of therapy, as in the case of Ha et al. [52]. In the latter case the accuracy achieved was 88% on a dataset of 141 patients from which 3107 volumetric slices of the lesion were extracted by an experienced radiologist and given as input to a VGG16-like architecture. The architecture developed in this thesis, in order to be able to stand comparison with such a dataset, must use the results coming from the experiments conducted to the maximum of its potential, that is with both pre- and post-NAC studies, obtaining a similar accuracy of 85.3%, but being able to count on less data and above all not previously processed by radiologists except for

the identification of the index slice.

A more direct comparison can be made with the studies of El Adoui et al. [54], [55], [59]. These studies in fact introduce the idea of a multi-branch architecture capable of considering and processing both pre-NAC and post-NAC studies. They also use a dataset comparable in size (42 patients) to the one used in this thesis (37 patients), but working with ROIs extracted around the lesion rather than a full slice.

In [54] an architecture using two distinct VGG-like blocks independently processes ROIs extracted and recorded from the two available studies, concatenating the extracted features before making the prediction. The AUC value recorded in the best performing model is 0.96 (obtained using 25% of the data as validation set, not specified if in k-fold cross-validation). In [55] the branches of the previous architecture are tripled, going to process separately ROIs extracted from axial, transversal and coronal slices, thus focusing on features extracted from all spatial dimensions of the DCE-MRI dataset. In this case the best recorded AUC is 0.92.

More interesting is the comparison with [59], where the same dataset of 42 patients used previously is expanded with 14 new studies used for validation. Again, the volume of interest containing the lesion is first extracted from axial DCE-MRI slices and used as input for an architecture that has two branches to independently analyse pre- and post-NAC before concatenation. The best result was recorded when the input, in addition to the lesion, also contained the immediate surrounding region, with a final AUC of 0.92, sensitivity of 92.2 and specificity of 79.1. Also in this study Grad-CAM is used, showing that for a correct prediction the network considers important not only the lesion but also the surrounding area. The architecture developed in this thesis is very similar in performance and, although based on full slices, it seems to confirm that other areas of surrounding tissue in addition to the lesion itself are indeed important according to Grad-CAM.

Network	Dataset (MRI Patients)	MRI Sequences Used	Input Type	Architecture	Performance
El Adoui et al. [54]	42	DCE	ROI crops	VGGNet-like	AUC: 0.96
El Adoui et al. [55]	42	DCE	ROI crops	VGGNet-like	AUC: 0.92
El Adoui et al. [59]	42 + 14 external for validation	DCE, DWI	VOI crops	Custom Multi-input CNN	AUC: 0.91
Proposed Architecture	37	DCE, DWI, T2	**Full slices**	Multi-task ensemble learning model	AUC: 0.90

Table 7.3: Characteristics and classification performance with other studies.

Network	Input Type	AUC	Accuracy	Sensitivity	Specificity
El Adoui et al. [59]	VOI crops	0.91	88	92.2	79.1
Proposed Architecture	**Full slices**	0.90	81.7	81.4	82.7

Table 7.4: Characteristics and classification performance with El Adoui et al. [59].

The impressive results obtained by Qu et al. [58] deserve a special mention: on a dataset using manually segmented lesions extracted from 244 patients for

training and 58 for validation, they obtained results of AUC: 0.97, sensitivity: 96, specificity: 100. The study proposes a multi-branch architecture that uses as input six lesions extracted from different time instants from the pre-NAC study and another six from the post-NAC study, before concatenating the extracted features for the final prediction. In addition to the values recorded, the study also states that the combined use of both studies for NAC treatment prediction is important, recording much lower values using only pre-NAC slices and good but not great results with post-NAC slices. This result also occurs in the proposed architecture.

In conclusion, no other studies in the field of treatment response prediction to NAC therapy have been found to date that make intensive use of the various information contained in different sequences of a multi-parameter MRI dataset, and in general the use of full slices is normally avoided in the literature, preferring an input with less noise but obtainable only through previous processing or through the work of a radiologist.

In spite of this, the performances recorded are in line with those reported in the other studies considered, behaving according to the most accredited hypotheses in this field and above all demonstrating the potential of an architecture that can still be developed in the future.

7.2 Limitations of the Study

The most obvious limitation to the results of the proposed architecture is due to the small number of patients available in the dataset. In addition to the usual difficulties in the field of medical imaging to obtain organised and structured data, participation in a study aiming at detecting treatment response requires an additional MRI study (the one after two cycles of NAC) to which the patient willing to participate must agree to undergo. Testing the architecture on a larger and more diverse dataset can provide a much more effective measure of the real capabilities of the network.

Furthermore, although the study minimised inter-observer and intra-observer variability by using human input only for the simple task of identifying the index slice, it uses the entire extracted slice as input. This, in addition to the various difficulties explained above, makes the imaging protocol used much more incisive (compared to using a ROI of the lesion as input). It would therefore be useful to test the architecture using mpMRI studies applying other acquisition protocols, possibly acquired in other centres, thus overcoming the retrospective and single centre nature of the study.

A prospective and multi-center study may help in the construction of a model capable of effective generalisation in predicting treatment response for NAC therapy,

adapting to different clinical situations.

7.3 Future Developments

The potential shown by this architecture, despite the non-optimal conditions of the dataset, demonstrates the possibility of further developments related to the proposed solutions. If, on the one hand, it would be desirable to resolve the limitations highlighted above, on the other hand, even with the current conditions it could be possible to develop new ideas linked to the progress of Deep Learning techniques.

Sophisticated and increasingly used techniques such as generative adversarial networks could be used to produce synthetic inputs and thus artificially increase the number of data available for training, with the added possibility of balancing the minority class (patients achieving pCR). But other more immediate and less performance-intensive approaches can also be explored, exploiting the multi-task nature of the network. The features already extracted in the baseline study could be used to add further tasks related to them, asupicably increasing the generalisation capabilities of the architecture, with a trend in the literature that sees excellent results when Deep Learning techniques are combined with a priori knowledge extracted by qualified personnel.

In addition, the most important "building blocks" of the network consisting of the ResNet50s pre-trained on ImageNet could be made more efficient by improving their ability to process medical images, using for example the Deep Colorization approach proposed by Morra, Piano et al. [74] that by transforming input slices into RGB helps the transfered learning component of the architecture to process data more compliant with those on which it was trained.

Chapter 8

Conclusions

The use of computer-aided systems in the medical field is an increasingly concrete and not negligible reality, and the pervasiveness of data-driven applications in every sector, whether academic, industrial or simply everyday, suggests that these tools will be even more decisive in the future. The parallel development, aided by adequate technological support, of techniques and approaches capable of exploiting this data, offers more and more solutions to be explored: the use of Deep Learning architectures to analyse breast cancer MRIs is certainly one of them, as highlighted by the important role it plays in the most recent literature.

Among the possible uses in the development of a detection and classification system, ranging from segmentation to distinguishing a lesion between malignant and benign, treatment response prediction is particularly important. Having a rapid response on the efficacy of neoadjuvant chemotherapy can prevent the patient from unnecessarily undergoing the wrong treatment, allowing more effective ones to be chosen. The architecture developed in this thesis aims to accomplish this task, being designed after a careful analysis of other state-of-the-art models used for this purpose and trying to solve the characteristic problems of this specific field using the most advanced Deep Learning techniques.

The multi-task ensemble learning model created succeeds in making the most of the information contained in a multi-parameter MRI dataset, yielding promising results in line with the most accredited studies in this field. To the best of our knowledge, the proposed architecture is the first to separately analyse different MRI acquisition modes, pre-NAC and post-NAC, starting from an input composed of full slices. These results, which can still be further developed, are achieved by limiting the need for prior knowledge and intervention on the dataset, trying to provide an approach capable of generalising as much as possible in a field which unfortunately still lacks a certain degree of standardization with regard to the data used, and are intended as a further aid in carrying out such a critical clinical task.

Bibliography

[1] Douglas Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rep. META Group, Feb. 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (cit. on p. 1).

[2] Caleb Garling. *Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines*. May 2015. URL: https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/ (cit. on p. 2).

[3] Debashis Ganguly, Srabonti Chakraborty, Maricel Balitanas, and Tai-hoon Kim. «Medical Imaging: A Review». In: vol. 78. Sept. 2010, pp. 504–516. ISBN: 978-3-642-16443-9. DOI: 10.1007/978-3-642-16444-6_63 (cit. on p. 4).

[4] Sabyasachi Dash, Sushil Shakyawar, Mohit Sharma, and Sandeep Kaushik. «Big data in healthcare: management, analysis and future prospects». In: *Journal of Big Data* 6 (June 2019). DOI: 10.1186/s40537-019-0217-0 (cit. on p. 4).

[5] Kunio Doi. «Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential». In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 31 (June 2007), pp. 198–211. DOI: 10.1016/j.compmedimag.2007.02.002 (cit. on p. 5).

[6] Michel Coleman et al. «Cancer survival in five continents: a worldwide population-based study (CONCORD)». In: *The lancet oncology* 9 (Aug. 2008), pp. 730–56. DOI: 10.1016/S1470-2045(08)70179-7 (cit. on p. 5).

[7] Anne F. Schott and Daniel F. Hayes. «Defining the Benefits of Neoadjuvant Chemotherapy for Breast Cancer». In: *Journal of Clinical Oncology* 30.15 (2012). PMID: 22508810, pp. 1747–1749. DOI: 10.1200/JCO.2011.41.3161. eprint: https://doi.org/10.1200/JCO.2011.41.3161. URL: https://doi.org/10.1200/JCO.2011.41.3161 (cit. on p. 5).

[8] Shahla Masood. «Neoadjuvant chemotherapy in breast cancers». In: *Women's Health* 12 (Sept. 2016), pp. 480–491. DOI: 10.1177/1745505716677139 (cit. on p. 5).

[9] Abi Berger. «Magnetic resonance imaging». In: *BMJ* 324.7328 (2002), p. 35. ISSN: 0959-8138. DOI: 10.1136/bmj.324.7328.35. eprint: https://www.bmj.com/content/324/7328/35.full.pdf. URL: https://www.bmj.com/content/324/7328/35 (cit. on p. 6).

[10] «Automation and anxiety». In: *The Economist* (). ISSN: 0013-0613. URL: https://www.economist.com/special-report/2016/06/23/automation-and-anxiety (visited on 04/03/2020) (cit. on p. 6).

[11] Emanuele Neri, Nandita de Souza, Adrian Brady, Angel Alberich Bayarri, Christoph D. Becker, Francesca Coppola, Jacob Visser, and European Society of Radiology (ESR). «What the radiologist should know about artificial intelligence – an ESR white paper». In: *Insights into Imaging* 10.1 (Apr. 2019), p. 44. ISSN: 1869-4101. DOI: 10.1186/s13244-019-0738-2. URL: https://doi.org/10.1186/s13244-019-0738-2 (visited on 04/03/2020) (cit. on p. 7).

[12] Thomas H. Davenport and D. O. Keith J. Dreyer. «AI Will Change Radiology, but It Won't Replace Radiologists». In: *Harvard Business Review* (Mar. 2018). ISSN: 0017-8012. URL: https://hbr.org/2018/03/ai-will-change-radiology-but-it-wont-replace-radiologists (visited on 04/03/2020) (cit. on p. 7).

[13] Ahmet Yurttakal, Hasan Erbay, Turkan Ikizceli, Seyhan Karacavus, and Gökalp Çınarer. «ARTICLE A COMPARATIVE STUDY ON SEGMENTATION AND CLASSIFICATION IN BREAST MRI IMAGING». In: *IIOAB Journal* 9 (Dec. 2018), pp. 23–33 (cit. on p. 8).

[14] Robert Gillies, Paul Kinahan, and Hedvig Hricak. «Radiomics: Images Are More than Pictures, They Are Data». In: *Radiology* 278 (Nov. 2015), p. 151169. DOI: 10.1148/radiol.2015151169 (cit. on p. 9).

[15] Elizabeth Cain, Ashirbani Saha, Michael Harowicz, Jeffrey Marks, Paul Marcom, and Maciej Mazurowski. «Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set». In: *Breast Cancer Research and Treatment* 173 (Oct. 2018). DOI: 10.1007/s10549-018-4990-9 (cit. on p. 9).

[16] Amirhessam Tahmassebi et al. «Impact of Machine Learning With Multiparametric Magnetic Resonance Imaging of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients». In: *Investigative Radiology* 54 (Oct. 2018), p. 1. DOI: 10.1097/RLI.0000000000000518 (cit. on p. 9).

[17] Amirhessam Tahmassebi, Anahid Ehtemami, Behshad Mohebali, Amir Gandomi, Katja Pinker, and Anke Meyer-Base. «Big data analytics in medical imaging using deep learning». In: May 2019, p. 13. DOI: 10.1117/12.2516014 (cit. on p. 9).

[18] Roberto Lo Gullo, Sarah Eskreis-Winkler, Elizabeth A. Morris, and Katja Pinker. «Machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy». In: *The Breast* 49 (2020), pp. 115–122. ISSN: 0960-9776. DOI: https://doi.org/10.1016/j.breast.2019.11.009. URL: http://www.sciencedirect.com/science/article/pii/S0960977619311014 (cit. on p. 9).

[19] Taye Girma Debelee, Friedhelm Schwenker, Achim Ibenthal, and Dereje W. Yohannes. «Survey of deep learning in breast cancer image analysis». In: 2019 (cit. on p. 9).

[20] Daniel Truhn, Simone Schrading, Christoph Haarburger, Hannah Schneider, Dorit Merhof, and Christiane Kuhl. «Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI». In: *Radiology* 290.2 (2019). PMID: 30422086, pp. 290–297. DOI: 10.1148/radiol.2018181352. eprint: https://doi.org/10.1148/radiol.2018181352. URL: https://doi.org/10.1148/radiol.2018181352 (cit. on p. 9).

[21] Azam Hamidinekoo, Erika Denton, Andrik Rampun, Kate Honnor, and Reyer Zwiggelaar. «Deep Learning in Mammography and Breast Histology, an Overview and Future Trends». In: *Medical Image Analysis* 47 (Mar. 2018). DOI: 10.1016/j.media.2018.03.006 (cit. on p. 9).

[22] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger. «Comparison of Breast MRI Tumor Classification Using Human-Engineered Radiomics, Transfer Learning From Deep Convolutional Neural Networks, and Fusion Methods». In: *Proceedings of the IEEE* 108.1 (Jan. 2020), pp. 163–177. ISSN: 1558-2256 (cit. on p. 9).

[23] Renée Granzier, Thiemo Nijnatten, Henry Woodruff, Marjolein Smidt, and Marc Lobbes. «Exploring Breast Cancer Response Prediction to Neoadjuvant Systemic Therapy using MRI-based Radiomics: A Systematic Review». In: *European Journal of Radiology* 121 (Nov. 2019), p. 108736. DOI: 10.1016/j.ejrad.2019.108736 (cit. on p. 10).

[24] Ahmet Yurttakal, Hasan Erbay, Turkan Ikizceli, and Seyhan Karacavus. «Detection of breast cancer via deep convolution neural networks using MRI images». In: *Multimedia Tools and Applications* (Apr. 2019). DOI: 10.1007/s11042-019-7479-6 (cit. on p. 10).

[25] Anna Vignati et al. «Performance of a Fully Automatic Lesion Detection System for Breast DCE-MRI». In: *Journal of magnetic resonance imaging : JMRI* 34 (Dec. 2011), pp. 1341–51. DOI: 10.1002/jmri.22680 (cit. on p. 11).

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-Net: Convolutional Networks for Biomedical Image Segmentation». In: (May 2015) (cit. on p. 11).

[27] Gabriele Piantadosi, Mario Sansone, and Carlo Sansone. «Breast Segmentation in MRI via U-Net Deep Convolutional Neural Networks». In: Aug. 2018, pp. 3917–3922. DOI: 10.1109/ICPR.2018.8545327 (cit. on p. 11).

[28] Gabriele Piantadosi, Mario Sansone, Roberta Fusco, and Carlo Sansone. «Multi-planar 3D breast segmentation in MRI via deep convolutional neural networks». In: *Artificial Intelligence in Medicine* 103 (2020), p. 101781. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2019.101781. URL: http://www.sciencedirect.com/science/article/pii/S0933365718306985 (cit. on p. 11).

[29] Xiaowei Xu, Ling Fu, Yizhi Chen, Rasmus Larsson, Dandan Zhang, Shiteng Suo, Jia Hua, and Jun Zhao. «Breast Region Segmentation being Convolutional Neural Network in Dynamic Contrast Enhanced MRI». In: vol. 2018. July 2018, pp. 750–753. DOI: 10.1109/EMBC.2018.8512422 (cit. on p. 11).

[30] Xinpeng Zheng, zhuangsheng Liu, Lin Chang, Wansheng Long, and Yao Lu. «Coordinate-guided U-Net for automated breast segmentation on MRI images». In: May 2019, p. 84. DOI: 10.1117/12.2524250 (cit. on p. 11).

[31] Homa Fashandi, Gregory Kuling, YingLi Lu, Hongbo Wu, and Anne Martel. «An investigation of the effect of fat suppression and dimensionality on the accuracy of breast MRI segmentation using U-nets». In: *Medical Physics* 46 (Jan. 2019). DOI: 10.1002/mp.13375 (cit. on p. 12).

[32] Antonio Galli, Michela Gravina, Stefano Marrone, Gabriele Piantadosi, Mario Sansone, and Carlo Sansone. «Evaluating Impacts of Motion Correction on Deep Learning Approaches for Breast DCE-MRI Segmentation and Classification». In: Aug. 2019, pp. 294–304. ISBN: 978-3-030-29890-6. DOI: 10.1007/978-3-030-29891-3_26 (cit. on p. 12).

[33] Shuyue Guan and Murray Loew. «Using generative adversarial networks and transfer learning for breast cancer detection by convolutional neural networks». In: *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*. Ed. by Po-Hao Chen and Peter R. Bak. Vol. 10954. International Society for Optics and Photonics. SPIE, 2019, pp. 306–318. DOI: 10.1117/12.2512671. URL: https://doi.org/10.1117/12.2512671 (cit. on pp. 12, 13).

[34] Hoo-Chang Shin, Neil Tenenholtz, Jameson Rogers, Christopher Schwarz, Matthew Senjem, Jeffrey Gunter, Katherine Andriole, and Mark Michalski. «Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings». In: Sept. 2018, pp. 1–11. ISBN: 978-3-030-00535-1. DOI: 10.1007/978-3-030-00536-8_1 (cit. on p. 13).

[35] Lei Zhang, Zhimeng Luo, Ruimei Chai, Dooman Arefan, Jules Sumkin, and Shandong Wu. «Deep-learning method for tumor segmentation in breast DCE-MRI». In: Mar. 2019, p. 14. DOI: 10.1117/12.2513090 (cit. on pp. 13, 14).

[36] Mehmet Dalmış, Suzan Vreemann, Thijs Kooi, Ritse Mann, Nico Karssemeijer, and Albert Gubern-Mérida. «Fully automated detection of breast cancer in screening MRI using convolutional neural networks». In: *Journal of Medical Imaging* 5 (Jan. 2018), p. 1. DOI: 10.1117/1.JMI.5.1.014502 (cit. on p. 13).

[37] Wenhuan Lu, Zhe Wang, Yuqing He, Hong Yu, Naixue Xiong, and Jianguo Wei. «Breast Cancer Detection Based on Merging Four Modes MRI Using Convolutional Neural Networks». In: May 2019, pp. 1035–1039. DOI: 10.1109/ICASSP.2019.8683149 (cit. on p. 14).

[38] Gabriele Piantadosi, Stefano Marrone, Antonio Galli, Mario Sansone, and Carlo Sansone. «DCE-MRI Breast Lesions Segmentation with a 3TP U-Net Deep Convolutional Neural Network». In: June 2019, pp. 628–633. DOI: 10.1109/CBMS.2019.00130 (cit. on p. 14).

[39] Mingjian Chen, Hao Zheng, Changsheng Lu, Enmei Tu, Jie Yang, and Nikola Kasabov. «A Spatio-Temporal Fully Convolutional Network for Breast Lesion Segmentation in DCE-MRI: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part VII». In: Jan. 2018, pp. 358–368. ISBN: 978-3-030-04238-7. DOI: 10.1007/978-3-030-04239-4_32 (cit. on p. 14).

[40] Gabriel Maicas, Gustavo Carneiro, Andrew Bradley, Jacinto Nascimento, and Ian Reid. «Deep Reinforcement Learning for Active Breast Lesion Detection from DCE-MRI». In: Sept. 2017, pp. 665–673. ISBN: 978-3-319-66178-0. DOI: 10.1007/978-3-319-66179-7_76 (cit. on p. 15).

[41] Natalia Antropova, B Huynh, and Maryellen Giger. «SU-D-207B-06: Predicting Breast Cancer Malignancy On DCE-MRI Data Using Pre-Trained Convolutional Neural Networks». In: *Medical Physics* 43 (June 2016), pp. 3349–3350. DOI: 10.1118/1.4955674 (cit. on p. 15).

[42] Reza Rasti, Mohammad Teshnehlab, and Son Phung. «Breast Cancer Diagnosis in DCE-MRI using Mixture Ensemble of Convolutional Neural Networks». In: *Pattern Recognition* 72 (Aug. 2017). DOI: 10.1016/j.patcog.2017.08.004 (cit. on p. 15).

[43] Natalia Antropova, Benjamin Q. Huynh, and Maryellen L. Giger. «A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets». In: *Medical Physics* 44 (2017), pp. 5162–5171 (cit. on pp. 15, 16).

[44] Natalia Antropova, Hiroyuki Abe, and Maryellen Giger. «Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks». In: *Journal of Medical Imaging* 5 (Feb. 2018), p. 1. DOI: 10.1117/1.JMI.5.1.014503 (cit. on p. 16).

[45] Qiyuan Hu, Heather Whitney, and Maryellen Giger. «Transfer Learning in 4D for Breast Cancer Diagnosis using Dynamic Contrast-Enhanced Magnetic Resonance Imaging». In: (Nov. 2019) (cit. on p. 16).

[46] Natalia Antropova, Benjamin Huynh, Hui Li, and Maryellen Giger. «Breast lesion classification based on dynamic contrast-enhanced magnetic resonance images sequences with long short-term memory networks». In: *Journal of Medical Imaging* 6 (Aug. 2018). DOI: 10.1117/1.JMI.6.1.011002 (cit. on p. 16).

[47] Michela Gravina, Stefano Marrone, Gabriele Piantadosi, Mario Sansone, and Carlo Sansone. «3TP-CNN: Radiomics and Deep Learning for Lesions Classification in DCE-MRI». In: Sept. 2019, pp. 661–671. ISBN: 978-3-030-30644-1. DOI: 10.1007/978-3-030-30645-8_60 (cit. on p. 16).

[48] Jiejie Zhou et al. «Diagnosis of Benign and Malignant Breast Lesions on DCE-MRI by Using Radiomics and Deep Learning With Consideration of Peritumor Tissue». In: *Journal of Magnetic Resonance Imaging* 51 (Nov. 2019). DOI: 10.1002/jmri.26981 (cit. on p. 17).

[49] Hongwei Feng, Jiaqi Cao, Hongyu Wang, Yilin Xie, Di Yang, Jun Feng, and Baoying Chen. «A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence MRI». In: *Magnetic Resonance Imaging* (Mar. 2020). DOI: 10.1016/j.mri.2020.03.001 (cit. on p. 17).

[50] Luyang Luo, Hao Chen, Xi Wang, Qi Dou, Huangjin Lin, Juan Zhou, Gongjie Li, and Pheng-Ann Heng. «Deep Angular Embedding and Feature Correlation Attention for Breast MRI Cancer Analysis». In: (June 2019) (cit. on p. 17).

[51] Benjamin Huynh, Natasha Antropova, and Maryellen Giger. «Comparison of breast DCE-MRI contrast time points for predicting response to neoadjuvant chemotherapy using deep convolutional neural network features with transfer learning». In: Mar. 2017, 101340U. DOI: 10.1117/12.2255316 (cit. on pp. 18, 20).

[52] Richard Ha et al. «Prior to Initiation of Chemotherapy, Can We Predict Breast Tumor Response? Deep Learning Convolutional Neural Networks Approach Using a Breast MRI Tumor Dataset». In: *Journal of Digital Imaging* 32 (Oct. 2018). DOI: 10.1007/s10278-018-0144-1 (cit. on pp. 18, 20, 53).

[53] Kavya Ravichandran, Nathaniel Braman, Andrew Janowczyk, and Anant Madabhushi. «A deep learning classifier for prediction of pathological complete response to neoadjuvant chemotherapy from baseline breast DCE-MRI». In: Feb. 2018, p. 11. DOI: 10.1117/12.2294056 (cit. on pp. 18, 20, 53).

[54] Mohammed El Adoui, Amine Larhmam, Stylianos Drisis, and Mohammed Benjelloun. «Deep Learning approach predicting breast tumor response to neoadjuvant treatment using DCE-MRI volumes acquired before and after chemotherapy». In: Mar. 2019, p. 90. DOI: 10.1117/12.2505887 (cit. on pp. 18, 20, 54).

[55] Mohammed El Adoui, Stylianos Drisis, and Mohammed Benjelloun. «Predict Breast Tumor Response to Chemotherapy Using a 3D Deep Learning Architecture Applied to DCE-MRI Data». In: Apr. 2019, pp. 33–40. ISBN: 978-3-030-17934-2. DOI: 10.1007/978-3-030-17935-9_4 (cit. on pp. 18, 20, 54).

[56] Michał Byra, Katarzyna Dobruch-Sobczak, Ziemowit Klimonda, Hanna Piotrzkowska-Wroblewska, and Jerzy Litniewski. «Early Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer Sonography Using Siamese Convolutional Neural Networks». In: *IEEE Journal of Biomedical and Health Informatics* PP (July 2020), pp. 1–1. DOI: 10.1109/JBHI.2020.3008040 (cit. on p. 19).

[57] Joon Ho Choi et al. «Early prediction of neoadjuvant chemotherapy response for advanced breast cancer using PET/MRI image deep learning». In: *Scientific Reports* 10 (Dec. 2020). DOI: 10.1038/s41598-020-77875-5 (cit. on p. 19).

[58] Yu-Hong Qu, Hai-Tao Zhu, Kun Cao, Xiao-Ting Li, Meng Ye, and Ying-Shi Sun. «Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using a deep learning (DL) method». In: *Thoracic Cancer* 11.3 (2020), pp. 651–658. DOI: 10.1111/1759-7714.13309. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1759-7714.13309. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1759-7714.13309 (cit. on pp. 19, 20, 54).

[59] Mohammed El Adoui, Stylianos Drisis, and Mohammed Benjelloun. «Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images». In: *International Journal of Computer Assisted Radiology and Surgery* 15 (June 2020). DOI: 10.1007/s11548-020-02209-9 (cit. on pp. 19, 20, 46, 54).

[60] Amirhessam Tahmassebi et al. «Impact of Machine Learning With Multiparametric Magnetic Resonance Imaging of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients». In: *Investigative Radiology* 54 (Oct. 2018), p. 1. DOI: 10.1097/RLI.0000000000000518 (cit. on pp. 20, 24, 25, 31, 52, 53).

[61] Christina Yau et al. «Abstract GS5-01: Residual cancer burden after neoadjuvant therapy and long-term survival outcomes in breast cancer: A multicenter pooled analysis». In: *Cancer Research* 80.4 Supplement (2020), GS5-01–GS5-01. ISSN: 0008-5472. DOI: 10.1158/1538-7445.SABCS19-GS5-01. eprint: https://cancerres.aacrjournals.org/content. URL: https://cancerres.aacrjournals.org/content/80/4_Supplement/GS5-01 (cit. on p. 21).

[62] W. Fraser Symmans et al. «Measurement of Residual Breast Cancer Burden to Predict Survival After Neoadjuvant Chemotherapy». In: *Journal of Clinical Oncology* 25.28 (2007). PMID: 17785706, pp. 4414–4422. DOI: 10.1200/JCO.2007.10.6823. eprint: https://doi.org/10.1200/JCO.2007.10.6823. URL: https://doi.org/10.1200/JCO.2007.10.6823 (cit. on p. 21).

[63] Yuka Asano et al. «Prediction of survival after neoadjuvant chemotherapy for breast cancer by evaluation of tumor-infiltrating lymphocytes and residual cancer burden». In: *BMC Cancer* 17 (Dec. 2017). DOI: 10.1186/s12885-017-3927-8 (cit. on p. 21).

[64] Katja Pinker et al. «A Combined High Temporal and High Spatial Resolution 3 Tesla MR Imaging Protocol for the Assessment of Breast Lesions». In: *Investigative radiology* 44 (Aug. 2009), pp. 553–8. DOI: 10.1097/RLI.0b013e3181b4c127 (cit. on p. 22).

[65] Lindsay Turnbull. «Dynamic contrast-enhanced MRI in the diagnosis and management of breast cancer». In: *NMR in biomedicine* 22 (Jan. 2009), pp. 28–39. DOI: 10.1002/nbm.1273 (cit. on p. 23).

[66] Claudio Spick et al. «Diffusion-weighted MRI of breast lesions: A prospective clinical investigation of the quantitative imaging biomarker characteristics of reproducibility, repeatability, and diagnostic accuracy». In: *NMR in Biomedicine* 29 (Aug. 2016). DOI: 10.1002/nbm.3596 (cit. on p. 23).

[67] Christine Westra, Vandana Dialani, Tejas Mehta, and Ronald Eisenberg. «Using T2-Weighted Sequences to More Accurately Characterize Breast Masses Seen on MRI». In: *AJR. American journal of roentgenology* 202 (Mar. 2014), W183–90. DOI: 10.2214/AJR.13.11266 (cit. on p. 24).

[68] Tianqi Chen and Carlos Guestrin. «XGBoost: A Scalable Tree Boosting System». In: Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785 (cit. on p. 25).

[69] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. «ImageNet Classification with Deep Convolutional Neural Networks». In: *Neural Information Processing Systems* 25 (Jan. 2012). DOI: 10.1145/3065386 (cit. on p. 28).

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. «Deep Residual Learning for Image Recognition». In: June 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on p. 28).

[71] Sebastian Ruder. «An Overview of Multi-Task Learning in Deep Neural Networks». In: (June 2017) (cit. on p. 31).

[72] Ramprasaath Rs, Michael Cogswell, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. «Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization». In: Oct. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74 (cit. on p. 33).

[73] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. «Focal Loss for Dense Object Detection». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (July 2018), pp. 1–1. DOI: 10.1109/TPAMI.2018.2858826 (cit. on p. 37).

[74] Lia Morra, Luca Piano, F. Lamberti, and Tatiana Tommasi. «Bridging the gap between Natural and Medical Images through Deep Colorization». In: (May 2020) (cit. on p. 56).