UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Physics

Final Dissertation

MR-radiomic-based pathological response prediction to neoadjuvant chemotherapy in breast cancer

Thesis supervisor

Prof./Dr. Laura De Nardo

Thesis co-supervisor

Dr. Luisa Altabella

Candidate

Marina Fedon Vocaturo

Academic Year 2023/2024

Contents

List of Figures

List of Tables

List of Abbreviations

ACR	American College of Radiology
AI	Artificial Intelligence
AUC	Area Under the Curve
BCE	Binary Cross Entropy
BI-RADS	Breast Imaging Reporting and Data System
CI	Confidence Interval
CL	Confidence Level
CNN	Convolutional Neural Network
CT	Computed Tomography
CV	Cross Validation
DCE-MRI	Dynamic Contrast-Enhanced MRI
DNN	Deep Neural Networks
DSC	Dice Similarity Coefficient
DWI	Diffusion-Weighted Imaging
ELU	Exponential Linear Unit
ER	Estrogen Receptor
FCM	Fuzzy C-Means
FN	False Negative
FNN	Feedforward Neural Network
FOV	Field Of View
FP	False Positive
GLCM	Grey-level Co-occurrence Matrix
GLRLM	Grey-level Run-length Matrix
GLSZM	Gray-level Size Zone Matrix
GMM	Gaussian Mixture Models
GPU	Graphical Processing Unit
HER2	Her-2/neu
IBSI	Image Biomarker Standardisation Initiative
ICC	Intraclass Correlation
IHC	Immunohistochemistry
IQR	Interquartile Range
LASSO	Least Absolute Shrinkage and Selection Operator
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LoG	Laplacian of Gaussian
MIM	Mutual Information Maximisation
ML	Machine Learning
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
mRMR	Minimum Redundancy Maximum Relevance
NAC	Neoadjuvant Chemotherapy

NGLDM	Neighbouring Grey-level Dependence Matrix
NGTDM	Neighbouring Grey-tone Difference Matrix
NN	Neural Network
PCA	Principal Component Analysis
pCR	Pathological Complete Response
PET	Positron Emission Tomography
PR	Progesterone Receptor
PReLU	Parametric Rectified Linear Unit
QIN	Quantitative Imaging Network
ReLU	Rectified Linear Unit
RF	Random Forest
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
RSD	Relative Size Difference
SBS	Sequential Backward Selection
SE	Standard Error
SFS	Sequential Forward Selection
SGD	Stochastic Gradient Descent
SPECT	Single-Photon Emission Computed Tomography
SVM	Support Vector Machines
TIC	Time Intensity Curve
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
UDFS	Unsupervised Discriminative Feature Selection
US	Ultrasound

Introduction

In the era of artificial intelligence (AI) and personalised and precision medicine, supervised and unsupervised learning techniques have gained an increasingly relevant role in the advanced analysis of medical images across various fields of modern medicine. This thesis specifically focuses on the application of AI in magnetic resonance imaging (MRI) to predict the pathological response to neoadjuvant chemotherapy in breast cancer.
Neoadjuvant chemotherapy (NAC) is administered in large operable or locally advanced breast cancers to facilitate tumour shrinkage, allowing for breast-conserving surgery. The ideal outcome of NAC is a pathological complete response; however, when this isn't achieved, adverse effects often outweigh the benefits. Therefore, the prior assessment of patients and potential complete responders to NAC holds significant clinical importance for personalized treatment. MRI is routinely used for dense breast imaging and screening in high-risk patients, demonstrating its potential to enhance the clinical diagnosis of breast cancer. Dynamic Contrast-Enhanced MRI (DCE-MRI) stands out as the most informative sequence for breast MR imaging.

This work will delve into the application of radiomics analysis, the process of extracting quantitative features from imaging data, to construct a model capable of predicting NAC outcomes. Radiomics analysis involves the initial stage of lesion delineation, which can be executed manually, semi-automatically, or fully automatically using AI, followed by the extraction of image features that reveal key components of the tumor phenotype. These features, combined with clinical variables, are used to train a classifier capable of discriminating between no response and complete response.

Several critical aspects emerge in the radiomics pipeline, encompassing variability in determining the region of interest (ROI) and the subsequent stability of the extracted features. Moreover, the multidimensional nature of radiomics features necessitates highly accurate and reliable methods to select the most relevant ones and build the predictive model.

Chapter 1 provides a general overview of the work context, defining the most generic radiomic workflow and its application in the context of breast cancer.

In Chapter 2, a general framework of machine learning for radiomics is presented, addressing both deep learning for lesion segmentation and simple classifier models for outcome prediction.

Chapter 3 is dedicated to evaluating observer variability in determining the ROI in the first part, and in the second part, it focuses on the development of a reliable algorithm for automatic segmentation of breast lesions on DCE-MR images. The algorithm's performance is then compared to manual segmentation by an experienced radiologist.

Finally, Chapter 4 is centred on building the predictive model, involving feature extraction and selection, training the classifier, and comparing different models, including radiomic features only and the impact of adding also clinical factors.

Chapter 5 encompasses a discussion of the results, drawing conclusions and providing further perspectives and possible improvements.

Chapter 1

Radiomics from basics to applications in breast cancer

The field of medical imaging commenced in 1895 with Wilhelm Conrad Roentgen's groundbreaking discovery of X-rays, marking a pivotal moment when physicians gained unprecedented access to visualising the internal structures of the human body. Initially, 2D radiography stood as the sole imaging modality. However, subsequent decades witnessed the advent of computed tomography (CT), MRI, ultrasound (US), single-photon emission computed tomography (SPECT), and positron emission tomography (PET). Throughout the last century, this field rapidly evolved, assuming a central role in the healthcare system. Technological advancements not only provided comprehensive views of the human body in 2D, 3D, and 4D but also enabled functional and molecular imaging [9]. These innovative tools swiftly became commonplace in clinical practice, contributing to diagnosis, screening, and treatment planning. Presently, techniques such as MRI and various tomography modalities are routine, particularly in oncology, aiding in preoperative staging and disease extent determination.

The success of medical imaging, however, owes not only to technological advancements but also to the digital revolution. This shift from analog to digital, exemplified for instance by the replacement of radiographic films with computed radiography and CT, not only empowers clinicians to edit images for improved visualisation as a diagnostic tool but also facilitates the storage of images and enables further analysis and comparisons.

Simultaneously, as imaging devices continue to improve in resolution, the healthcare sector concurrently generates an increasing volume of high-resolution data. In the digital age, a single examination can amount to several gigabytes, making a substantial contribution to the overall volume of generated data. It is projected that by 2025, the total generated data will reach up to 175 zettabytes, with a significant portion originating from the medical sector [87]. The burgeoning realm of medical big data presents an opportunity to apply machine learning (ML) techniques, creating computer-aided systems and predictive models to assist specialists in clinical decision support, diagnosis, monitoring, and therapy planning. This, in turn, propels precision medicine towards personalised care. AI's application and the processing of vast datasets might in fact unveil characteristics not easily discernible to the human eye.

Within this framework, a new research area emerges: radiomics, the extraction of quantitative features from imaging data for decision support. Radiomics finds application across various fields, with a significant frontier in oncology. Given the frequent imaging examinations undergone by cancer patients, radiomics enables the extraction of pathophysiological information, reducing the need for invasive procedures. The potential applications encompass automated cancer detection, diagnosis, and prognosis especially for treatment response prediction. Furthermore, extracted information aids in disease status monitoring and distinguishing between benign and malignant tumours. Radiomics emerges as a potent ally for specialists, providing a comprehensive insight into individual patients.

In line with these advancements, this thesis delves into the application of radiomics in the context of breast cancer, aiming to construct a predictive model for assessing the response to neoadjuvant chemotherapy. This chapter offers a comprehensive background on radiomics, its workflow, insights into its specific application in breast cancer, and outlines the objectives of this study.

1.1 Radiomics

In recent years, the growing storage of medical imaging data, coupled with rapid advancements in AI technologies, has paved the way for leveraging this data to obtain quantitative information, enhancing the clinician's experience. Radiomics, defined as the conversion of imaging data into higher-dimensional data followed by data mining and the application of ML or statistical models, stands at the forefront of this transformative approach [29].
This field is particularly intriguing as it operates under the premise that quantitative features extracted from images may encode details pertaining to the underlying pathophysiology of tissues, including the phenotype and genotype of a tumour.
Radiomic analysis complements traditional manual image viewing by extracting features that elude quantitative description by the human eye, offering the potential to uncover correlations with clinical and pathological parameters of patients [101]. This approach facilitates the use of mathematical models capable of making predictions in various fields, fostering personalised treatment for each patient.
Unlike simple computer-aided systems, radiomics involves the calculation of an extensive number of features, capable of addressing diverse issues from disease classification to treatment planning. In oncology, where tumours exhibit significant heterogeneity, radiomics transcends the limitations of subjective semantic features and radiologist experience-based clinical choices. It serves as a valuable complement to more invasive examinations like biopsies, offering a non-invasive means of obtaining comprehensive information without the limitations associated with sampling errors.
The extraction of precise information from images through defined mathematical expressions facilitates an objective and repeatable process in both diagnosis and prognosis, advancing the realms of precision and personalised medicine. Ultimately, the value of radiomics lies in the identification and selection of key quantitative features capable of accurately assessing pathological information. With a foundational understanding of the general concept of radiomics, let us now turn our attention to the practical workflow that underpins its application in oncology.

1.1.1 Radiomics workflow

Despite the simple key concept, radiomics encompasses a series of steps, each beset by its unique challenges. The fundamental stages are depicted in Figure 1.1, involving:

- image acquisition and reconstruction;
- identification and manual or automatised segmentation of the region of interest (ROI);
- image preprocessing;
- extraction of radiomic features.

The outcome is a comprehensive database containing all extracted radiomic features, primed for information mining. Depending on the specific task at hand, the need to retain only key features arises, requiring the workflow to conclude with:

- dimensionality reduction techniques;
- model development.

The last stages allow to establish the relationship between features and the required pathological elements, providing the final radiomic model.

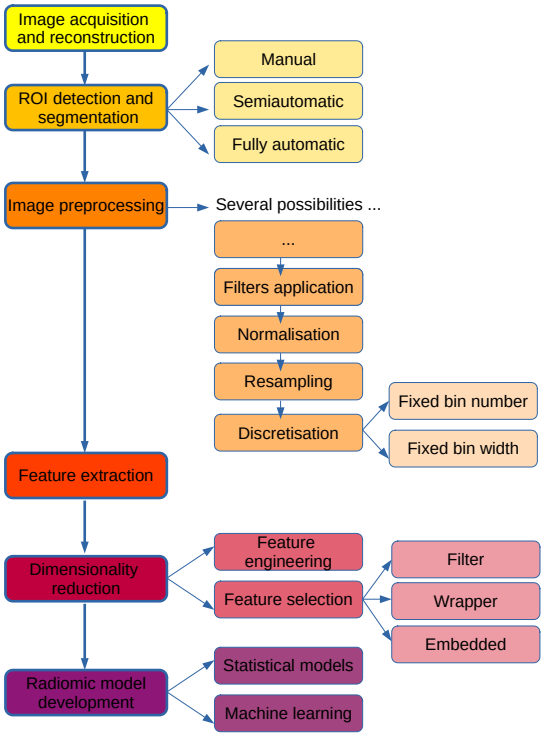


FIGURE 1.1: Example of radiomics workflow, illustrating the different steps and options from image acquisition to model building.

1.1.2 Image acquisition and reconstruction

In the realm of radiomics, the focus is often directed towards extracting expressive features from input data obtained from different medical imaging modalities (e.g. CT, PET, MRI, US, etc.).
Medical images are commonly stored following the widely used DICOM® standards. Depending on the acquisition technique, each image can be represented by a 2D or 3D array, filled with elements known as pixels or voxels, encoding information on intensity with standard or arbitrary units (such as Hounsfield units for CT and arbitrary signal intensity for MRI). For a 2D image M or 3D image V, each intensity value respectively m_{ij} or v_{ijk} is referred to as a grey level or grey tone. Since digital images have finite resolution, intensity values correspond to regular spatial intervals over the acquisition grid. The physical distance between the centres of two elements in any direction, defining the image resolution, is referred to as spacing and is typically measured in millimetres.
Notably, the modern landscape of acquisition units is diverse, encompassing various parameters and reconstruction protocols. This diversity is influenced by many factors, in primis scanner vendors, leading to images with disparate resolutions and quality. Standardisation of acquisition and reconstruction protocols across different centres remains a challenge, as remarked by Gillies et al. [29]. While lack of standardisation poses minimal issues in clinical practice where radiologists primarily view images for semantic features, it becomes a critical concern in the radiomic field. Image variability across different centres or scanners can introduce quantitative differences unrelated to underlying biological effects. Efforts to mitigate such effects include initiatives aimed at establishing definitions for acquisition and reconstruction standards. Additionally, features with high reproducibility, insensitive to acquisition and reconstruction protocols, can be selectively chosen [101]. In this pursuit, the Quantitative Imaging Network (QIN) plays a crucial role. The QIN is actively involved in developing quantitative imaging methods, with specific teams working to validate robust radiomic features for clinical use [79].

1.1.3 Identification and segmentation of the ROI

After image acquisition, expert radiologists analyse images to identify tumoural tissues and suspected areas, which may be located in a single site or multiple locations, especially in patients with metastasis. The subsequent goal is to delineate the borders of the volume of interest, crucial for feature extraction. Considering a medical image of volume V, with w, h, and d, respectively, width, height, and depth in terms of the number of voxels, the whole volume can be represented as a set of $n = w \cdot h \cdot d$ points $X = \{x_1, ..., x_n\}$. Different ROIs can be delineated in terms of segmentations.
One can define the segmentation as the partition $S = \{S^{t0}, .., S^{tm}\}$ of X associated with the membership function

$$f^i(x) = \begin{cases} 1 & \text{if } x \in S^i \\ 0 & \text{if } x \notin S^i \end{cases} \qquad \text{for } i = t0, ..., tm$$

each one identifying a different tissue or region, usually consisting in two classes discerning the background from the main organ or lesion.
The segmentation stage stands out as perhaps the most critical component of radiomics, given that the extracted features are directly linked to the segmented volumes. This task is notably challenging in oncological data due to the indistinct and irregular borders of many tumours. Manual accurate segmentation, typically performed by experienced radiologists, introduces a degree of operator subjectivity, potentially leading to significant inter-operator differences. Manual segmentation is also highly time-consuming, often requiring radiologists to meticulously draw ROIs over hundreds of 2D slices.
To address these challenges, ML-based solutions have been proposed to automate the procedure. These solutions range from semi-automatic methods, where the operator only needs to choose an intensity threshold or place seed points for region growth, to recent advancements in deep learning that enable fully automatic segmentation of organs or lesions. Automation serves as a valuable tool to reduce operator variability and generate a larger amount of labelled data within an acceptable timeframe, thereby increasing database sizes and enhancing the reliability of radiomic models. However, even automated segmentation requires operator supervision to ensure correct performance, particularly in cases of high intersubject variability or differences in image acquisition parameters, with the possibility of manual refinements and corrections. Different organs and tissues exhibit varying challenges for segmentation, necessitating a tailored approach for each task. The segmentation approach remains a topic of debate, with some studies favouring manual segmentation while others, focusing on reproducibility, opt for automatic segmentation.

1.1.4 Image preprocessing

Following the image acquisition, a standard practice involves processing the acquired images to enhance their quality by reducing noise, addressing artifacts, and when necessary applying corrections for motion and other specific effects. It's noteworthy that often these procedures are often executed by scanner vendors directly after image reconstruction. Nevertheless, prior to extracting radiomic features, additional preprocessing steps may prove beneficial in obtaining an optimal set of features. Some of the most common stages involve image normalisation, interpolation and grey level discretisation, but also in some cases the application of filters.

Normalisation
Depending on the imaging modality, grey level intensity might be on a scale with calibrated units or might be arbitrary signal intensity. In general, in the latter case, to account for different ranges, different temporal cohorts, different centres, or scanners, a good approach is to standardise the images by normalising the grey levels to fit in a common range of intensity (I_{min}, I_{max}). This procedure is not

mandatory in the case of calibrated units; however, some studies have evidenced that normalisation might still be beneficial, improving the results of radiomic analysis [73]. Moreover, in this stage, cut-off values might be applied to remove outliers.

Interpolation

It is worth noting that the majority of 3D images do not exhibit isotropic spacing, with the spacing in the cranio-caudal direction typically larger than in the other directions. Calculating features related to texture, which involves groups of multiple voxels in all directions, necessitates rotational invariance, requiring the same spacing in all directions. Therefore, interpolation to isotropic voxel size is always a necessary step. Furthermore, maintaining consistent voxel spacing facilitates the comparison of images from different samples, cohorts, or centres, ensuring reproducibility.

According to the Image Biomarker Standardisation Initiative (IBSI) [123], there are no clear suggestions on whether upsampling or downsampling schemes are more beneficial for data analysis. Upsampling is based on inference, introducing artificial information, while, on the contrary, the choice of downsampling inevitably incurs in loss of information and might produce aliasing artifacts.

Another choice that needs attention is the use of 3D or 2D interpolators for 3D images. In most cases, 3D interpolators are preferred; however, 2D interpolation that avoids interpolating voxels between slices is suggested and beneficial in cases where the spacing between slices is particularly large with respect to the spacing in-plane and/or the final desired voxel size [123]. In fact, the use of a 3D interpolator would lead to the inference of a large number of voxels in-between slices or losing much of the in-plane information. In the end, for such cases, since the spacing is no more isotropic, texture features can only be calculated in-plane.

Also, for the interpolation algorithm, multiple choices are possible, with the most common being the nearest neighbour, trilinear convolution, and tricubic spline interpolation [123]. The former methods assign intensity according to the most nearby voxels in the original image grid, producing blocks of similar intensity, inducing a possible bias in feature calculation. On the other hand, other choices based on polynomial interpolation on a larger neighbourhood provide a smoother intensity transition. However, also trilinear and tricubic methods have their own disadvantages, with, respectively, the possible presence of artefacts when upsampling and the possibility to have out-of-range intensities. Overall, one has to mention that feature reproducibility depends also on the choice of the interpolator, with some features more reproducible using a particular algorithm with respect to another [49].

Coherently if the image is interpolated after the ROI segmentation, also the ROI binary mask should be interpolated to have the same dimensions. In this case, the suggestion [123] is to use the nearest neighbour or trilinear interpolator to obtain meaningful masks; the former choice is the simplest and avoids the presence of new voxels containing fractions of the original ones and the need to choose binarisation thresholds.

Discretisation

The use of discretisation is often necessary to have a tractable extraction of texture features and reduce noise. Depending on the nature of the imaging modality, different methods are preferred.

In the case of arbitrary intensity units, as for MRI images, the recommended choice involves the use of discretisation using a fixed bin number. Considering an imagine with n pixels or voxels, I_k^d the discretised intensity of the element k, obtained by fixing the number of bins N_g is defined as:

$$I_k^d = \begin{cases} \left[N_g \dfrac{I_k - I_{min}}{I_{max} - I_{min}} \right] + 1 & I_k < I_{max} \\ \\ N_g & I_k = I_{max} \end{cases}$$

This method breaks the direct link between intensity and physiological meaning but is beneficial

in cases where contrast is important, moreover, it makes possible to compare features from different samples.
On the contrary, in cases where the intensity is calibrated in units and there is a well-defined minimum in the intensity range, another approach is also possible, involving the use of fixed bin width. In this cases, by fixing the width of the binning w_b, the discretized intensity is obtained as:

$$I_k^d = \left\lfloor \frac{I_k - I_{min}}{w_b} \right\rfloor + 1, \quad \text{maintaining a direct relationship with the original scale of intensity.}$$

Increasing the bin size and decreasing the bin number produce coarser images, allowing computing less noisy features. Depending on the specific task and modality, the most suitable discretisation method and parameters differ. The choice of discretisation has an overall crucial impact on the feature values and reproducibility.

Filters application
In order to enhance the predictive power of radiomic features, one can try to increase image expressiveness by applying filters to the image before extracting features. Various filters can be used, for instance, smoothing filters reducing image noise, wavelet filters accounting for the spectral dimension of the data, exponential filters to highlight subtle differences in the data, and Laplacian of Gaussian filters to enhance edges [30]. In general, the use of preprocessing filters has an impact on radiomic analysis, and despite increasing dimensionality of the dataset, they might improve the predictive performances of radiomic models in some cases [21].
One should, however, pay attention to the application of filters stage since the choice of application after or before resampling might be crucial depending on the filter type. When the analytical expression of the filter is available, and it's defined on the continuous, it can be applied before resampling, allowing anisotropic voxel grids. In the opposite case, applying filters directly with anisotropic image resolution generates response maps that are not directly comparable since having different frequency response. To avoid this effect, it is necessary to perform filtering after resampling the image to uniform spacing. In general, after filtering, image intensities of the response maps have no longer an evident physical meaning, so also the discretisation method should use a fixed bin number accordingly.

1.1.5 Radiomic features extraction

After defining the ROI together with its segmentation mask and undergoing appropriate preprocessing, the focus finally shifts to the extraction of radiomic features from the voxels within the mask as illustrated in the schematic view in Figure 1.2. Radiologists often employ semantic features to describe the ROI, offering qualitative descriptors such as shape, margin spiculation, and vascularisation. In contrast, radiomic features are agnostic, serving as quantitative descriptors with a precise mathematical description. Over the years, hundreds of features have been defined, prompting an attempt to classify them based on their origin. In the end three main distinct feature families emerge: morphological features, first-order features, and textural features.

Morphological features
This category encapsulates the fundamental geometric characteristics of the ROI, such as volume, area, and related dimensions. The representation of the ROI varies depending on the context, offering diverse perspectives for the volume [123], interpreted as:

- collection of voxels, with each single voxel contributing with its own micro volume, particularly useful for raw volume approximations;

- set of points, each one represented by the coordinates of the center of each voxel, suitable in analysing the inner structure;
- surface triangular mesh, ideal approach when the external surface structure holds significance.

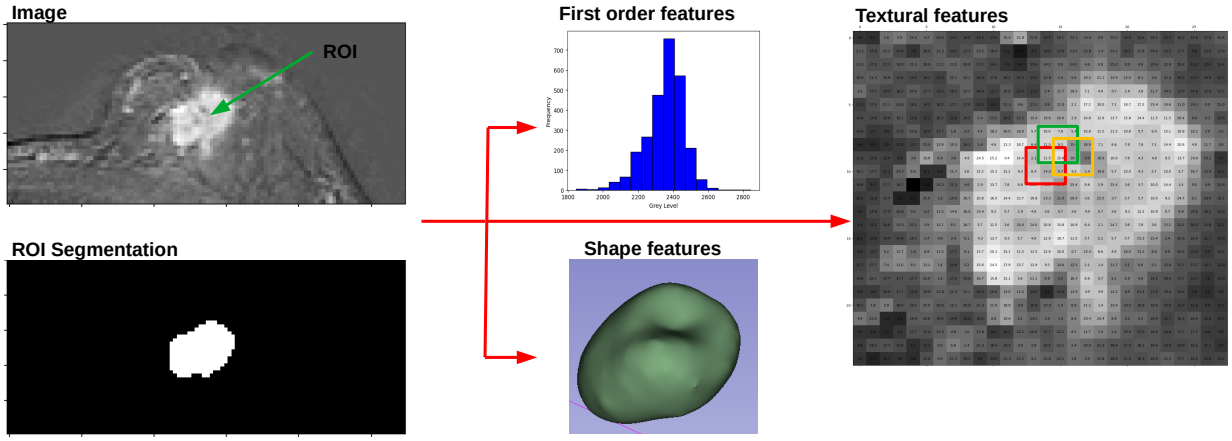


FIGURE 1.2: Schematic view of segmentation and extraction of radiomic features.

Features related to shape are calculated from the 3D mask and are expressed in terms of the standard unit length of 1 mm. Among these features, one can derive both an overestimated voxel-based volume, obtained through simple voxel counting, and a more precise mesh-based volume, computed by approximating with tetrahedrons.
The mesh representation further allows estimation of the surface and the surface-to-volume ratio. Additional shape-related features can be derived from deviations from a sphere-like volume, including compactness and sphericity. Multiple derived definitions can provide highly correlated yet nuanced information.
Moreover, constructing the convex hull of the mesh representation enables the determination of the maximal distance between the most distant vertices in the ROI. For a comprehensive view of the dimensions, principal component analysis (PCA) comes into play, determining an ellipsoid that encloses the volume. The eigenvectors offer orientation insights, while eigenvalues provide information on axis length, elongation, and flatness.

First order features

First order features, often referred to as histogram features, are computed from the distribution of individual voxel values within the ROI mask, with no consideration for spatial relationships among them [4].
Following the discretisation step mentioned in earlier stages, the intensity mask of a ROI with n voxels can be characterised by a discrete set of N_g intensity values $\mathbf{I}^d = \{I_1^d, ..., I_n^d\}$.
One can thus define the histogram $\mathbf{H} = \{f_1, ..., f_{N_g}\}$ where f_i is the frequency count for the grey value $i \in [1, N_g]$, and get an approximated occurrence probability as $p_i = \dfrac{f_i}{n}$.

From the frequency and occurrence probability of each intensity, various metrics can be obtained as: minimum, maximum, range, sample median, mean, variance, standard deviation, skewness, kurtosis, percentiles, interquartile range, Shannon entropy , uniformity , and many other derived measures. For formal definitions of all validated features, please refer to the IBSI guidelines [123].

First-order features provide a comprehensive overview of the distribution and characteristics of voxel intensities within the ROI, forming a basis for quantitative analysis.

Textural features
Textural features form an extensive set of characteristics that concentrate on image texture. Initially devised to assess the texture of surfaces in 2D images, their extension to 3D images is straightforward. This family of features is itself composed of different subsets: second-order features, delving into the intensity relationships between pairs of neighbouring voxels obtained from the Gray-level Co-occurrence Matrix Features (GLCM), and higher-order features accentuating connections among a greater number of voxels obtained from Gray-level Run-length Matrix (GLRLM), Gray-level Size Zone Matrix (GLSZM), Neighbouring Gray-tone Difference Matrix (NGTDM), and Neighbouring Gray-level Dependence Matrix (NGLDM).

- **GLCM features**
  The GLCM is a matrix that captures the distribution of discretised intensities among connected neighbouring pixels or voxels along a specified direction. In 2D images, an 8-connected neighbourhood is constructed using four orthogonal direction vectors. In the case of 3D images, a 26-connected neighbourhood is formed, utilising 13 unique direction vectors. Let N_g represent the number of discretised grey levels. For each direction vector $\mathbf{v}$, the GLCM is defined as a matrix $\mathbf{M}^v$ of size $N_g \times N_g$. Each element m_{ij} in this matrix signifies the frequency of the combination of grey levels i and j in neighbouring voxels across directions $v_+ = \mathbf{v}$ and $v_- = -\mathbf{v}$. The probability distribution for grey level co-occurrences can be obtained by normalising $\mathbf{M}^v$, resulting in $\mathbf{P}^v$. Each element p_{ij} in $\mathbf{P}^v$ represents the joint probability of having grey levels i and j in neighbouring voxels in the direction $\mathbf{v}$. It is important to note that both $\mathbf{M}^v$ and $\mathbf{P}^v$ are symmetric by definition. Then features are obtained from the probability of gray-level co-occurrence, providing among the most commons joint average and variance, entropy, contrast, correlations, cluster tendency, and various measures of homogeneity like inverse variance.

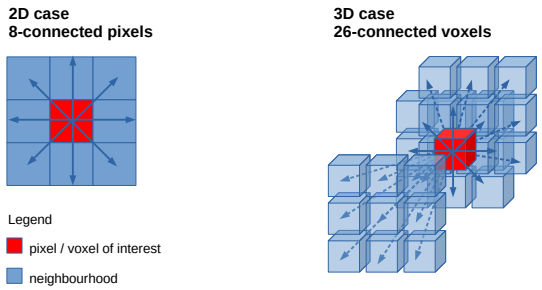


FIGURE 1.3: Example of 2D and 3D neighbourhood definition for GLCM. Pixel and voxels are neighbours with any other touching one of their faces, edges, or corners.

- **GLRLM features**
  Similar to GLCM, GLRLM also assesses the distribution of grey levels in the image. It is based on the use of run lengths, i.e. consecutive sequences of pixels/voxels with the same grey level along a given direction $\mathbf{v}$. The matrix elements are in fact the occurrences of runs with length j for a discretised grey level i. Metrics about uniformity, variance, entropy, emphasis of both high and low gray levels, and short and long run lengths, can be estimated.

- **GLSZM features**
  GLSZM quantifies instead the number of groups (zones) of linked voxels in the neighbourhood, where voxels share the same discretised grey level values. In 3D, a voxel is linked to its neighbourhood if all 26 neighbouring voxels have the same grey level, while in 2D, 8-connectedness is used. The GLSZM is a matrix with elements represented by s_{ij}, denoting the number of zones with discretised grey level i and size j. Features are defined similarly to the previous cases, encompassing metrics about uniformity, variance, entropy, emphasis of both high and low grey levels, and small and large zones, along with combinations of these.

- **NGTDM features**
  NGTDM serves as an alternative to GLCM, encompassing the sum of grey levels differences with discretised grey level i and the average discretised grey level of neighbouring pixels up to a given Chebyshev distance δ. A significant feature within this class is coarseness, evaluating the spatial rate of change in intensity, and assessing the presence of coarse textures and large-scale patterns. Also contrast can be obtained from the dynamic range and spatial frequency of grey level changes. Additionally, the presence of large changes in grey levels and non-uniformity in intensity changes can be evaluated as texture busyness, complexity, and strength.

- **NGLDM features**
  NGLDM serves as a rotationally invariant alternative to GLCM and GLRLM. It describes dependency of intensity, evaluating relationships between any central pixel and its neighbours within a window. The matrix represents the number of pixel pairs (central and neighbouring pixels) that have intensity differences lower than a given threshold. Features within this category are defined similarly to previous cases, including the emphasis of low and high dependence, grey level count, with all possible combinations, and the usual uniformity, variance, and entropy.

Features interpretation
When it comes to interpreting features, morphological features offer straightforward visualisability and a direct correlation with tumour phenotype. Deviations from sphericity at varying degrees can be associated with different tumour types. Similarly, first-order features can be linked to diseases displaying either heightened or reduced intensity enhancement, with entropy variations contingent on the tissue type involved. On the other hand, features of higher order may lack immediate interpretation but can be connected to specific textural characteristics in diverse scenarios [4]. For example, a high correlation might indicate the presence of linear or honeycomb patterns, while coarser or finer textures could be indicative of biological differences in tissue properties. Unravelling these connections might enhance our understanding of underlying functioning, contributing to more nuanced and informed medical interpretations.

1.1.6 Dimensionality reduction

After the extraction phase, each image is characterised by a substantial and challenging number of features. Indeed, the volume of features frequently exceeds the number of cases in the dataset, necessitating careful considerations in their selection. An initial criterion for selection could be based on robustness and reproducibility across various parameters, such as imaging techniques, different operator segmentations, and filtering. Additionally, many features exhibit high correlation and redundancy by definition.
Efficient analysis thus requires reducing the total number of features, achieved through either feature selection or feature engineering. Feature selection involves choosing the most informative features from a larger set, while feature engineering entails crafting new features through combinations of existing ones.

In the realm of feature selection, diverse methods exist. Simple filter methods eliminate for instance non-reproducible, highly correlated, and low-variance features. Other approaches, integrated into the model development stage, use information scores, dependencies with the desired outcome, wrapper methods with iterative selection based on predictive model improvements, and embedded methods tied to the prediction model itself.
A pictorial example of feature engineering for dimensionality reduction is instead the PCA. This method provide principal components as new orthogonal features with higher variance, obtained from linear combinations of the original ones. Through this process, PCA transforms the data into a more condensed form, preserving essential information while alleviating the curse of having more features than data.

1.1.7 Model development

The ultimate objective of radiomic analysis lies in establishing a meaningful relationship between image features and the phenotype of tissues, including their molecular characteristics. Analysis of the extracted features serves as the vital bridge connecting images to clinical outcomes, with the aspiration that these radiomic features are not only reliable but also reproducible for application in clinical diagnostics [101].
The overarching strategy involves preserving as much data as possible upfront, utilising data mining downstream to identify features with the highest prognostic value. This approach stems from the belief that filtering at the input stage would be inefficient and presuppose a prior understanding of the biological meaning of features before testing the model [29]. Robust features that survive the preceding selection stages are then employed to construct models that relate them to specific biomarkers or clinical endpoints.
This form of analysis can either focus on inference and building statistical models to assess correlations between data or prioritise accurate predictions using ML models to address specific questions.
Many applications convert the goal of finding correlations between radiomics and clinical information into a classification problem. For instance, determining whether a tumour is malignant or benign, assessing the expression of a specific gene, or discriminating between different therapy outcomes. ML offers a diverse array of models for various tasks, and the choice of a specific model often hinges on the desired outcome.
In practice, employing multiple models and selecting them based on performance is a common strategy. Regardless, any generated model must undergo evaluation on an independent validation set.
However, even in the end, the efficacy of a successful radiomic model hinges on whether it can offer more clinical benefits than the judgment of clinicians. Models that establish a connected biological meaning and undergo biological validation are highly preferred [103]. Without a comprehension of the biological rationale, radiomic features and models may appear as black boxes, impeding widespread adoption and making validation and acceptance particularly challenging. Incorporating a biological context into radiomic models not only fortifies conclusions but also opens up additional avenues for validation and further investigation. Providing a deeper understanding of the underlying biological mechanisms enhances the interpretability of the models, fostering trust and facilitating their integration into clinical decision-making processes.

Radiomics challenges

The intrinsic strength of any model lies in having sufficient statistics, emphasizing the need to collect and integrate large-scale medical image data for constructing high-quality datasets. A general guideline is that each feature requires at least a tenth of the samples to be encoded, highlighting the crucial role of gathering and sharing data to build larger datasets and generate more reliable models.
However, this undertaking is far from straightforward. Medical images are dispersed across various

institutions, and the quantity of available medical images is directly linked to the size of the local population. Furthermore, achieving accurate prognostic estimates necessitates monitoring patients over several years. The complexity is heightened by variations in instruments, parameters, and methods across different healthcare facilities, leading to highly inconsistent and, at times, incomplete imaging data. Ambiguous findings add another layer of complexity, further reducing the dataset size and requiring precise dataset curation.
Despite these challenges, radiomics harbours immense potential to become a valuable and applicable tool. This potential can be realized through collaborative efforts across institutions, with the implementation of well-defined protocols to standardise data collection and ensure the reliability of radiomic models across diverse healthcare settings.

1.2 Background on breast cancer

As previously highlighted, radiomics research is flourishing, particularly in the field of oncology. The object of this study is the use of radiomics for breast cancer. However, before delving into the details of the radiomic model, it is essential to understand the clinical perspective and the key factors influencing the diagnostic and prognostic process.

1.2.1 Breast cancer prevalence and distribution

The global incidence of all cancer types annually approaches almost 20 million cases, resulting in over 9 million deaths, according to estimates up to the year 2020 [98]. More recent estimates of worldwide cancer incidence and mortality are collected by the Global Cancer Statistics (GLOBOCAN2022 `https://gco.iarc.who.int/today` accessed 18-03-24.), as reported in Figure 1.4. Among these, breast cancer stands out as one of the most prevalent, with an annual occurrence surpassing 2.3 million patients, ranking second only to lung cancer. In terms of mortality, it holds the fourth position, following lung cancer, colorectal cancer, and liver cancer, with over 666 thousand deaths each year.

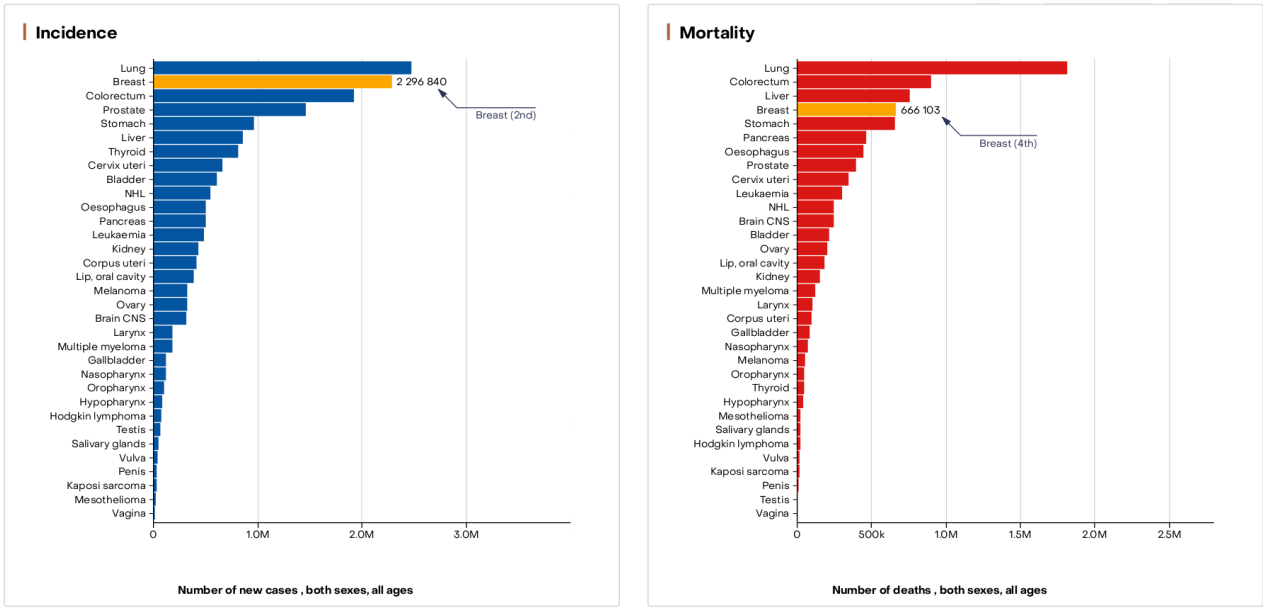


FIGURE 1.4: Estimates of the absolute number of incidences and mortalities for all cancers are respectively shown on the left and right sides. Breast cancer ranks second in terms of absolute incidence and fourth in terms of mortality. Adapted from GLOBOCAN2022 `https://gco.iarc.who.int/today` accessed 18-03-24.

Primarily affecting women, breast cancer stands as the most commonly diagnosed cancer, comprising 24.5% of all cancers in women and ranking among the leading causes of mortality, representing 15.5% of female deaths attributed to cancer. Current estimates in Italy [1] indicate an even higher local prevalence, accounting for about 30% of all tumour sites in women, with a projection of over 55 thousand new cases per year.
Recent studies [51] suggest a rising trend in global breast cancer incidence, attributed to increased risk factors, population growth and ageing, and enhanced screening detection. While the incidence is higher in highly developed countries, driven by effective screening modalities, mortality displays a decreasing trend due to advancements in medical treatment. Conversely, in less developed countries lacking efficient prevention, screening, and treatment, there is an alarming increase in mortality.
Risk factors for breast cancer in women are diverse, encompassing reproductive and hormonal elements (e.g. early age at menarche, later age at menopause, advanced age at first birth, reduced number of children, reduced breastfeeding, use of hormone therapy in menopause, use of oral contraceptives, etc.), lifestyle factors (e.g. excessive body weight, sedentary lifestyle, alcohol consumption), and genetic mutations (especially BRCA1, BRCA2 mutations). Family history of breast cancer in other women also elevates the risk.
Considering the global burden of this cancer, heightened awareness, preventive strategies, and improved access to therapy are increasingly imperative. Moreover breast cancer exhibits significant variability in phenotype, genotypes, and therapy choices. Ongoing efforts are directed towards personalised oncological treatments, aiming to provide more targeted therapy tailored to each tumour's specific phenotype and genetic pattern. Radiomic research in this field plays a crucial role in different fronts both in classifying tumour types, distinguishing between malignant and benign tumours,and/or different phenotypes, and predicting prognosis and response to specific treatments in early stages, providing suggestions for more informed therapeutic decisions.

1.2.2 From diagnosis to therapy

The diagnosis of breast cancer typically arises from either diagnostic exams prompted by specific symptoms (such as pain or the detection of a palpable mass) or preventive screening.
Mammography is the most commonly used imaging technique for preventive screening, contributing to a significant reduction (19%) in overall breast cancer mortality. To enhance the accuracy of mammography and reduce false positives, complementary examinations such as digital tomosynthesis, MRI, or US are often employed.
Upon the discovery of suspicious tissue, invasive pathologic evaluation becomes crucial. Techniques such as fine-needle aspiration, core biopsy, or surgical excision are employed for tissue sampling. Histological analysis, including immunohistochemistry (IHC), in situ hybridisation, and molecular tests, provides valuable information for differentiating closely related diseases. Histologic markers like oestrogen receptor (ER), progesterone receptor (PR), and Her-2/neu (HER2) are pivotal in determining treatment responses to targeted agents.
However, the intrinsic variability within tumours poses limitations to the accuracy of these markers, as a single biopsy specimen may not fully capture the heterogeneity of the tumour site. To address the sampling error inherent in biopsy-based assessments, imaging techniques offer a more comprehensive overview. For this aim, radiomic analysis, which extracts quantitative features from the entire ROI, can offer valuable insights into the overall tumour characteristics, contributing to a thorough understanding of its heterogeneity.
In tumour staging, the usual screening tools (mammography, ultrasound, etc.) are commonly employed. However, in specific scenarios, additional imaging modalities like MRI, CT, or PET/CT might be utilised to evaluate distant metastases, depending on the stage.

The role of MRI in breast cancer screening and treatment has become increasingly significant, particularly in cases where mammography may be less effective. While mammography remains a widely accessible and cost-effective screening technique with proven benefits in reducing breast cancer mortality, its sensitivity diminishes in individuals with dense breast tissue, especially among younger women and those with BRCA gene mutations [68]. An example is illustrated in Figure 1.5, showcasing a woman with dense breast tissue who underwent both mammography and DCE-MRI. On the left side, mammography fails to detect any malignancies, while on the right side, suspicious lesions with heightened contrast uptake are observed in both the right and left breasts.

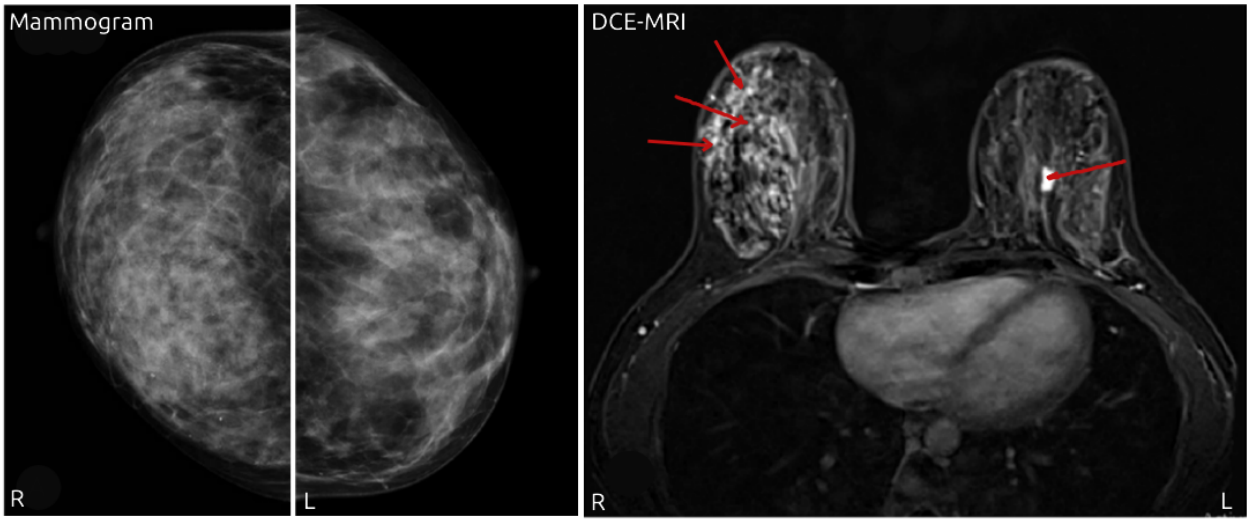


FIGURE 1.5: On the left side, a bilateral digital mammography of a woman with dense breasts shows no suspicious lesions but warrants further investigation. On the right side, a dynamic contrast-enhanced image of the same woman reveals suspicious contrast uptake in both breasts, with a clumped area indicating potential malignancy on the right breast and a small enhanced mass on the left breast. Adapted from [45].

MRI emerges as a valuable alternative due to its higher sensitivity and independence from breast density. Moreover, another advantage of MRI is that it does not use ionising radiation, unlike mammography, which utilizes X-rays. Therefore, MRI provides a safe means to investigate structures without the risks associated with ionising radiation. Its application as a screening tool becomes particularly pertinent for high-risk patients with a family history of breast cancer or suspected BRCA mutations. MRI's ability to detect cancer foci not easily visible through physical examination, mammography, or ultrasound further enhances its diagnostic utility. MRI also aids in the investigation of breast cancer with axillary metastases, identifying the primary tumour site that may remain undetected by other imaging techniques.

Regarding therapy, historically, radical mastectomy was a prevalent surgical intervention for breast cancer. However, nuanced approaches, such as breast-conserving strategies, radiotherapy, and chemotherapy, have emerged over time, tailored to the histopathologic characteristics of each case.

Beyond screening, MRI plays a crucial role in assessing the efficacy of pre-operative therapy, known as neoadjuvant therapy, which can potentially enable a breast-conserving approach instead of a full mastectomy. Moreover, it assists in evaluating the precise extension of residual cancer after therapy administration. The true potential of MRI lies in its predictive capacity for biological behaviour. Several studies have investigated MRI's capability to predict pathological complete response to neoadjuvant therapy. Early changes in intracellular metabolism, detectable through MRI, appear to be indicative of treatment response [68].

The landscape of breast cancer treatment is rapidly evolving from a one-size-fits-all approach to personalised medicine. With advancements in diagnostic tools such as genomic expression profiles and molecular imaging, the ability to characterise breast cancer more accurately has improved, leading to the emergence of personalised treatment strategies.
The exploration of gene expression profiling has been instrumental in categorising breast cancer into a heterogeneous group of diseases characterised by diverse molecular aberrations. This distinction into various subtypes is crucial as different histological features exhibit distinct clinical behaviours and treatment responses [83].
In the early stages, dating back to the 1970s, the initial categorisation into subsets was primarily based on oestrogen receptor expression, highlighting variations in clinical subgroups. Treatment decisions relied on clinical variables such as tumour size, lymph node metastasis, and histological grade. Subsequently, predictive markers became essential for choosing between endocrine therapy and the use of specific monoclonal antibodies (e.g. trastuzumab). ER and PR were employed for endocrine therapy, while HER2 protein expression guided trastuzumab therapy. The revelation that treatment response is linked to intrinsic molecular characteristics rather than anatomical prognostic factors, such as size and lymph node status, marked a significant shift in understanding breast cancer dynamics.
Beyond the broad categorisation into ER-positive and ER-negative cancers, further molecular distinctions can be made. At least four molecular subtypes have been identified: luminal, HER2-enriched, basal, and normal breast-like. These distinctions at the RNA level involve ER, PR, proliferation-related genes, and HER2-related genes, with the additional potential contribution of Ki67 expression [83].
The inherent heterogeneity of breast cancers necessitates an individualized prognosis evaluation to determine the most effective therapy. However, sampling errors and unclear dependencies can lead to incomplete and limited information, posing challenges in predicting the response to a specific therapy. The ongoing research in radiomics aims to overcome these challenges and enhance the precision of treatment predictions. While genomic analysis of breast cancer subtypes has yielded valuable insights, the practical focus in clinical settings often revolves around discerning which patients benefit from specific types of therapy rather than pinpointing the particular molecular subtype [19]. Additionally, given the increasing incidence of breast cancer and the limited availability of complex molecular tests in less developed regions, it may be more pragmatic to initially assess the potential therapy response using less expensive tests. These include immunohistochemical tests for oestrogen and progesterone receptors, in situ hybridization for HER2 overexpression, and in some cases, the evaluation of Ki67 expression.
Standard categorisations typically involve tumour types as follows:

- Triple negative (both hormone receptor-negative and HER2-negative)
- Hormone receptor-negative, HER2-positive
- Hormone receptor-positive, HER2-negative

However, treatment recommendations, especially for the last type, can sometimes be controversial, highlighting the complexities involved in tailoring therapies to individual patients. This underscores the importance of refining predictive models, such as those in radiomics research, to enhance the accuracy of therapy response predictions and guide treatment decisions effectively.

1.2.3 Neoadjuvant chemotherapy

In the context of locally advanced breast cancer, where tumors have a significant extent, are fixed to the chest wall or skin, or exhibit extensive axillary nodal disease, surgical treatment becomes challenging. Neoadjuvant chemotherapy provides assistance and hope by downsizing or eradicating such

primary tumors, enabling routine surgical procedures. Patients achieving a complete or significant partial response to therapy may undergo surgery for previously inoperable diseases. Additionally, NAC is particularly beneficial for operable tumors with an unfavorable tumour-to-breast size ratio, increasing the potential for breast conservation, avoiding mastectomy in favour of local excision. Combining NAC with standard surgery and radiotherapy enhances the overall survival and control of malignancy [82]. Assessing a clinical complete response necessitates imaging evaluation. However, not all cancers respond uniformly to cytotoxic chemotherapy, and some exhibit incomplete and fragmented responses. Pinder et al. [78] proposed a classification of pathological response with three primary classes: pathological complete response (pCR), partial response, and no response. Further subclassification within the same response class is possible based on the percentage of residual tumor remaining and the presence or absence of ductal carcinoma in situ, as detailed in Table 1.1.

Pinder class	**Description**
1-i	pathological complete response, no ductal carcinoma in situ
1-ii	pathological complete response, included ductal carcinoma in situ
2-i	response $> 90\%$ (or 10% of invasive tumour left)
2-ii	response of $50 - 90\%$ (or 10-50 % of invasive tumour left)
2-iii	response $< 50\%$ (or $>0\%$ of invasive tumour left)
3	no sign of response

TABLE 1.1: Pinder classification for pathological response.

On the other hand, like all therapies, NAC has side effects, and in cases of no response, it may pose more disadvantages for patients. Indeed, NAC may result in various side effects, as outlined by the World Health Organization and the National Cancer Institute Common Terminology Criteria for Adverse Events, including but not limited to febrile neutropenia, neutropenia, cardiac and pulmonary toxicity, neurotoxicity, hematological malignancy, and even treatment-related death [114]. Considering the potential adverse effects associated with NAC, it becomes crucial to weigh the risks and benefits of the treatment. For cases where a complete or partial response is not achieved, patients might experience only adverse events without reaping the therapeutic benefits. Therefore, it would be beneficial to predict the likely outcome of NAC in advance. Such predictive capabilities could help exclude patients from treatment when the potential benefits are deemed to be outweighed by the risks and adverse effects.
In this context, the purpose of this work, i.e. radiomic research on high resolution MRI images aiming to predict NAC response, offers the prospect of reducing toxicity and expenses associated with continuing chemotherapy when not beneficial.

1.2.4 Prognostic and predictive factors

The choice of the most appropriate treatment for individual breast cancer patients involves considering various prognostic clinical factors. Identifying the most influential factors is crucial to determine which patients are likely to benefit from specific treatments or to differentiate groups that can potentially avoid certain therapies. Some of the well-established factors include:

- **tumor size**, typically correlated with survival outcomes; nonetheless, clinical and radiological assessments may occasionally be inaccurate, additionally, even small symptomatic tumors measuring ≤ 1 cm in size might exhibit nodal metastasis in 20% of cases;

- **axillary node status**, indicating the presence or absence of nodal involvement, stands as one of the most crucial predictors of survival, with the number of involved axillary lymph nodes highly correlated with survival outcomes.

Clinical staging is in fact typically determined based on tumor size and nodal status. Furthermore, multifocality and multicentricity, representing the presence of multiple tumoral foci, frequently align with nodal metastases, thus acting as unfavorable prognostic indicators.
Moreover, other factors play an important role:

- **histological grade**, determined by a combination of scores for mitotic rate, nuclear grade, and morphological appearance, is strongly associated with lymphovascular invasion and permeation, which correlate with a worse prognosis even in the absence of nodal metastases. Conversely, lower grades are typically indicative of a better prognosis. Certain specific types of invasive breast cancer, such as tubular, papillary, and cribriform, exhibit significantly better prognoses compared to ductal cancer of no special type;

- **patient age**, individuals under 35 years old often present with high-grade tumours and exhibit poorer survival rates. Age frequently serves as a crucial predictor of responses to both chemotherapy and hormone therapy;

- **oestrogen and progesterone receptor status** play a significant role in breast cancer prognosis and treatment. Positivity to hormonal receptors allows for the use of endocrine therapy, which in turn improves survival rates. Their presence is indicative of a higher response rate to endocrine therapy. Conversely, if one or both receptors are negative, the response rate diminishes progressively;

- **proliferation markers**, such as the Ki67/MIB1 index and the number of cells in active cell division (S phase), serve as prognostic factors. Although they may not have a well-defined threshold, they provide valuable insights into epithelial proliferation rates;

- **overexpression of HER2**, frequently observed in invasive breast cancers, is linked to higher rates of recurrence and poorer overall survival. Additionally, tumours with elevated HER2 levels often exhibit resistance to hormonal therapy and chemotherapy. This marker holds both prognostic significance and predictive value for assessing response to therapy.

Other factors with potential prognostic value, such as the presence of tumour-infiltrating lymphocytes, are also present, although their predictive power for therapy has not been fully assessed [10].
Ultimately, the multitude of clinical factors that offer insights into the response to therapy allows for the integration of both clinical and radiomic information, thereby enhancing the development of more robust predictive models.

1.2.5 DCE-MRI

Among the various imaging techniques mentioned previously, the emphasis on evaluating and predicting NAC outcome will be placed on MRI, given its numerous advantages. MRI is becoming increasingly used in both screening, diagnosis and therapy assessment phases.
The MRI signal is generated by the resonance of protons, mainly present in water and fatty tissues, contributing to the overall brightness of the image. Resulting images show parenchyma, fat, and lesions when present. To better detect lesions, a paramagnetic contrast agent (often gadolinium-based) is often injected.
Since signals from fat and lesions are similar, requiring fat suppression and subtraction of images before and after the contrast medium injection is necessary to better visualise the lesions [89].

Dynamic contrast-enhanced MRI is one of the most sensitive techniques, providing information on both morphology and, in a certain sense, functional information about perfusion and vascularity [60]. Usually, DCE-MRI consists of a collection of a pre-injection image and a sequence of images after contrast injection, capable of capturing differences in uptake and washout of the contrast agent by evaluation at multiple time points, typically in the range of 1 to 3 minutes after contrast injection. The evaluation of contrast uptake and washout patterns can, in fact, be used to identify different conditions [89].
A useful tool to evaluate the enhancement kinetics is the use of a semiquantitative time intensity curve (TIC), as recommended by The American College of Radiology (ACR) in the Breast Imaging-Reporting and Data System (BI-RADS). Also, quantitative analysis can be performed by means of specific pharmacokinetic models based on different distributions of contrast agents between intravascular and interstitial space. In general, models based on DCE-MRI have demonstrated to be useful in both classification between malignant and benign lesions, assessing the response to NAC, and predicting the pathological complete response in breast cancer [60].
MRI imaging is essential in therapy response evaluation. The most important features to describe a lesion according to the ACR, BI-RADS involve the characterisation of the morphology and the enhancement patterns. Regarding morphology, suspicious areas can be defined as focus or foci (with a diameter < 5mm), mass (3D lesion with a convex margin), or non-mass. The characterisation of a mass involves characteristics of shape (irregular, round, oval, lobular), margin (smooth, spiculated, irregular), and the pattern of internal enhancement (homogeneous, heterogeneous, central, septal). Non-mass lesions should include distribution characteristics and symmetry. Usually, benign lesions are more related to regular shape and margins.

Enhancement patterns are analysed by means of TIC, e.g., a curve obtained from signal intensity in the breast tissue ROI as a function of time after contrast material injection. In particular, three different enhancement patterns can be distinguished:

- **Type I:** progressive continuous increase in signal intensity, usually associated with benign findings;
- **Type II:** plateau pattern, consisting of an initial increase in intensity followed by a plateau, often associated with malignancy;
- **Type III:** washout enhancement pattern, having an initial increase and successive decrease in intensity, also associated with malignancy.

An illustration depicting the different types of kinetic curves is presented in Figure 1.6. The kinetic curve information is contained in both the initial peak and delayed phase; however, the curve alone does not provide enough information, needing integration with morphological characteristics.
Many tumours can be detected only after contrast agent injection. To achieve optimal sensitivity, high-resolution T1-weighted images are acquired. Morphological features are better evaluated on high spatial resolution images, with the acquisition of thin sections and a small field of view. Whereas enhancement time course can be estimated also in cases with decreased temporal resolution [59].
Moreover, further improvement in functional imaging MRI, including diffusion-weighted imaging (DWI) and spectroscopy, can be combined to provide multiparametric information able to quantify the development and progression of breast cancer, assisting in the prediction of NAC.

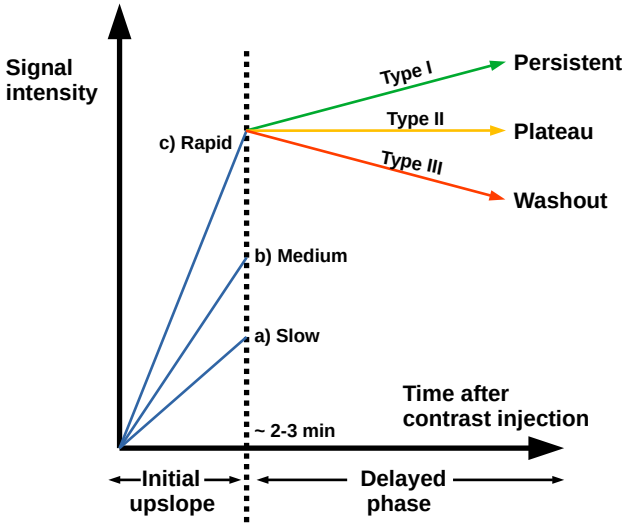


FIGURE 1.6: Example of kinetic curves, i.e., signal intensity as a function of time after contrast agent injection. Curve interpretation involves two phases: the initial upslope and the delayed phase. The former lasts for about 2-3 minutes and can be a) slow, b) medium, c) rapid. Then the trend changes with three possibilities: type I with continuous increase in enhancement (persistent pattern), type II with steady leveling (plateau pattern), and type III with a decrease in signal intensity (washout pattern). Type II and III are usually associated with malignancy, whereas type I is detected for benign lesions. Adapted from [23].

1.2.6 Purpose of the study

After providing an overview of radiomics and exploring the challenges associated with breast cancer, the focus now shifts to the specific objectives of this study. Amidst the nuanced considerations surrounding neoadjuvant chemotherapy, its potential benefits must be balanced against the risk of adverse effects and deteriorating clinical conditions in certain scenarios.
This study aims to address this balance by developing a predictive classifier for NAC outcomes.
Leveraging a dataset comprising dynamic contrast-enhanced magnetic resonance images of women with malignant breast lesions, a comprehensive radiomic pipeline will be employed.
Machine learning models will be trained at various stages, encompassing lesion segmentation and outcome prediction. Further elucidation on the machine learning methodologies adopted will be provided in the subsequent chapter.

Chapter 2

General framework in Machine Learning for Radiomics

Machine learning, akin to data science and statistics, is a field dedicated to understanding how to extract valuable information from data and make predictions. Specifically, machine learning represents a subset of artificial intelligence focused on developing algorithms capable of autonomously learning from data. It has emerged as a cornerstone of modern research, finding applications across diverse sectors. Given the breadth of this field, this chapter aims to offer a succinct overview of general concepts and commonly used techniques within the context of radiomics. Special attention will be given to binary classifiers for predictive models, segmentation models for image analysis, and associated evaluation metrics. Additionally, the chapter will delve into the challenges posed by high-dimensional data and explore methods for feature selection.

2.1 Introduction to Machine Learning

Machine Learning can address various problems, all of which can be conceptualised within a common framework. In this framework, we start with a generic object of study and define an observable quantity $\mathbf{x}$ of the system, along with a model $p(\mathbf{x}|\theta)$ that describes the probability of observing $\mathbf{x}$ given some parameters θ. Once we have collected a dataset $\mathbf{X}$ consisting of observations of $\mathbf{x}$, we can use this data to 'fit the model' by finding the best set of parameters $\hat{\theta}$ that explain the data [63]. At this juncture, we can distinguish between:

- **estimation problems**, concerning the accuracy of the parameter estimates $\hat{\theta}$;
- **prediction problems**, focusing on the model's ability to predict new observations, aiming to maximise the accuracy of $p(\mathbf{x}|\theta)$.

Another distinction can be made by considering supervised and unsupervised problems. In supervised learning, the algorithm learns from labelled data, where each input is associated with a corresponding target output. This allows the model to make predictions on new, unseen data by learning patterns from the labelled examples. Conversely, in unsupervised learning, the algorithm works with unlabelled data, seeking to find hidden patterns or structures within the data itself. Given the emphasis of this work on predictive models, our primary focus will be on addressing prediction problems through supervised learning techniques.
Let's delve into the details of setting up a prediction problem in machine learning. The first component is the dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, which comprises a matrix of input variables (also referred to as predictors or independent variables)[1] denoted by $\mathbf{X}$, and a vector of response variables (also known as outcomes or dependent variables) denoted by $\mathbf{y}$. The second component is the model $f(\mathbf{x}; \boldsymbol{\theta})$, defined as the function $f : \mathbf{x} \to y$ of the parameters $\boldsymbol{\theta}$. The final fundamental element is the cost function (or loss function) $\mathcal{C}(\mathbf{y}, f(\mathbf{X}; \boldsymbol{\theta}))$, which enables us to assess how the model performs on the given observations by measuring the similarity between the response produced by the model from observations $\mathbf{x}$ and the

[1] Note the ambiguity, as input variables may exhibit high correlation among themselves.

available data $\mathbf{y}$. Fitting the model then entails searching for the optimal parameter values $\boldsymbol{\theta}$ that minimise the cost function. To build a useful model suitable for prediction, several steps are involved:

- **Dataset splitting:** the dataset $\mathcal{D}$ must be randomly divided into two mutually exclusive groups, typically with a more substantial portion of data allocated to the training group ($\mathcal{D}_{\text{train}}$) and the rest to the testing group ($\mathcal{D}_{\text{test}}$).
- **Model training:** the model is trained by minimizing the cost function using only the training set, resulting in optimal parameter estimates $\hat{\theta} = \arg\min_{\boldsymbol{\theta}} \{\mathcal{C}(\mathbf{y}_{\text{train}}, f(\mathbf{X}_{\text{train}}; \boldsymbol{\theta}))\}$.
- **Assessing model performance**: the cost function is evaluated on the test set data, $\mathcal{C}(\mathbf{y}_{\text{test}}, f(\mathbf{X}_{\text{test}}; \hat{\theta}))$. Considering the best-fit model $f(\mathbf{X}_{\text{test}}; \hat{\theta})$, one can define the in-sample error as $E_{\text{in}} = \mathcal{C}(\mathbf{y}_{\text{train}}, f(\mathbf{X}_{\text{train}}; \hat{\theta}))$ and the out-of-sample (or generalisation) error as $E_{\text{out}} = \mathcal{C}(\mathbf{y}_{\text{test}}, f(\mathbf{X}_{\text{test}}; \hat{\theta}))$. It is fundamental to note that $E_{\text{out}} \geq E_{\text{in}}$.

The first step, also known as cross-validation, is essential to provide an unbiased estimate of the predictive performance of the obtained model. In traditional statistical approaches, the starting point is a mathematical model assumed to be true, and the goal is to estimate the parameters. In contrast, the essence of machine learning lies in inference about more complex systems, often with high-dimensional data, for which the true form of the mathematical model is unknown. In this context, there are multiple candidate models that need to be compared. The metric of comparison is often based on selecting the best model with the minimum E_{out} [63].
It's important to distinguish between fitting existing data and making predictions about new data. The model with the lowest in-sample error E_{in}, which fits the training data best, often doesn't have the lowest out-of-sample error E_{out}. This discrepancy becomes more pronounced as both the model and data complexity increase.
Increasing model complexity, i.e., adding more parameters, introduces high-dimensional spaces where the 'curse of dimensionality' has significant effects not seen in low-dimensional spaces. For example, critical points in high-dimensional spaces can become saddle points rather than maxima or minima, posing challenges for optimisation problems. The decision to increase complexity should be carefully considered, as it may improve predictive power only if the sample size is large enough to accurately learn the new parameters.
In small datasets, noise can cause fluctuations that resemble genuine patterns. Simpler models, with fewer parameters, are unable to represent complex patterns and must ignore such fluctuations. In contrast, models with many parameters may inadvertently capture noise-generated patterns, believing them to be real information. This phenomenon, known as overfitting, results in excellent in-sample performance but poor generalisation to other samples. Addressing overfitting can be approached in two ways: using simpler models with fewer parameters or increasing the dataset size to reduce the likelihood of noise patterns. Often, simpler models yield better predictive performance by introducing a bit more bias but less dependence on the particular training dataset, striking a balance known as the bias-variance tradeoff. While having an infinite amount of training data would reduce bias and improve predictive performance, in practice, using simpler models is advisable given finite training sets.

Concerning the training of the model, the objective is to determine the parameters that minimise the cost function. Practically, this is typically achieved by employing gradient descent methods, which involve iteratively adjusting the parameters θ in the direction where the gradient of the cost function is large and negative. This approach allows finding parameters θ corresponding to a local minimum of the cost function. However, despite the simplicity of the concept, a major challenge arises from the complexity of most cost functions, which are often non-convex in the parameter space and contain numerous local minima in high-dimensional spaces. Furthermore, there is no direct access to the true

cost function, and it must be empirically estimated from the data.
In many cases, the cost function can be expressed as a sum over the n data points:

$$\mathcal{C}(\theta) = \sum_{i=1}^{n} c_i(\mathbf{x}_i, \theta) \tag{2.1}$$

Here, $c_i(\mathbf{x}_i, \theta)$ represents the cost function calculated for data point i. In the simplest form of gradient descent (GD), parameters are initially set to some value θ_0 at time step t. Then, after computing the cost function on the training data, they are updated at time $t+1$ as follows:

$$\begin{cases} \mathbf{v_t} = \eta_t \nabla_\theta \mathcal{C}(\theta_\mathbf{t}) \\ \theta_{\mathbf{t+1}} = \theta_\mathbf{t} - \mathbf{v_t} \end{cases} \tag{2.2}$$

Here, $\nabla_\theta \mathcal{C}(\theta_t)$ denotes the gradient of the cost function in the parameters space, and η_t is the learning rate, which controls the step length in the direction of the gradient at time t. The choice of η requires caution: a small value would necessitate many iterations to reach a local minimum, while excessively large values may cause overshooting the minimum, resulting in oscillations around it or divergence. It's important to note that this method only ensures finding local minima, and to escape them, stochasticity is required (which can be introduced, for instance, by computing the cost function over a smaller subset of data). Furthermore, it is sensitive to initial conditions and treats all directions uniformly. Depending on the specific case, various improvements can be applied to enhance convergence. Overall, the learning process aims to converge to a local minimum of the cost function. However, it can also be halted prematurely by monitoring the out-of-sample performance. If the error begins to increase, indicating potential overfitting, the process may be stopped.

2.2 Classification methods

Overall, the radiomic pipeline, as described in the previous chapter, incorporates ML in various stages, including segmentation and predictive model building. In both cases, classifiers are commonly employed. Segmentation often involves voxel-wise classification to distinguish background voxels from other class voxels within an image. Similarly, model building typically revolves around constructing binary or multi-class classifiers, such as discriminating between benign and malignant tumours or determining complete versus incomplete responses.
Classification problems entail discrete outcome variables representing different categories. The dependent variables, denoted as discrete numbers $y_i \in \mathbb{Z}$, range from $m = 0$ to $m = M-1$ where M signifies the total number of classes. With a dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ with $X \in \mathbb{R}^{n \times p}$, comprising n samples and p features, the goal is to train a model capable of predicting the output class for unseen data.

2.2.1 Perceptron

The simplest case involves dichotomous classification, distinguishing between just two classes. One of the most basic examples is the perceptron, a classifier based on a weighted linear combination of the p features plus an offset $s_i = \boldsymbol{x_i}^T \boldsymbol{w} + b_0 \equiv \mathbf{x}_i^T \mathbf{w}$ with $\mathbf{x}_i = (1, \boldsymbol{x}_i)$ and $\mathbf{w} = (b_0, \boldsymbol{w})$, where the output is mapped from the real axis to discrete values using the sign function as:

$$f(x_i) = \text{sign}(s_i) = \begin{cases} 1 & \text{if } s_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

In these simple cases, the training procedure resembles basic linear regression, with weights obtained through least squares. This kind of classification can be defined as hard classification since the outcome is strictly assigned to a well-defined class. However, softer definitions may be more suitable for noisy data, predicting instead the probability of belonging to a certain class.

2.2.2 Logistic regression

One of the most popular and representative cases of soft classification is logistic regression, which is based on the sigmoid function $\sigma(s) = \dfrac{1}{1+e^{-s}}$, where the probability of the point $\mathbf{x}_i$ belonging to category $y_i = \{0, 1\}$ is given by :

$$\begin{cases} P(y_i = 1|\mathbf{x_i}, \mathbf{w}) = \dfrac{1}{1+e^{-\mathbf{x_i^T w}}} \\ P(y_i = 0|\mathbf{x_i}, \mathbf{w}) = 1 - P(y_i = 1|\mathbf{x_i}, \mathbf{w}) \end{cases} \tag{2.4}$$

Here $\mathbf{w}$ corresponds to the weights that need to be learned during the training of the model. Figure 2.1 presents an example illustrating the sigmoid function's behaviour in logistic regression. The curves depict how changes in predictor variable weights affect the probability of a binary outcome. As mentioned earlier, the weights are learned by minimising a cost function, that in this case is derived from a maximum likelihood estimation (MLE), aiming to maximise the probability of observing the given data. Assuming a dataset $\mathcal{D} = \{(y_i, \mathbf{x_i})\}$ where the labels are binary values $y_i = \{0, 1\}$ and x_i are independent of each other, one can write the likelihood of observing the data as:

$$\mathcal{L}(\mathbf{w}) = P(\mathcal{D}|\mathbf{w}) = \prod_{i=1}^{n} \left[\sigma(\mathbf{x_i^T w})\right]^{y_i} \left[1 - \sigma(\mathbf{x_i^T w})\right]^{1-y_i} \tag{2.5}$$

and subsequently the log-likelihood:

$$\log \mathcal{L}(\mathbf{w}) = \log P(\mathcal{D}|\mathbf{w}) = \sum_{i=1}^{n} y_i \sigma(\mathbf{x_i^T w}) + (1 - y_i)[1 - \sigma(\mathbf{x_i^T w})] \tag{2.6}$$

The final set of chosen parameters is the one that maximises $\log \mathcal{L}(\mathbf{w})$, $\hat{\mathbf{w}} = \text{argmax}_\theta \log \mathcal{L}(\mathbf{w})$.
To use it in a minimisation problem instead, one can simply take the negative log-likelihood as the cost function:

$$\mathcal{C}(\mathbf{w}) = -\log \mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} -y_i \sigma(\mathbf{x_i^T w}) - (1 - y_i)[1 - \sigma(\mathbf{x_i^T w})] \tag{2.7}$$

This cost function is usually known as cross-entropy and is one of the most widely used in different models. One of the advantages of cross-entropy is that it is a convex function of the weights $\mathbf{w}$, so local and global minimisation coincide. The minimisation problem takes the transcendental form:

$$0 = \nabla \mathcal{C}(\mathbf{w}) = \sum_{i=1}^{n} \left[\sigma(\mathbf{x_i^T w} - y_i)\right] \mathbf{x_i} \tag{2.8}$$

and requires numerical methods to be solved.

2.2.3 Multinomial logistic regression

When considering a problem with multiple classes, the approach is to extend the sigmoid model to handle multi-class outcomes. Utilising a one-hot encoding for the outcome and having M classes,

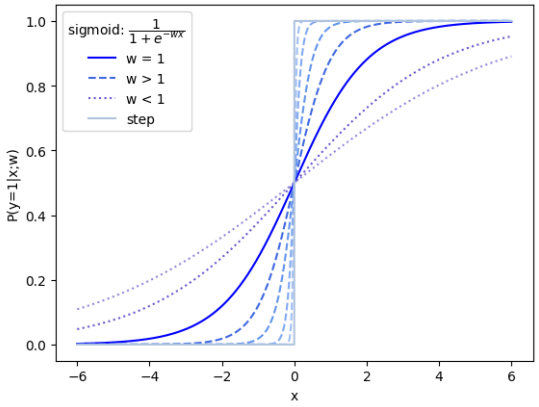


FIGURE 2.1: Example of the sigmoid function for the logistic model as a function of the weights. As the weights increase, the sigmoid tends more and more to the limit of the step function of the hard classifier. Beyond a certain threshold, typically 0.5, the class probability increases significantly, indicating greater confidence in the classification decision.

where $y_i \in \mathbb{Z}_2^M$ with the j-th component of y_i being 1 while all others are 0 to indicate class j, the probability of $\mathbf{x_i}$ belonging to class m' among the M classes can be expressed as:

$$P(y_{im'} = 1|\mathbf{x_i}, \{\mathbf{w}_k\}_{k=0}^{M-1}) = \frac{e^{-\mathbf{x_i^T w_{m'}}}}{\sum_{m=0}^{M-1} e^{-\mathbf{x_i^T w_m}}} \tag{2.9}$$

Here, $y_{im'}$ represents the m'-th component of y_i. This expression is commonly known as the Softmax function. Using a maximum likelihood estimation process similarly, the cost function can be derived as:

$$\mathcal{C}(\mathbf{w}) = -\sum_{i=1}^{N}\sum_{m=0}^{M-1} y_{im} \log P(y_{im} = 1|\mathbf{x_i}, \mathbf{w_m}) + (1 - y_{im}) \log\left(1 - P(y_{im} = 1|\mathbf{x_i}, \mathbf{w_m})\right) \tag{2.10}$$

For $M = 1$, this formulation reverts back to the binary cross-entropy defined previously.

2.2.4 Ensemble models

Ensemble models are among the most powerful techniques in machine learning, leveraging combinations of multiple models to enhance predictive performance. Common ensemble methods include Random Forest, boosted gradient trees (such as XGBoost), and more sophisticated models like neural networks. The fundamental concept behind ensembles is to harness the "wisdom of the crowds" by aggregating predictions from diverse models. However, while this approach offers advantages, it can also exacerbate the deficiencies of individual predictors and correlations among them. From the perspective of the bias-variance tradeoff, correlations among ensemble models can pose challenges. With a fixed ensemble size, correlated models may not effectively reduce variance, potentially leading to an increase in bias due to correlated errors. Ideally, using larger ensembles of uncorrelated models can substantially suppress variance. In scenarios with small training sets, various models may yield similar performance on data. Averaging predictions from multiple models mitigates the risk of selecting inappropriate models. Nonetheless, the increased representational power of ensembles comes with the drawback of more parameters. The overarching strategy is to introduce as much randomness as possible in ensemble construction to reduce correlations and prevent bias inflation. Such an approach is

particularly beneficial for high-variance scenarios but may not be effective for high-bias cases. Various approaches to model ensembling can be employed.

Bagging

Bagging, or Bootstrap Aggregating, is a technique used when the dataset $\mathcal{D}$ is sufficiently large. It involves splitting the dataset into N smaller subsets $\mathcal{D}_1, ..., \mathcal{D}_N$ and training a predictor on each subset. The final predictor is then an aggregate of all the individual results.
For classification tasks with M classes, the final predictor employs a majority vote of all the predictors. Mathematically, this can be represented as:

$$f(\mathbf{X}) = \text{argmax}_j \sum_{i=1}^{N} I\left[f_{\mathcal{D}_i}(\mathbf{X}) = j\right], \quad j \in \{0, ..., M-1\} \tag{2.11}$$

Here, I is the indicator function, equal to 1 when the predictor $f_{\mathcal{D}_i}$ predicts response j. Similarly, for continuous regression tasks, the overall result is the average of the individual predictors. Other variations of this procedure, suitable for smaller datasets, include empirical bootstrapping. This involves sampling with replacement to create new subsets from the original dataset. While useful for unstable predictors as it reduces variance, it can increase bias. Conversely, if the procedure yields stable predictions, bootstrapping may not enhance the model.

Boosting

In boosting, unlike bagging where all models vote with equal weight, each weak classifier is associated with a weight α_k. The total classifier is then a weighted sum of the individual classifiers, given by:

$$f(\mathbf{X}) = \sum_{i=1}^{N} \alpha_i f_i(\mathbf{X}) \quad \text{where } \sum_i \alpha_i = 1 \tag{2.12}$$

In this formulation, α_i represents the weight assigned to the i-th weak classifier $f_i(\mathbf{X})$. These weights are typically determined during the training process, where classifiers that perform better are assigned higher weights. The final prediction is then made based on the weighted combination of the individual classifier predictions, with the weights ensuring that more accurate classifiers have a greater influence on the overall prediction.

Random Forests

Random Forests (RF) is among the most widely used ensemble algorithms for classification tasks. It relies on a collection of tree-based classifiers, typically decision trees, to make predictions. Each decision tree in a Random Forest operates by posing a series of specific questions to recursively partition the data into different categories. Visualising it as a tree structure, each branch corresponds to a question about a particular feature, such as whether a certain feature j exceeds a specific threshold t_j. These questions divide the data into subgroups, and this process continues until reaching terminal nodes, also known as leaves, which provide the final outcome or prediction. A schematic example of a single decision tree is depicted on the left side of Figure 2.2. The primary objective is to construct trees that effectively partition the classes. However, as decision trees grow deeper (i.e., have more branches), they tend to become overly complex, leading to overfitting. Since the partitioning decisions are primarily based on the input data, individual decision trees are susceptible to noise, resulting in high variance. To mitigate these issues, RF employ an ensemble of trees with lower complexity. This ensemble approach can involve methods such as bagging, where trees are trained on different subsets of the

training data, or feature bagging, where trees are constructed using random subsets of features for each split. By diversifying the trees in the ensemble, correlations are reduced, consequently lowering the overall variance. On the right side of Figure 2.2, an example of a tree ensemble is illustrated, where multiple individual trees contribute to a final decision through majority voting. Additional enhancements have been made by utilising gradient boosting to construct tree ensembles, as seen in algorithms like XGBoost. However, it's important to note that increasing the number of parameters in these methods may increase the risk of overfitting, especially when dealing with small datasets [63].

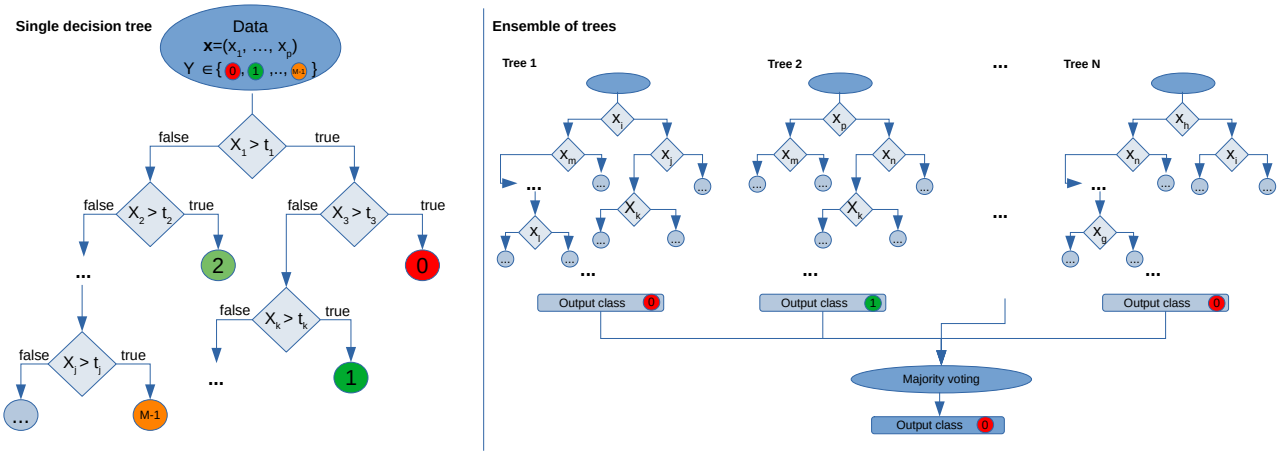


FIGURE 2.2: Illustration of a Random Forests model: On the left, a single decision tree sequentially queries features to classify instances. On the right, an ensemble of decision trees aggregates predictions through majority voting to yield the final classification outcome.

2.2.5 Support vector machines

Support Vector Machines (SVM) is a versatile and widely-used model in machine learning for classification problems. It stands out for its ability to perform well with minimal features, robustness to errors, and computational efficiency compared to more complex models like neural networks [28]. SVM can be applied to both classification and regression tasks.
In classification scenarios, SVM handle different data types differently. When data are linearly separable, SVM aims to find the optimal hyperplane that effectively separates the data while generalising well to new instances. In a dataset with samples x_i where $i = 1, ..., n$ and two classes, the linear hyperplane is defined as

$$w^T x + b = 0 \tag{2.13}$$

where w represents the coefficient vector and b is the bias term. The optimisation process for finding the optimal hyperplane involves minimising the classification error while maximising the distance from the hyperplane to the closest data points of each class. By defining the margins of separation for the two classes as $w^T x + b \geq 1$ for class 1 (with label $y_i = 1$) and $w^T x + b \leq -1$ for the other (with label $y_i = -1$), maximising the margin distance:

$$d(w, b; x) = \frac{(w^T x + b - 1) - (w^T x + b + 1)}{||w||} = \frac{2}{||w||} \tag{2.14}$$

becomes equivalent to minimising the norm of the vector w. Thus, the problem boils down to minimising the convex function:

$$\mathcal{C}_{lin} = \frac{1}{2} w w^T \tag{2.15}$$

Once the parameters w and b are determined, the optimal hyperplane is defined.
In cases where the data are not separable with an hard margin, SVM can still be employed by incorporating a penalty term into the cost function to account for classification errors. The cost function in such cases becomes:

$$\mathcal{C}_{ns} = \frac{1}{2} w w^T + c \sum_{i=1}^{N} \xi_i \tag{2.16}$$

where ξ_i represents the distances between misclassified data points and the margin, and c is a trade-off parameter that balances margin maximisation and error minimisation. An illustration of a hyperplane and margin in binary classification scenarios is presented in Figure 2.3. One of the most advantageous properties of SVM is their ability to find optimal hyperplanes even for data that is not linearly separable. In such cases, SVM employs a technique to map the input data to a higher-dimensional feature space where linear separation becomes feasible. This mapping $\phi : \mathcal{X} \subset \mathbb{R}^p \to \mathcal{V} \subset \mathbb{R}^q$ transforms the original input data x into a new feature space representation $\phi(x)$. Crucially, this technique relies solely on the knowledge of inner products, without the need of knowing the precise functional form of ϕ, allowing the definition of a kernel function $K(x_i, x_j) = \langle x_i, x_j \rangle_{\mathcal{V}}$ associated to the inner product. The general equation for the hyperplane of separation can then be expressed as $d(x) = \sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b$, where α_i are Lagrange multipliers, eliminating the need to determine the weighting parameters w. SVM offers flexibility in choosing different kernels, including radial, polynomial, sigmoid, and Gaussian radial basis functions, among others. However, more sophisticated kernels introduce additional parameters, leading to increased model complexity. For smaller and simpler datasets, linear kernels are often recommended to avoid unnecessary complexity. It is important to note that data should be rescaled before employing SVM, as features with very large values may exert undue influence on the model. Additionally, it is crucial to consider that only points in close proximity to the decision boundary carry significant weight, while those farther away have less impact. Furthermore, meticulous tuning of the model's parameters is essential to attain optimal performance.

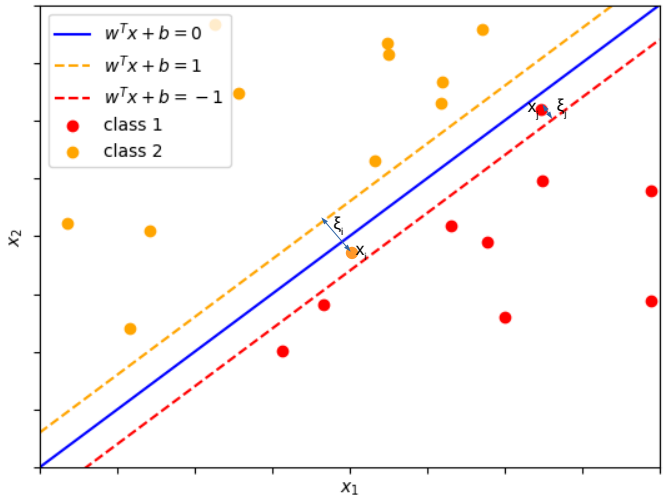


FIGURE 2.3: Example of separating hyperplane and margins in the case of binary classification. Distances ξ_i between misclassified data points x_i and the corresponding true class margin provides the penalisation terms for SVM optimisation.

2.3 Segmentation techniques

In the segmentation stage, various methods are employed depending on the complexity of the task at hand. For straightforward tasks, such as segmenting whole organs, simpler techniques not requiring machine learning are often used. These include image thresholding and atlas-based segmentation. However, when the object of interest is a small lesion or portion of tissue, methods involving machine learning are more effective. These may include clustering techniques or more sophisticated deep learning algorithms.

2.3.1 Image thresholding

Thresholding is one of the simplest segmentation methods, where segmentation is determined based on a membership function:

$$f(x) = \begin{cases} 1, & \text{if } I_x > I_{thr} \\ 0, & \text{if } I_x \leq I_{thr} \end{cases} \tag{2.17}$$

Here, x represents a generic pixel or voxel of the image, I_x corresponds to its intensity, and I_{thr} is the threshold value. Multiple thresholds can be defined depending on the tissue to be identified. Various methods exist for defining the threshold intensity, such as global fixed thresholds, local thresholds based on the pixel's neighbourhood, or adaptive thresholds that are specific functions of x.
One commonly used method is Otsu thresholding, which exploits the statistics of histogram of grey levels. It assumes that well-separated classes have distinct grey levels, and the optimal threshold should minimise within-class variance or equivalently maximise between-class variance [70].
For an image with n pixels or voxels and N_g grey levels, one can analyse the normalised histogram of intensity values $\mathbf{H} = \{p_1, ..., p_{N_g}\}$ with $p_i = \dfrac{f_i}{n}$ and f_i the occurrences of grey level i.
By separating the image into two classes c_0 and c_1 at a grey value k (where $k \in [1, N_g]$), class probabilities are determined as:

$$\omega_0 = \Pr(C_0) = \sum_{i=1}^{k} p_i, \quad \omega_1 = \Pr(C_1) = 1 - \Pr(C_0) \tag{2.18}$$

The class mean levels, along with class variances can thus be defined as:

$$\mu_0(k) = \sum_{i=1}^{k} i\Pr(i|C_0) = \sum_{i=1}^{k} i\frac{ip_i}{\omega_0}, \quad \mu_1(k) = \sum_{i=k+1}^{N_g} i\Pr(i|C_1) = \sum_{i=1}^{k} i\frac{ip_i}{\omega_1} \tag{2.19}$$

$$\sigma_0^2(k) = \sum_{i=1}^{k}(i-\mu_0)^2\Pr(i|C_0) = \sum_{i=1}^{k}(i-\mu_0)^2\frac{ip_i}{\omega_0}, \quad \sigma_1^2(k) = \sum_{i=1}^{k}(i-\mu_1)^2\Pr(i|C_1) = \sum_{i=1}^{k}(i-\mu_1)^2\frac{ip_i}{\omega_1} \tag{2.20}$$

The within-class variance is then defined as the weighted average of the variances of the two classes

$$\sigma_W^2 = \omega_0\sigma_0^2 + \omega_1\sigma_1^2 \tag{2.21}$$

and the between-class variance is the variance of the class means around the combined mean

$$\sigma_B^2 = \omega_0(\mu_0 - \mu_T)^2 + \omega_1(\mu_1 - \mu_T)^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2, \quad \text{with } \mu_T = \mu(N_g) \tag{2.22}$$

It is important to note that $\sigma_B^2(k) = \sigma^2(N_g) - \sigma_W^2(k)$ thus the optimal threshold k_{opt} can be chosen as the one that minimises the within-class variance or equivalently the one that maximises the between-class variance:

$$k_{opt} = \text{argmax}_k(\sigma_B^2(k)) \tag{2.23}$$

While image thresholding is fast and does not require prior knowledge, it may yield approximate results for images with multiple peaks in the histogram. Additionally, assuming the same variance for background and object segments may not be robust to noise or distinguish objects with similar intensities.

2.3.2 Atlas based segmentation

One of the methods that leverages prior knowledge in segmentation involves utilising a reference image with a pre-existing full segmentation, referred to as an atlas. Much like geographical atlases, these atlases provide descriptions of the object of segmentation, encompassing its shape and texture. Atlases are grounded in the consistent anatomical structures of the same type and can be employed to characterise groups of individuals [6]. Deterministic atlases, derived from a single subject, serve as representatives of average size, shape, and intensity. However, a single subject may not fully encapsulate the population, necessitating the creation of statistical atlases. Statistical atlases are based on images from a larger group, capturing the variability inherent in anatomical structures. These atlases can be constructed by registering, normalising, and averaging voxel-wise images to generate probabilistic maps. Similarly, atlases tailored to specific diseases can be developed using images of affected individuals. Atlas-based segmentation often frames the segmentation task as an image registration problem, where the intensities of the atlas and the target image are aligned. This process involves optimising the alignment by defining allowable transformations and selecting appropriate similarity measures. Atlas-based segmentation techniques, although informative, can pose computational challenges and may lack representation, particularly in the context of diseases or anatomical variations.

2.3.3 Clustering

Before the emergence of sophisticated neural networks, clustering stood out as one of the most widely used machine learning methods for unsupervised tasks. Clustering, along with other unsupervised techniques, is designed to reveal underlying structures within unlabelled datasets. Its primary objective is to organise data points into clusters, leveraging measures of distance or similarity to identify patterns or structures shared among points. One of the most common clustering algorithms is the K-means algorithm, which involves grouping data into K clusters starting from K centroids. Also in this case, learning can be viewed as an optimisation problem.
Considering an image with n pixels or voxels $\{\mathbf{x}_j\}_{j=1}^n$, and given a fixed integer K corresponding to the number of clusters, the objective is to find the cluster means $\{\boldsymbol{\mu}_k\}_{k=1}^K$ and assign each element of the image to a cluster by minimising the sum of the squared distances between points and cluster centres. This can be expressed as minimising the following cost function:

$$\mathcal{C}(\{\mathbf{x}, \boldsymbol{\mu}\}) = \sum_{k=1}^{K} \sum_{j=1}^{n} a_{jk}(\mathbf{x}_{jk} - \boldsymbol{\mu}_k)^2 \tag{2.24}$$

Here, a_{jk} is the cluster assignment function defined as: $a_{jk} = \begin{cases} 1, & \text{if } \mathbf{x}_j \in \text{ cluster } k \\ 0, & \text{otherwise} \end{cases}$.
This objective aims to minimise the variance within each cluster. In practice, the number of clusters is fixed, and at the first stage, the centroids are randomly chosen, and points are assigned to the

closest centroid. The algorithm then follows an expectation-maximisation procedure, with an initial step (expectation) where the cluster assignments $\{a_{jk}\}$ are fixed, and the cost function $\mathcal{C}$ is minimised with respect to the centroids, resulting in new centroids given by:

$$\boldsymbol{\mu} = \frac{1}{N_k} \sum_{j=1}^{n} a_{jk} \mathbf{x}_j, \quad \text{with} \quad N_k = \sum_{j=1}^{n} a_{jk} \tag{2.25}$$

where N_k is the number of elements assigned to cluster k. The subsequent maximisation step involves fixing the new cluster means and minimising $\mathcal{C}$ with respect to the cluster assignments, resulting in:

$$a_{jk} = \begin{cases} 1, & \text{if } k = \text{argmin}_{k'}(\mathbf{x}_n - \boldsymbol{\mu}_{k'})^2 \\ 0, & \text{otherwise} \end{cases} \tag{2.26}$$

The algorithm iteratively runs these two steps until a certain convergence criterion is met, such as requiring that the difference in the cost function or centroids is less than a specified threshold. Although convergence to a local minimum is always guaranteed, the cost function $\mathcal{C}$ is typically not convex. Therefore, the choice of different initialisation can strongly influence the local minimum found through expectation maximisation. Thus, using various initialisation or manually adjusting the values may be necessary to address delicate segmentation tasks.
However, it's essential to note that this model's primary assumption is that the clusters have similar variances. When this condition is not met, issues may arise. To address this limitation, more flexible models allowing for different variances, such as Gaussian Mixture Models (GMM), have been developed. Furthermore, K-means can be considered a hard clustering method because each element can only belong to a single cluster. In contrast, softer versions with the possibility of membership in multiple clusters have been developed. One widely used method is Fuzzy C-means (FCM), where the membership function is derived from a matrix indicating the degree to which an element x_i belongs to a certain cluster c_j. In FCM, cluster centroids are in fact determined by a weighted average based on the degree of belonging to the respective cluster.

2.3.4 Deep Neural Networks

With the advancement in computational capabilities, particularly the development of specialised processors like Graphical Processing Units (GPUs), deep neural networks (DNN) have emerged as one of the most powerful and widely used learning techniques. The field of neural networks is extensive, encompassing various types such as general-purpose networks for supervised learning, networks for unsupervised learning like Restricted Boltzmann Machines, recurrent neural networks for sequential data, and those specifically designed for image processing, such as Convolutional Neural Networks (CNNs). In domains like medical image segmentation, CNNs have shown promising results, although much of the progress is empirical and heuristic, driven by rapid advancements in technology.
Before exploring CNNs, let's review the basics of general neural networks. Neural networks are nonlinear models that extend common supervised learning methods like linear and logistic regression. As the name suggests, neural networks are based on the concept of the neuron unit. In practice, a neuron i is fed with an input vector $\boldsymbol{x} \in \mathbb{R}^p$, where p represents the number of features. The neuron then produces a scalar output $a_i(\boldsymbol{x})$. The entire network consists of multiple layers as depicted in Figure 2.4 on the right, each containing numerous neurons. The output of one layer serves as the input to the next layer, culminating in the final layer that produces the ultimate output. Intermediate layers are often referred to as hidden layers. The function a_i that generates the output scalar is a customised choice depending on the specific network architecture. It typically comprises two main components, as illustrated on the left in Figure 2.4:

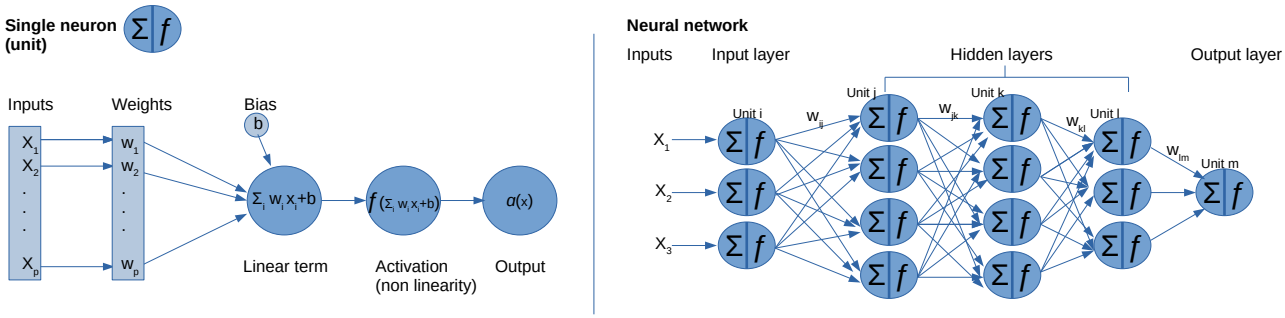


FIGURE 2.4: On the left is an example illustrating the working principle of a single neuron, including the linear term and the non-linear activation term. On the right is an example of a fully connected neural network with 3 units in the input layer and 3 hidden layers, consisting of 4 and 3 units respectively. The output of each layer is processed as new input for the subsequent layers

- **linear operation:** calculating weights for the various inputs of the neuron i, it is usually performed as a dot product between the input and neuron-specific weights $\boldsymbol{w}^{(i)} = (w_1^{(i)}, ..., w_p^{(i)})$, followed by the addition of a bias $b^{(i)}$ specific to the neuron

$$z^{(i)} = \boldsymbol{w}^{(i)} \cdot \boldsymbol{x} + b^{(i)} = \mathbf{x}^T \cdot \mathbf{w}^{(i)} \quad \text{with } \mathbf{x} = (1, \boldsymbol{x}), \quad \mathbf{w}^{(i)} = (b^{(i)}, \boldsymbol{w}^{(i)}) \tag{2.27}$$

- **activation:** often consisting of a non-linear transformation f_i applied to the weighted input,

$$a_i(\mathbf{x}) = f_i(z^{(i)}) \tag{2.28}$$

  which is usually common for all neurons $f_i \equiv f$.

These components allow neural networks to capture complex relationships and non-linearities in data, enabling them to perform well on various tasks. Like other methods, neural networks are trained by optimising the parameters for the weights $\boldsymbol{w^{(i)}}$ and the bias $b^{(i)}$. This optimisation is typically achieved using gradient descent-based methods, which involve computing the derivatives of the input-output function with respect to the weights and biases. Therefore, the choice of the appropriate nonlinearity function is crucial.
Various functions can be used for f, some examples are presented in Figure 2.5 with historically well-known options including step functions, sigmoids, and hyperbolic tangents. However, step functions are not suitable due to their discontinuous derivatives, while sigmoids and hyperbolic tangents suffer from saturating behaviour, resulting in vanishing derivatives for large inputs. Vanishing gradients are a common issue with saturating nonlinearities.
To address this problem, alternative functions with non-saturating behaviour have been proposed. These include Rectified Linear Units (ReLUs), Leaky Rectified Linear Units (Leaky ReLUs), and Exponential Linear Units (ELUs). These functions have the advantage of gradients that do not vanish, making them more suitable for deep neural networks.

In terms of network architecture, the simplest form of a neural network is the feedforward network (FNN), which is characterized by hierarchical layers. The network begins with an input layer, which generates an output that serves as the input for subsequent hidden layers, and so on, until reaching the final output layer. Typically, the output layer consists of a classifier, such as logistic or softmax for categorical data, or linear regression for continuous outputs. In essence, the network takes an input

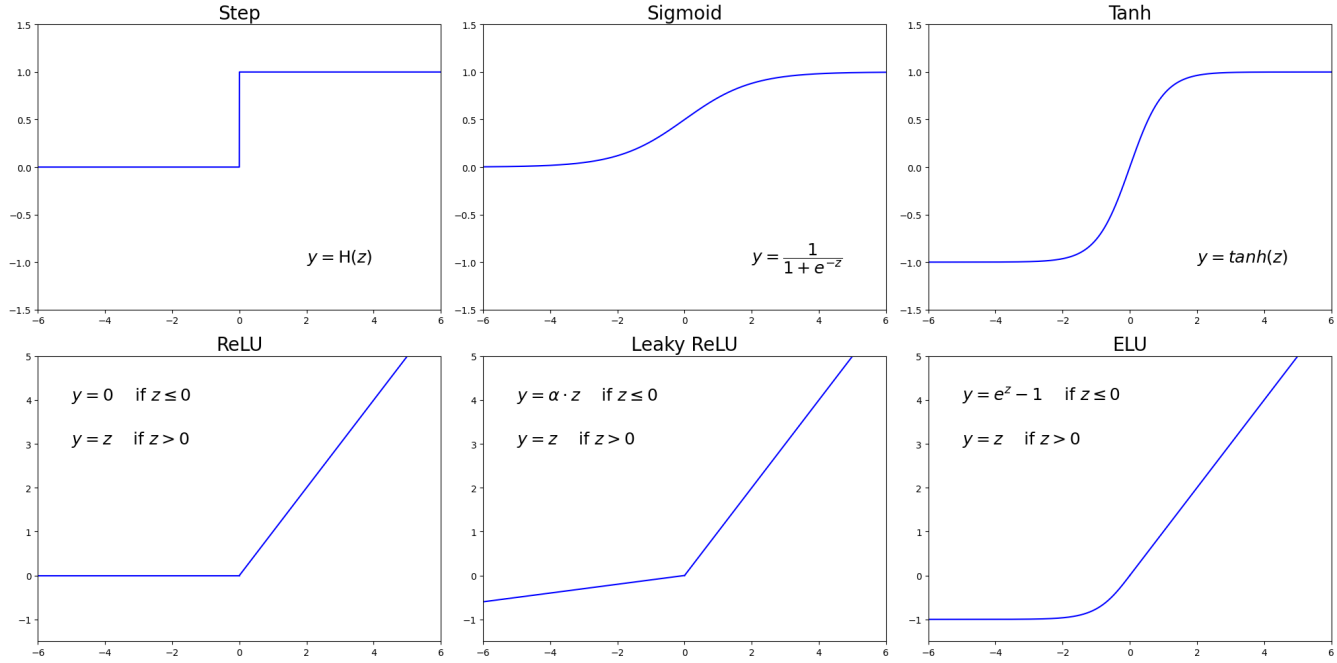


FIGURE 2.5: Examples of activation functions. At the top step, sigmoid, and hyperbolic tangent functions, which exhibit a saturating behaviour implying vanishing gradients. Below are alternative definitions, ReLU, leaky ReLU and ELU, aimed at overcoming vanishing gradients issues.

vector $\boldsymbol{x}$ and produces an output y that depends on all the layers and their respective weights and biases. Increasing the number of layers enhances the network's representational power, allowing it to approximate arbitrary functions. However, the number of layers chosen depends on factors such as the specific problem, the available data, and the desired complexity. Often, a larger number of parameters is retained to prevent underfitting. There is ongoing debate regarding whether it is more effective to train deeper networks or shallower but wider networks [63]. However, in certain cases, connections between layers can be skipped, allowing the output of one of the top layers to propagate directly to deeper layers, bypassing intermediate layers. This technique, known as skip connections, is commonly employed in image processing tasks to improve performance.

During the training phase of a deep neural network, similar to previous supervised methods, a suitable loss function is defined, and gradient descent is employed to find optimal parameters. However, the presence of a large number of parameters and layers in DNNs significantly increases computational cost. To address this challenge, the problem is often reformulated in a more convenient manner.

For categorical data, the last layer typically employs a sigmoid or softmax function, and the most common loss function is cross-entropy, as discussed in the previous section on classifiers. Computing the gradient of the cost function for updating parameters via gradient descent can be computationally intensive, as it requires calculating gradients for each parameter at every step. Therefore, a brute force approach is often impractical. To overcome this challenge, a technique called backpropagation, which leverages the network's layer structure, has been developed.

In a network with L layers ($l = 1, ..., L$), where w^l_{jk} represents the weight connecting neuron k in layer $l-1$ to neuron j in layer l, and b^l_j denotes the corresponding bias, the activation a^l_j of neuron j in layer l is related to activation of previous layers $l-1$ as follows:

$$a^l_j = f\left(\sum_k w^l_{jk} a^{l-1}_k + b^l_j\right) \equiv f(z^l_j) \tag{2.29}$$

where f represents the activation function. Backpropagation allows efficient computation of gradients by propagating errors backwards through the network, enabling parameter updates that minimise the loss function. The error of neuron j in layer l, denoted as Δ_j^l, can be defined as the change in the cost function with respect to the weighted input z_j^l:

$$\Delta_j^l = \frac{\partial \mathcal{C}}{\partial z_j^l} = \frac{\partial \mathcal{C}}{\partial a_j^l} f'(z_j^l) \tag{2.30}$$

where f' represents the derivative of the non-linearity with respect to its input. Similarly, the error can be interpreted as a variation with respect to the bias:

$$\Delta_j^l = \frac{\partial \mathcal{C}}{\partial z_j^l} = \frac{\partial \mathcal{C}}{\partial b_j^l} \frac{\partial b_j^l}{\partial z_j^l} = \frac{\partial \mathcal{C}}{\partial b_j^l} \tag{2.31}$$

Utilising the chain rule, the error in layer l can also be expressed as a function of the activation of layer $l+1$:

$$\Delta_j^l = \frac{\partial \mathcal{C}}{\partial z_j^l} = \sum_k \frac{\partial \mathcal{C}}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l} = \left(\sum_k \Delta_k^{l+1} w_{kj}^{l+1} \right) f'(z_j^l) \tag{2.32}$$

Finally, differentiating with respect to the weights one obtains:

$$\frac{\partial \mathcal{C}}{\partial w_{jk}^l} = \frac{\partial \mathcal{C}}{\partial z_j^l} \frac{\partial z_j^l}{\partial w_{jk}^l} = \Delta_j^l a_k^{l-1} \tag{2.33}$$

The backpropagation algorithm can be summarised as follows [63]:

1. Activation of the input layer a_j^1 $(l = 1)$.
2. Feedforward: calculation of the weighted inputs z^l and outputs a^l for all subsequent layers $(l = 2, ..., L)$ following the model architecture.
3. Calculation of the error Δ_j^L on the output layer $(l = L)$ using Equation 2.30, utilising information about the derivatives of the cost function and activation function.
4. Backpropagation of the error for all the other layers $(\Delta_j^l, l = L-1, ..., 1)$ using Equation 2.32.
5. Calculation of all the gradients with respect to weights and biases using Equations 2.31 and 2.33.

In summary, backpropagation involves a forward pass from input layer to output layer, calculating weighted inputs and activations of all neurons, and a backward pass, where error is backpropagated from the last layer to the input layer, and the errors are used to estimate all the gradients. Despite its convenience, the calculation of gradients is still computationally extensive for large networks. Moreover, different issues might occur in backpropagation, including vanishing or exploding gradients. Strategies such as using non-saturating activation functions, good weight initializations, and gradient clipping for large values are used to mitigate these issues.

2.3.5 Convolutional Neural Networks

Convolutional Neural Networks take advantage of properties such as locality and translational invariance, which are common in physical systems. These networks are designed to be translationally invariant and respect the locality of input data, making them particularly useful for processing images. A typical CNN consists of a succession of two basic layers:

- **Convolutional layers:** these layers apply convolution operations using a set of filters. Filters serve as receptive fields for local areas, capturing small spatial patches of the image and detecting specific features, such as vertical and horizontal edges. A filter is represented by a 2D or 3D array of learnable weights, which is applied to specific locations in the input data, producing an output. The filter then shifts by a certain amount known as the stride before being applied again, repeating the process until the entire input data is covered. The result of this convolution operation is a feature map. During the convolution operation, the weights of the filter remain fixed, meaning that all the neurons corresponding to that filter share the same parameters.

- **Pooling layers:** these layers downsample the input while preserving local structure. Like convolutions, filters are applied in pooling, but without containing any learnable parameters. The goal is to coarsen the image. In this step, a fixed operation is performed by the kernel. Common pooling operations include max pooling, which selects the maximum value of a receptive field, and average pooling, which calculates the average.

Each filter produces its own feature map, encoding the presence or absence of certain features at various locations of the input data. High values indicate the presence of the searched feature, while low values indicate its absence. Then the output feature maps of convolutional and pooling layers typically go through a nonlinearity, usually a ReLU.
After several convolutional and pooling layers, CNN often include fully connected layers and a classifier, such as softmax or sigmoid, similar to feedforward neural networks. Parameters in CNN are shared among neurons corresponding to the same filter applied in different locations of the input, reducing the computational costs required by backpropagation.
The hierarchical structure of CNN is well-suited for capturing abstract features, making them effective for tasks such as classification and segmentation. Lower layers encode low-level features like edge detection, while deeper layers combine these features to represent higher-level abstract features.

2.4 Feature selection techniques

Among the various preprocessing strategies, one of the most commonly used techniques for high-dimensional data in various fields is the reduction of dimensionality in the feature space. For instance, in medical imaging, a set of 169 standard radiomic features has been identified by IBSI [123], and the application of filters to images can potentially increase the number of features to several thousands.
However, medical datasets typically consist of only a few hundred cases. Therefore, before developing a model, it is crucial to address the typical challenges of "large p small n problems," where p represents the number of features and n represents the number of samples.
The main concern is the curse of dimensionality, a phenomenon where data sparsity in high-dimensional spaces leads to various negative effects. This includes increased computational complexity, as each additional dimension escalates the computational requirements for processing and analysing the data. Additionally, data sparsity makes it challenging to discern meaningful patterns or relationships within the data, hindering the ability to extract valuable insights. Moreover, the heightened risk of overfitting arises, wherein machine learning models may inadvertently capture noise or random fluctuations in the training data rather than the underlying patterns. Consequently, the model's ability to generalise to new cases is compromised, as it may struggle to accurately predict outcomes on unseen data. Furthermore, the difficulty in visualising and interpreting results exacerbates the problem, making it arduous for analysts to gain actionable insights from the data.
While one approach to mitigate this issue is to increase the dataset's size to balance n and p, this is often impractical. Hence, reducing the dimensionality of the feature space through feature selection or feature engineering methods is a common solution [72]. Additionally, reducing the feature space offers

advantages in terms of interpretability, storage requirements and computational costs.
In this context, reducing the number of dependent variables, i.e., the features used to train the models, is essential. This process aims to eliminate redundant and irrelevant features to improve model performance and generalisability [116]. The goal is to retain essential information while minimizing the number of predictors. Feature selection is widely used in ML and various methods have been developed, each with its own advantages and drawbacks.
In radiomics, there are typically no predefined assumptions regarding the clinical significance of specific radiomic features over others. Consequently, manually selecting a subset of informative features beforehand is not feasible. It's essential to note that not all extracted features contain useful information for the intended task. Therefore, feature selection is a necessary step that must be tailored to the specific research question. Feature sets can be categorised into several types [116]:

- **relevant features:** directly related to the research question and contribute to model performance;
- **irrelevant features:** not correlated with the desired outcome and not impacting the learning process;
- **weakly correlated but non-redundant features:** exhibiting discrete intercorrelation with other features, but still encoding valuable information;
- **highly correlated and redundant features:** containing similar information that can be inferred from other features, offering no additional value to the model.

The presence of irrelevant or highly correlated features can adversely affect models predictive capabilities. Multicollinearity can in fact overemphasise the importance of certain features while neglecting others, leading to poor generalisation on new data.
Various methods aim to retain a minimal subset of features while effectively predicting the target variable, thereby enhancing model accuracy without sacrificing it (to avoid underfitting or overfitting). However, there is currently no consensus on the best feature selection method for radiomics, leading to a certain degree of arbitrariness in this step of the process.

Feature selection methods can be categorised as supervised, unsupervised, or semi-supervised, depending on the availability of outcome information. Supervised approaches are commonly used for classification and regression problems, focusing on selecting discriminative features or those that best approximate the target variable. Unsupervised methods evaluate feature importance differently, catering to cases where data are not labelled. Hybrid semi-supervised methods can also be employed to leverage both labelled and unlabelled data.
Another approach to classifying feature selection methods involves their relationship with the learning algorithm. This classification reveals three main classes of feature selection methods: filter, wrapper and embedded methods.
From an alternative viewpoint, feature selection methods can be classified according to various methods for ranking features, including similarity based methods, information-based methods, sparse learning methods and statistical based methods [53].

2.4.1 Filter methods

Filter methods are independent of any specific learning algorithm; they rely on intrinsic data characteristics to assess feature importance. Evaluation metrics are directly derived from the data without feedback from the ML model. The advantage of these methods lies in their easy applicability to very large datasets and feature spaces, with no bias towards a particular model. Filter methods can score

each feature independently or account for relationships between features, leading to univariate and multivariate methods, respectively.
Examples of filters include: removing variables with a high percentage of missing data, eliminating features with zero or near-zero variance, assessing correlation with other features or with the outcome using metrics like Pearson correlation coefficient, Spearman's rank correlation coefficient, or Kendall's tau rank correlation coefficient, testing for significant separation between feature group averages using ANOVA or chi-squared tests, considering mutual information and redundancy, analyzing intraclass and interclass variances using Fisher Score and many others.

2.4.2 Wrapper methods

Wrapper methods depend on a chosen ML algorithm to evaluate the feature subset, utilising model performances to assess feature importance. These methods involve searching for the optimal feature subset and evaluating selected features based on accuracy or other metrics on the validation dataset. The search strategy to generate the subset can vary:

- **Complete search**: iterates over all possible combinations of subsets to select the best-scoring one. It is computationally expensive.
- **Heuristic search**: avoids the iteration over all combinations, includes Sequential Forward Selection (SFS), starting from an empty set and adding a new feature; Sequential Backward Selection (SBS), which begins with the full set of features and removes one at a time; and Bidirectional search, which combines SFS and SBS until convergence to a common subset. In this case it is guaranteed only to converge on a local optimum, reducing computational costs.
- **Random search**: evaluates random subsets of subsets, but may lead to non-reproducible results and is limited by computational costs.

Nevertheless, this search-based framework is computationally expensive and time-consuming because it requires a full training of the model to evaluate any subset. With a full search space for an initial set of p features being $2^p - 1$, these methods become impractical for very large p. While they offer more accurate predictions by considering the bias of a specific algorithm, their use is limited due to computational costs. Additionally, incomplete searches often involve random factors, diminishing result reproducibility.

2.4.3 Embedded methods

Embedded methods combine the advantages of both filter and wrapper classes. They are computationally efficient, as feature selection occurs during model construction without the need for additional evaluation. Embedded methods consider the bias of a given algorithm and might provide more accurate estimates. Examples of embedded methods include decision trees, which assess feature importance based on impurity reduction on each split (Gini importance) or permutation effects, and regularised models like the Least Absolute Shrinkage and Selection Operator (LASSO), which use penalised linear regression to obtain sparse coefficients capable of rejecting unuseful features.

2.4.4 Similarity-based methods

These methods determine feature importance by preserving data similarity, often encoding information in an affinity matrix and selecting top features that preserve manifold structure. They assume that data of the same class are usually close to each other. Various affinity measures have been defined, including Laplacian Score, Fisher Score, Spectral Feature Selection, and Relief Feature Selection. These methods

work well in both supervised and unsupervised settings, moreover they typically are very efficient acting as filters independent of any learning algorithm. However, they may struggle with handling redundancy, often resulting in highly correlated features.

2.4.5 Information-based methods

Information-based methods define hand-crafted information criteria aimed at maximising feature relevance while minimising redundancy. These methods often start from Shannon entropy definitions, considering both information gain and conditional information gain. Examples include methods for maximising mutual information (MIM), maximising relevance and minimising redundancy (mRMR), and other linear or non-linear combinations of Shannon-based information. While these methods also act as filters allowing the use of any algorithm afterwards, they address both relevance and redundancy questions compared to similarity methods, which mostly focus on relevance.

2.4.6 Sparse learning-based methods

These methods employ regression algorithms with regularisation terms aimed at minimising fitting errors. Feature selection is facilitated through $\mathcal{L}_p$ and $\mathcal{L}_{p,q}$ norm regularisation terms, where sparse regularisation terms force some feature coefficients to vanish, automatically eliminating irrelevant features. Methods like LASSO (based on $\mathcal{L}_1$ norm regularisation) have shown effectiveness in both supervised and unsupervised contexts, providing good performance and interpretability. However, some methods are tailored to specific learning algorithms and may not perform as well on other tasks. Additionally, optimisation in some cases involves non-smooth operations, requiring expensive computational matrix operations.

2.4.7 Statistical-based methods

These methods rely on statistical measures of the data, including variances, chi-square scores, and Gini indices, corresponding to most of the filter categories mentioned previously. They are often simple and straightforward with low computational costs, making them commonly used in preprocessing steps before applying more sophisticated feature selection techniques.

2.4.8 Other methods

It is important to note that apart from feature selection, other techniques for dimensionality reduction can also be employed, particularly through feature engineering. This involves creating new features from linear or nonlinear combinations of existing ones.
For example, Principal Component Analysis (PCA) can generate new features as linear combinations of the original ones, capturing most of the variance. On the other hand, Linear Discriminant Analysis (LDA) aims to create combinations of features in a supervised manner, enhancing the differences between classes for improved classification.

Feature selection preserves the physical meaning of the features and is often preferred for better model readability and interpretability. In contrast, feature engineering is applied directly to raw data without immediate interpretability. In the realm of radiomics, feature selection is commonly chosen as it facilitates an easier connection between specific groups of features and physiological or pathological data.

A thorough understanding of feature selection methods is essential when dealing with high-dimensional datasets. While numerous algorithms for feature selection have been proposed over time, a comprehensive review is beyond the scope of this work, for a more exhaustive overview see [53] [34].

In the following subsections, details about the dimensionality reduction methods to be used in this work will be provided, focusing on Minimum Redundancy Maximum Relevance, LASSO and Unsupervised Discriminative Feature Selection.

2.4.9 Minimum Redundancy Maximum Relevance

The Minimum Redundancy Maximum Relevance (mRMR) criterion, proposed by [75], falls under the category of supervised filter information-based methods. Since information is typically estimated on discrete variables, a discretisation process is necessary when working with radiomic features. Information is then obtained from some basics entropy concepts. For a discrete random variable X, its entropy, quantifying its uncertainty, is defined as:

$$H(X) = -\sum_{x_i \in X} P(x_i) \log(P(x_i)) \tag{2.34}$$

where x_i is the specific value taken by the random variable and $P(x_i)$ the probability of the given value over all the possible values. Similarly, the conditional entropy of X given another variable Y is defined as:

$$H(X|Y) = -\sum_{y_j \in Y} P(y_j) \sum_{x_i \in X} P(x_i|y_j) \log(P(x_i|y_j)) \tag{2.35}$$

where $P(y_j)$ is the prior of y_j and $P(x_i|y_j)$ the conditional probability of x_i given y_j.
The information gain, also known as mutual information shared between variables X and Y, can be expressed as:

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x_i \in X} \sum_{y_i \in Y} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \tag{2.36}$$

where $P(x_i, y_j)$ is the joint probability. When the two variables are independent, the information gain is null. Considering Y as the labels of the outcome, $\mathcal{S}$ as the current feature subset, and $X_j \in \mathcal{S}$ as a specific feature of the subset, a feature selection score for introducing a new feature X_k to the subset can be established based on a linear combination of information terms:

$$J_{mRMR}(X_k) = I(X_k;Y) - \frac{1}{|\mathcal{S}|} \sum_{X_j \in \mathcal{S}} I(X_k; X_j) \tag{2.37}$$

The aim is to select features that have a strong correlation with the outcome (represented by the first term $I(X_k, Y)$) while minimising correlation with other features (as penalised by the second term). This effectively reduces redundancy. Higher scores (J_{mRMR}) indicate greater feature importance, while lower scores suggest features that can be discarded.
The number of selected features is not fixed and does not depend on the ML algorithm applied afterwards. Hence, different classifiers may yield similar results with varying numbers of selected features. Generally, the preference is for the lowest number of selected features that still yield high model performance. In summary, the mRMR criterion maximises relevance while minimising redundancy, as its name suggests.

2.4.10 Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator technique, commonly referred to as LASSO, was pioneered by [102] within the domain of penalised linear regression. It offers a compelling characteristics: automated feature selection through the imposition of zero coefficients in the regression process. Considering a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, where observations $\boldsymbol{x}^{(i)} \in \mathbb{R}^p$ are characterised by p features,

an ordinary linear regression can be described as $y_i = f(\boldsymbol{x}^{(i)}; \boldsymbol{w}) + w_0 = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + w_0$, and optimal parameters $\hat{\boldsymbol{w}}$ are derived by minimising the $\mathcal{L}_2$ norm, or ordinary least squares:

$$\hat{\boldsymbol{w}} = \text{argmin}_{w \in \mathbb{R}^p} ||\mathbf{X}\boldsymbol{w} - \mathbf{y}||_2^2 \tag{2.38}$$

Penalised regression introduces another term representing the $\mathcal{L}_p$ norm of the weights.
In LASSO regression, this is achieved using the $\mathcal{L}_1$ norm:

$$\hat{\boldsymbol{w}} = \text{argmin}_{w \in \mathbb{R}^p} ||\mathbf{X}\boldsymbol{w} - \mathbf{y}||_2^2 + \lambda ||\boldsymbol{w}||_1 \tag{2.39}$$

Here, λ governs the strength of the penalty. This formulation equates to a constrained optimisation problem:

$$\hat{\boldsymbol{w}} = \text{argmin}_{||w||_1 \leq t} ||\mathbf{X}\boldsymbol{w} - \mathbf{y}||_2^2 \tag{2.40}$$

Practical constraints are imposed on learned parameters, with a narrower range of permissible coefficients as the penalty intensifies. Sparse coefficients can be intuitively identified by examining the constrained parameter space, which assumes a rotated square shape defined by $||\mathbf{w}||_1 \leq t$ as can be seen in Figure 2.6. Meanwhile, the least squares quadratic form exhibits an elliptical contour centred on the least squares parameters. Typically, the elliptical contour intersects the allowed region at one of its corners, leading to a selection of zero coefficients. This geometric trait renders the $\mathcal{L}_1$ penalty apt for feature selection, nullifying coefficients for non-informative features. Non-zero coefficients can then be ranked based on the magnitude of their estimated values.
LASSO can be directly applied to regression problems or utilised to generate a feature subset for subsequent classification models. It's important to note that for coefficients to hold significant values, data should be rescaled to a similar range. Additionally, the penalisation parameter requires tuning. This typically involves a grid search across different parameters and iterative evaluation on test data using mean square error for prediction assessment. Moreover, in addressing collinearity, the shrinkage operator facilitates feature selection by nullifying the coefficient of one of the correlated variables.

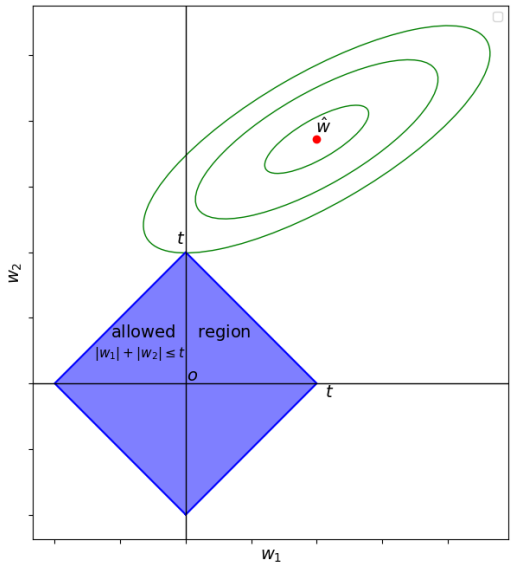


FIGURE 2.6: Example of the parameter space in penalised regression. The rotated square region represents the constraints imposed by the $\mathcal{L}_1$ norm, limiting the allowed region for the weights $\boldsymbol{w}$. The contour of the unpenalised least squares solution is shown with green ellipses centred at the best parameter $\hat{\boldsymbol{w}}$. One can notice that the shape of the constrained region enhances the probability of finding intersections with the axes with null weights, thus discarding the corresponding features.

2.4.11 Unsupervised Discriminative Feature Selection

Another sparse learning approach, known as $\mathcal{L}_{2,1}$ norm regularised discriminative feature selection for unsupervised learning (UDFS), combines discriminative analysis with the minimisation of the $\mathcal{L}_{2,1}$ norm. This method leverages local information to extract discriminative features while also preserving the data manifold structure. Local information has been demonstrated to be more crucial than global information in both classification and clustering tasks [113]. It offers an unsupervised approach to feature selection, assuming linear separability of class variables without requiring label information.
Considering a dataset as a matrix $X \in \mathbb{R}^{n\times p}$ with n, p respectively the number of samples and features, one can define the centred data as:

$$\tilde{X} = H_n X \quad \text{with} \quad H_n = \mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \in \mathbb{R}^{n\times n} \tag{2.41}$$

and $\mathbf{1}_n \in \mathbb{R}^n$ is a column vector with all unit elements. Then one can define the label matrix in terms of one hot encoding as $Y = [y_1, ..., y_n]^T \in \{0, 1\}^{n\times M}$ where M is the number of classes.
In a similar fashion one can define the scaled label matrix $G = [G_1, ..., G_n]^T = Y(Y^TY)^{-1/2}$.
From those matrices one can define the total scatter matrix encoding the overall variability or dispersion of the data in the entire dataset as:

$$S_t = \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T = \tilde{X}\tilde{X}^T \tag{2.42}$$

where μ is the mean of all samples, and the between scatter matrix encoding variations between different classes in the dataset:

$$S_b = \sum_{i=1}^{M} n_i(\mu_i - \mu)(\mu_i - \mu)^T = \tilde{X}GG^T\tilde{X}^T \tag{2.43}$$

where μ_i, n_i are respectively means and number of samples of class i.
With the objective of maximising S_b while minimising S_t, a local set of k-nearest neighbours is constructed for each data point x_i, denoted as $N_k(x_i)$, comprising x_i and its k nearest neighbours.
This set enables the definition of a local data matrix, $X_i = [x_i, x_{i1}, ..., x_{ik}]$, along with the local scatter matrices:

$$S_t^{(i)} = \tilde{X}i^T\tilde{X}i^T, \quad S_b^{(i)} = \tilde{X}_i G(i)G(i)^T\tilde{X}_i^T, \quad \text{where } \tilde{X}_i = X_i H_{k+1} \tag{2.44}$$

and a selection matrix $P_i \in \{0, 1\}^{n\times(k+1)}$ such that $G(i) = P_i^T G$.
Subsequently, in the absence of label information, UDFS assumes a linear classifier $W \in \mathbb{R}^{p\times M}$ that maps each instance $x_i \in \mathbb{R}^p$ to its class $G_i = W^T x_i \in \mathbb{R}^M$.
Finally, a local discriminative score, denoted as DS_i, can be defined as:

$$DS_i = Tr\left[(S_t^{(i)} + \lambda\mathbb{I}_p)^{-1}S_b^{(i)}\right] = Tr\left[W^T X S_i \tilde{X}_i^T(\tilde{X}_i\tilde{X}_i^T + \epsilon\mathbb{I})^{-1}\tilde{X}_i S_i^T X^T W\right] \tag{2.45}$$

with ϵ a parameter for invertibility. A larger value of DS_i indicates that W possesses strong discriminative ability for the datum x_i. The objective is to find the optimal W that achieves the highest discriminative score for all instances, while incorporating an $\mathcal{L}_{2,1}$ penalty for sparse feature selection, thus minimising the following function:

$$\mathcal{C}_{UDFS} = \sum_{i=1}^{n}\left\{Tr\left[G_{(i)}^T H_{k+1} G_{(i)}\right] - DS_i\right\} + \lambda||W||_{2,1} \tag{2.46}$$

where λ is the parameter controlling data sparsity. Similar to LASSO, once the optimal W is found, features corresponding to null coefficients can be discarded or ranked based on their weights.

2.5 Evaluation metrics for performance assessment

The most common approach to evaluating a model's predictive capabilities is by estimating its generalisation error. Depending on the dataset's size and characteristics, different evaluation methods can be employed. When ample data is available, a straightforward approach involves dividing the dataset into two subsets: a training set for model training and a test set for evaluating generalisation performance. Typically, the majority of the data, such as 80-90%, is allocated to the training set, with a smaller portion reserved for evaluation. This setup allows for a reliable estimate of the generalisation error in cases with a large dataset. In cases where models require extensive hyperparameters tuning, a portion of the training set is often set aside for hyperparameters validation. This step helps in selecting the optimal hyperparameters before assessing the model's performance on the test set.
However, when data is limited, this partitioning may not be feasible. To address this challenge, k-fold cross-validation comes to the rescue. In k-fold cross-validation, the dataset is divided into k nearly equal-sized portions or folds. The model is trained using $k-1$ folds and validated on the k-th fold to estimate the prediction error. This process is repeated for all fold combinations, resulting in k different estimates of the prediction error. These estimates are then averaged to produce a single estimate of the average generalisation error. Moreover, fold subdivision for small datasets can also have an impact. For this reason, repeated k-fold cross-validation is often performed, generating the sets of folds N times, resulting in $N \cdot k$ estimates of model performance.
When comparing different models, it's essential to consider not only the average estimation error but also its variance. A model with slightly lower performance but lower variance may be preferred, as it indicates greater reproducibility and robustness. Thus, models with lower variance are often favoured, even if their performance is slightly inferior.

Various evaluation metrics have been defined and widely used for both classification and segmentation problems. Segmentation can be viewed as a pixel or voxel-wise classification, allowing for the application of common definitions. In binary classification, each instance of a dataset is assigned to one of two classes, typically corresponding to positive or negative class labels. Many common classification models provide a continuous output in terms of membership probability. To obtain the final outcome class, a binarisation threshold must be applied.
Considering the true class labels, different situations arise:

- a positive instance correctly classified as positive, corresponding to a true positive (TP);
- a positive instance incorrectly classified as negative, leading to a false negative (FN);
- a negative instance correctly classified as negative, resulting in a true negative (TN);
- a negative instance incorrectly classified as positive, resulting in a false positive (FP).

Given the various scenarios outlined above, several metrics have been defined and are commonly employed to evaluate classification performance:

1. **Accuracy:** the most common metric used in classification, defined as the ratio of correctly classified samples over the total number of samples:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.47}$$

However, accuracy may not be suitable for all datasets due to its sensitivity to class imbalance. Even with very poor results on the minority class, accuracy can be high if the class of interest represents a small subset.

2. **Precision:** defined as accuracy but limited to positive predictions, the ratio of true positive predictions to the total positive predictions (TP + FP):

$$\text{precision} = \frac{TP}{TP + FP} \tag{2.48}$$

Precision is particularly useful in medical fields where the impact of false positive predictions is high, ensuring reliable predictions. However, precision alone is insufficient for performance estimation, as models generating very few positive predictions can still yield high precision.

3. **Sensitivity** (Recall, or True Positive Rate TPR): The ratio of true positive samples over the total number of positive instances (TP + FN), measuring how well the model captures positive instances:

$$\text{sensitivity (recall, TPR)} = \frac{TP}{TP + FN} \tag{2.49}$$

Precision and sensitivity have an inverse relationship; as one increases, the other decreases. In medical contexts, high sensitivity is often preferred to identify all possible diseases, even at the expense of some extra false positives.

4. **F1 Score** (Dice Index): A metric combining information from both precision and recall, defined as the harmonic mean of precision and recall:

$$\text{F1 (dice)} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2.50}$$

A high F1 score indicates a good balance between precision and recall, crucial for unbalanced datasets where a similar weight is desirable for both metrics.

5. **Specificity:** Also known as the true negative rate (TNR), the percentage of false samples correctly labelled:

$$\text{specificity (TNR)} = \frac{TN}{TN + FP} \tag{2.51}$$

This metric is essential to evaluate the rate of false positives, especially in medical diagnostics. It can be used to derive also the False Positive Rate (FPR), defined as FPR = 1 - TNR.

All these metrics serve as valuable tools for monitoring model performance.
However, the most commonly employed method for assessing the predictive accuracy of classifiers is the receiver operating characteristic (ROC) curve. ROC curves plot the true positive rate on the y-axis against the false positive rate on the x-axis, providing a concise representation of the trade-off between benefits (true positives) and costs (false positives). When a discrete classifier is utilised, a single point in the ROC space is presented. When probabilistic classifiers are utilised, multiple points are defined in the ROC space, each corresponding to a choice of the binarisation threshold. By continuously varying the binarisation threshold, a curve can be traced in the ROC space. An example is depicted in Figure 2.7. Several notable points in the ROC space are worth mentioning. For instance, the origin (0,0) represents a classifier that makes no positive predictions, thereby avoiding both errors and true positive predictions. Conversely, point (1,1) indicates a model that unconditionally predicts positive classes. The point (0,1) represents the ideal scenario of perfect classification, with a null false positive rate and a perfect true positive rate. Classifiers with performances located in the top left side of

the ROC space demonstrate good performance. It is important to note that a ROC curve provides an estimate of the relative prediction score, as calibrated predictions are not necessary for simply discriminating positive from negative instances [25].
While the ROC curve offers a comprehensive 2D representation of model performance, comparing models may require a single performance value. A common approach to condensing the information from the ROC curve into a single scalar is to calculate the area under the curve (AUC). The AUC, which ranges from 0 to 1, reflects the classifier's discriminatory power, with higher values indicating better performance. Notably, even a random classifier achieves an AUC of at least 0.5, with increasing AUC values indicating progressively better classifier performance. Generally, classifiers with an AUC above 0.8 are considered good, indicating strong predictive power, while values below this threshold suggest more discrete predictive ability.

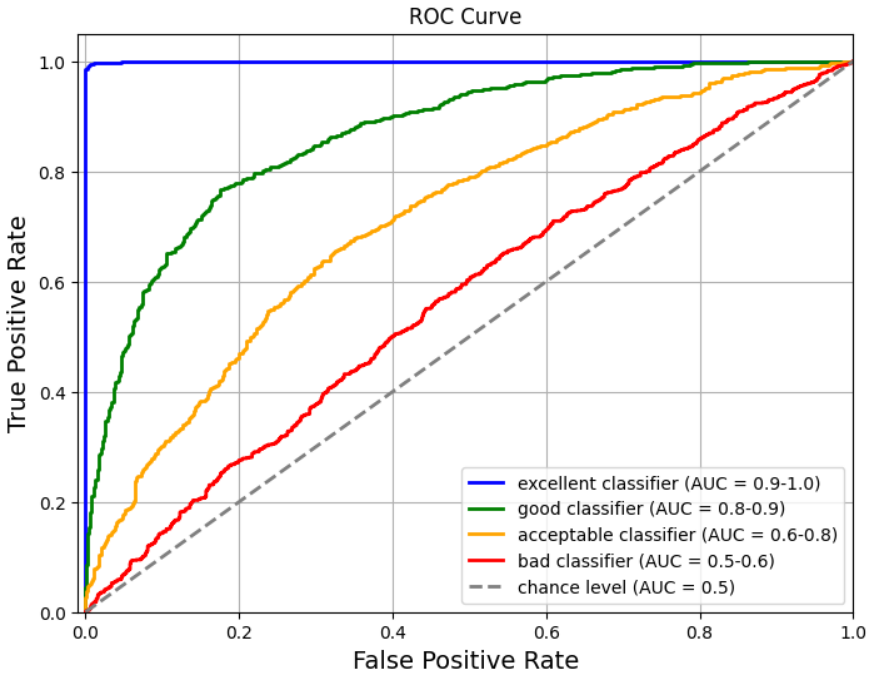


FIGURE 2.7: Example of ROC curves and corresponding AUC levels. The diagonal line represents the performance of a random classifier, while the various curves depict increasingly superior performances. The top-left corner represents the perfect classification scenario, while the other curves illustrate varying levels of AUC.

Overall, depending on the specific case, the choice of one metric over another will be more appropriate. In the subsequent chapters covering segmentation and classification, various metric choices will be employed to address different tasks.

Chapter 3

Development of the lesion segmentation model

In the realm of medical imaging, lesion segmentation has emerged as a crucial tool for monitoring tumour progression and delineating the region of interest to extract radiomic features, facilitating the characterisation of lesions and the development of radiomic models. Traditionally, this task has been performed manually by expert radiologists. However, the inherent variability between operators can significantly impact subsequent image analysis, particularly in radiomic feature extraction. Therefore, assessing compatibility and variability among operators is essential to ensure reproducibility and minimise bias in further analysis. Consequently, the development of automated models has become increasingly popular to mitigate the effects of human variability.
This study seeks to establish a comprehensive radiomic pipeline, commencing from the segmentation phase, providing an automated segmentation approach specifically designed for malignant breast lesions in DCE-MRI using a limited dataset comprising 131 cases.
This chapter will initially focus on selecting appropriate evaluation metrics to compare segmentations, crucial for analysing interobserver variability and establishing benchmarks to evaluate AI-generated results against ground truth, laying the foundation for subsequent model development. Among the most used scores emerges the Dice index representing the degree of overlap between two segmentations. A subset of segmentations, evaluated by two different radiologists, will be compared and analysed, revealing generally good overlap (median Dice score of 0.79 with an interquartile range of 0.70-0.85). However, systematic differences in size will be observed, with one operator tending to maintain larger margins and identify more small components.
After evaluating the variability among different operators, indicating the necessity of transitioning to automated segmentation methods, a comprehensive literature review is performed, revealing a growing interest in deep learning techniques, especially convolutional neural networks like the U-Net architecture. With these insights and considering available hardware resources, a customised model based on the V-Net architecture will be proposed, comprising two stages for segmenting whole breast tissue and lesions. The proposed model's characteristics and limitations will be discussed, revealing its moderate performance in lesion segmentation (median Dice score of 0.73 with an interquartile range of 0.53-0.84), which still aligns with operator performance. Despite the challenges presented by the limited size of the dataset, there is still room for enhancing the segmentation process to achieve more unbiased results, independent of the radiologist's level of experience.

3.1 Manual segmentation challenges

Before delving into the segmentation stage, let's first introduce the dataset of interest for the study. The retrospective study deals with a cohort of 131 women who underwent DCE-MRI before receiving NAC under the supervision of the Azienda Ospedaliera Universitaria Integrata Verona between January 2016 and November 2021. These women presented at least one malignant breast lesion, including cases with multicentric and multifocal tumours. The imaging protocol for all patients involved DCE-MRI with a 3T field on the same scanner (Achieva Philips Medical Systems, Cleveland, Ohio, USA) utilising the THRIVE sequence (T1-weighted high-resolution isotropic volume examination), fat saturated (SPAIR, TE= 2 ms, TR= shortest) acquired before and after the intravenous administration of a

gadolinium-based contrast agent, 0.2 mL/kg of gadobenate dimeglumine or 0.1 mL/kg of Gadoteridol (Bracco Imaging, Milan, Italy), followed a 20 mL saline flush, with a temporal resolution of 90 s. The study was approved by the Institutional Review Board of the Hospital. Patients provided consent for the anonymised processing of their data, and images were processed in NRRD file format, ensuring the absence of any personal information. After image subtraction, segmentation was performed on the second subtracted post-contrast image, the one that presents the maximum contrast enhancement for malignant lesions.
A portion of the dataset is used to investigate interobserver variability, with 72 cases for which two expert radiologists independently delineated ROIs according to a common procedure. The remaining cases were segmented only by one radiologist. Lesions were delineated slice-by-slice using the software 3D Slicer (https://www.slicer.org) and finally interpolated, smoothed, and merged to form a unique volume of interest exported as a binary mask.
The segmentation process is inherently time-consuming, mainly due to the significant number of slices within each image. The images were acquired in 3D at high resolution across all planes, with voxel spacing ranging from a minimum of [0.7, 0.7, 0.5] mm to a maximum of [1, 1, 1.1] mm, and an overall median spacing of [0.89, 0.89, 0.95] mm. The sizes of the images varied from a minimum of [352, 384, 120] to a maximum of [528, 528, 392], with a median size of [384, 384, 180]. This means that radiologists were required to examine up to 392 slices to locate a lesion for segmentation. However, in practice, operators often annotated only a subset of these slices, resorting to interpolation techniques to fill the gaps in the annotation process. Additionally, the process allows for a variety of morphological operations, such as the application of smoothing kernels with varying shapes (e.g., Gaussian, median, etc.) and sizes. Unfortunately, this practice diminishes the feasibility of building extensive datasets with comprehensive manual annotations, limiting the potential for in-depth lesion feature analysis. Hence, automation becomes crucial in augmenting the number of annotated databases.
The entire dataset will be leveraged to construct an automated segmentation pipeline. The segmentation model also necessitated delineating the entire breast ROI, a process similarly conducted by two different operators. The automated process will utilise delineated ROIs for both lesions and breast area in a balanced proportion from the two operators to mitigate human bias.

3.1.1 Measures for segmentation comparison

Considering a medical image of volume $V = w \cdot h \cdot d = n$ with w, h, and d, respectively, width, height, and depth in terms of the number of voxels, the whole volume can be represented as a set of points $X = \{x_1, ..., x_n\}$.
One can define the segmentation performed by the operator $j = \{1, 2, ...\}$ as the partition $S_j = \{S_j^L, S_j^B\}$ of X associated with the membership function $f_j^i(x) = \begin{cases} 1 & \text{if } x \in S_j^i \\ 0 & \text{if } x \notin S_j^i \end{cases}$ for $i = L, B$, where S^L represents the identified lesion and S^B is the background.

To evaluate the agreement between different segmentations, various metrics can be used depending on the specific task [99]. One can distinguish distinct categories, each with its own advantages:

- **overlap based metrics:** compare both the location and the size of the lesion;
- **volume based metrics:** compare just volumes neglecting shape, alignment and position (implicitly assuming that segments are always aligned), leading to a meaningful comparison only when alignment and overlap are high;
- **distance based metrics:** focussing on accuracy of the boundaries and shape, provide more precise information on position and distances where overlap-based metrics fail. In general quite

sensitive to outliers with possible tendency to over-penalisation (e.g. the case of the Hausdorff distance).

The requirement to focus more on alignment and extent rather than accuracy on boundaries led to the choice of the most common overlap-based measure, the Sørensen–Dice coefficient [22],[93], defined as follows:

$$DSC = 2 \cdot \frac{|S_1^L \cap S_2^L|}{|S_1^L| + |S_2^L|} \tag{3.1}$$

or it can be equivalently expressed in terms of the related Jaccard index [41] J:

$$DSC = 2 \cdot \frac{J}{1+J} \quad \text{with} \quad J = \frac{|S_1^L \cap S_2^L|}{|S_1^L \cup S_2^L|} \tag{3.2}$$

Despite being numerically different, DSC and J measure the same aspects and provide an equivalent ranking of the agreement between the segmentations. It should be noted that unlike the Jaccard index, the Dice score is not a metric as it does not satisfy the triangular inequality. Instead, it should be regarded as an overlap measure. However, despite this limitation, the Dice score is more commonly used due to its more immediate interpretability.
However also volume based metrics can be useful especially to discover possible bias between the operators, for this reason one can introduce also the relative size difference between the two segmentations defined as follows:

$$RSD = \frac{|S_2^L|}{|S_1^L|} - 1 \tag{3.3}$$

where S_1^L is considered the reference segmentation. The distribution of RSD provides information on possible systematic differences between observers.

Other sources of variability between operators refer to the intrinsic topology and geometry of segmented regions; in particular, one can identify and compare the number of disconnected components defined in each segmentation. Finally, a visual inspection of the segmentations also plays an important role in confirming the agreement of the operators.

Proposed method
The evaluation of the interobserver variability was performed according to the following procedure:

- **preliminary check:** before comparing the segmentations, it is necessary to verify that the masks are defined in a space of the same dimensionality (the same number of voxels and the same voxel size) and aligned (the same spatial origin). In case the operators performed further operations on the original images varying those attributes, one has to perform subsampling of the mask in the space with higher dimensionality, to match with the other, and apply a translation to get spatial alignment. These steps clearly may introduce a little source of error related to the discrete nature of binary segmentation.

- **overlap-based comparison:** masks are compared voxel-by-voxel to obtain an estimate of the overlap by means of the Sørensen–Dice coefficient;

- **volume-based comparison:** segmentation volumes are calculated as the cardinality of each point set and compared using the relative size difference;

- **visual inspection:** masks are visualised as 3D volumes in Slicer to check further anomalies that have not been captured by the previous metrics (possible identification of lesions located at different sites, discrepancies number of disconnected components, etc.).

3.1.2 Interobserver variability

The proposed method for estimating variability among different operators was applied to a subset comprising 72 lesion segmentations provided by two different radiologists. In some cases, resampling and alignment were necessary, assuming that any error introduced during these corrective procedures is negligible; the results are presented as follows. One can observe in Table 3.1 the summary of the results of the selected metrics. A raw estimate of the mean scores, along with their standard deviation, is not sufficient for both the Dice index and the relative size difference. It is necessary to examine the distribution, as both distributions are highly skewed and do not resemble a Gaussian. In such cases, simply taking the mean would underestimate the performance. Therefore, for completeness, the median together with the interquartile range is also reported. With regard to the overlap estimated

metric	**mean**	σ_s	**skewness**	**median**	**IQR**	**min**	**max**
DSC	0.74	0.18	-2.1	0.79	[0.70-0.85]	0.09	0.92
RSD	0.6	1.8	5.2	0.2	[0.1-0.5]	-0.9	12.5

TABLE 3.1: Summary statistics of Dice similarity coefficients and relative size differences between two operators

using the Dice index, overall good agreement appears to be achieved between the two observers, with a median value of $\overline{DSC} = 0.79$ and an interquartile range of 0.70-0.85. An histogram of the obtained DSC is illustrated in Figure 3.1, revealing a negatively skewed distribution with a long tail at very low values. There are 6 of 72 masks that are significantly in poor agreement with $DSC < 0.5$, and a small component with $0.5 \leq DSC < 0.7$. The former cases might be related to ambiguous lesions that were differently interpreted by the two operators, while the latter could be attributed to intrinsic interobserver variability. Despite the few pathological cases, one can assume that the two observers have optimal agreement for most of the cases, as one can see from the peaks in the histogram above 0.8. However, the origins of the discrepancies should be further investigated using other metrics.

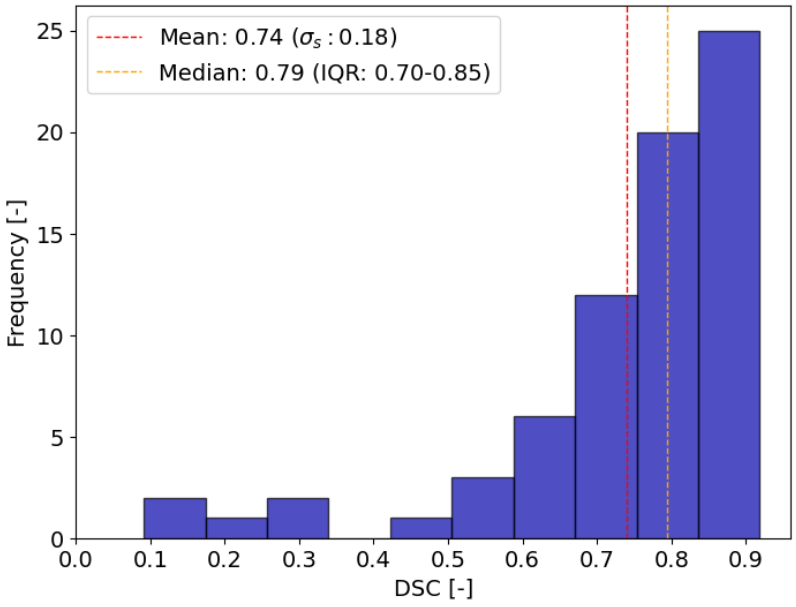


FIGURE 3.1: Distribution of Sørensen-Dice coefficient, showing an overall good agreement despite the presence of some anomalies with very low DSC.

Systematic effects were investigated by comparing mask volumes and calculating the relative size difference. A scatter plot illustrating the sizes of the two segmentations in terms of the number of voxels (using the same spacings) is depicted in Figure 3.2. Here, it is evident that most of the points lie on or above the line representing equal sizes ($|S_2^L| = |S_1^L|$), indicating that one of the two radiologists (operator 2) consistently generates larger segmentations. In Figure 3.3 also the histogram of RSD values is shown, and one can see that in most cases RSD is positive, which implies that the masks produced by the second operator are always larger than the ones produced by the first.

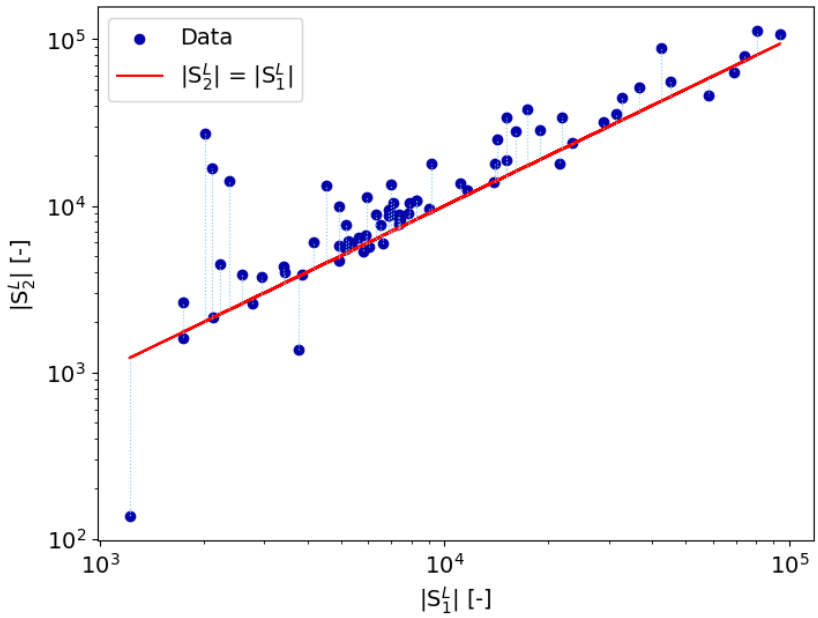


FIGURE 3.2: Scatterplot illustrating segmentation sizes, revealing a consistent trend where the sizes generated by operator 2 are consistently larger than those generated by operator 1.

This can be regarded as a systematic difference in one operator keeping larger boundaries around lesions or analogously in the other keeping a minimal distance. Some extreme points can again be observed at very positive values, synonyms of important differences in size between the delineated tissues, possibly due to the ambiguity of the lesion leading to misidentification of noncancerous tissues in some regions by one of the two operators.

In the end, both cases with good agreement according to the previous metrics and the anomalies were examined by visual inspection. Different situations occurred and the main possibilities can be summarised with the examples reported in Figure 3.4. Each figure shows the 3D visualisation of the segmentation in Slicer, respectively, in red, the mask depicted by the first operator, and in transparent green, the mask defined by the second operator.

Most instances with high DSC and almost zero RSD are characterised by a very compact lesion, as shown in Figure 3.4a. Such cases seem to be easily identified and less subject to interobserver variability; however, when the number of lesions increases and the shape becomes more irregular, the variability between observers becomes more and more evident.

An example of more complicated segmentation is shown in Figure 3.4b, where both operators identify two separated lesions. The second operator selects a much larger volume and maintains a coherently larger boundary with respect to the first operator, leading to a modest agreement on overlap since the first mask is almost entirely enclosed in the second. However, despite the overlap the volume difference becomes relevant.

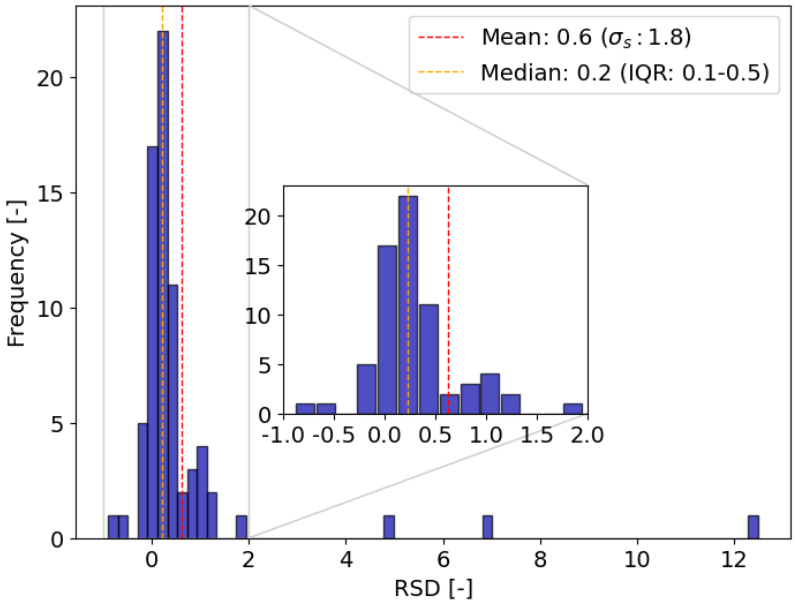
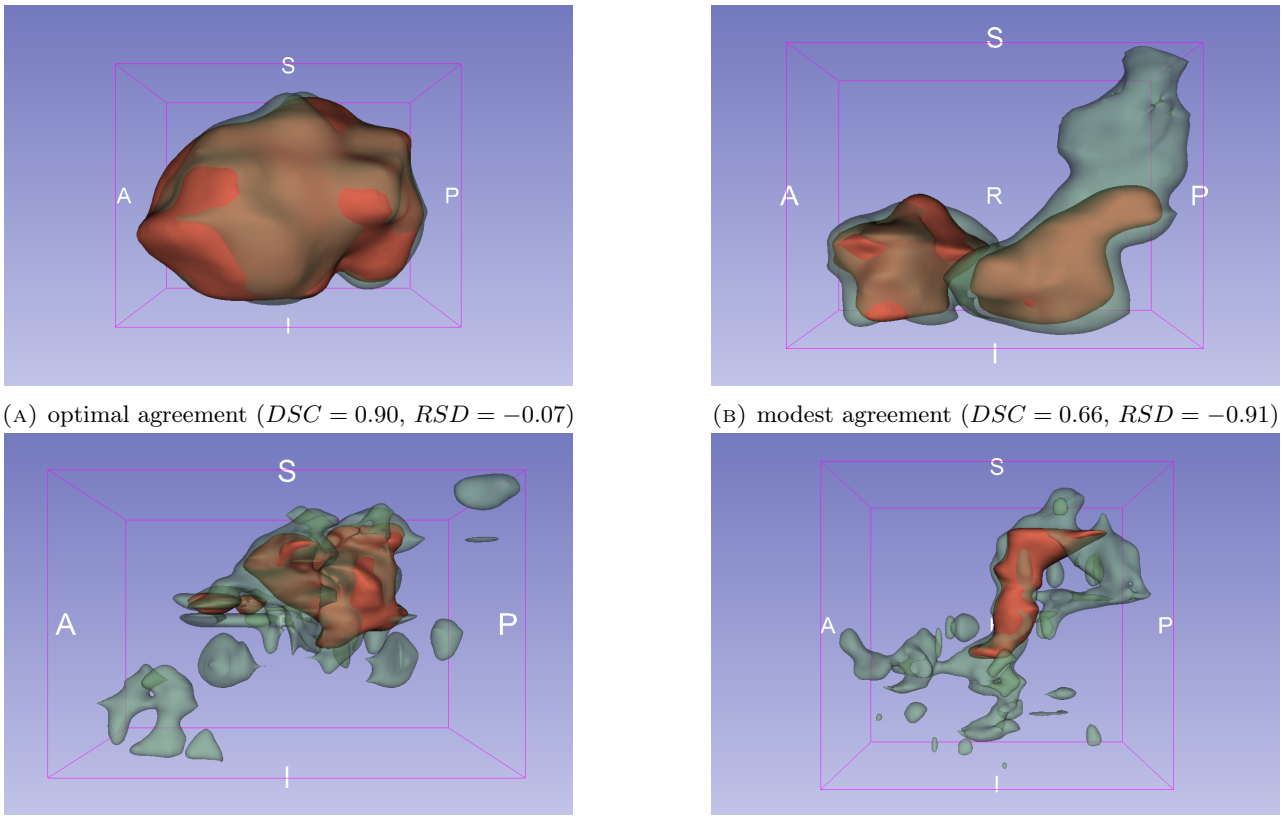


FIGURE 3.3: Histogram of relative size difference showing a general positive trend corresponding to bias between operators. Some outliers are present at very high positive values, whereas a zoom of the peak region shows that the median value is not centred in 0 but instead at positive values.

Another situation occurs when the shape of the lesion is quite irregular, even with an acceptable overlap, one can see in Figure 3.4c that the topology of the lesion is actually different, while the main larger components have similar segmentation, the number of smaller components identified by the two observers no longer matches, and the second operator tends to always select a larger region with much more small fragments.
Analogously one can see that in other cases, as shown in Figure 3.4d, the overlap is minimal, the volume difference is quite important, and even boundaries of the main component are significantly different between the two operators.
Overall, visual inspection confirmed good compatibility in most of the cases, but also underlined the presence of defined differences between operators, that can be characterised as:

- **Systematic size differences:** tendency of placing always larger or smaller boundaries around lesions, producing actually very different volumes also for the main lesion;
- **Differences on small components:** disagreement on number, size and location of smaller ambiguous components, possibly dependent on the personal experience achieved by the observer.
- **Shape differences:** tendency to prefer more convex disk-shaped segments or more irregular star-shaped domains rich in irregular protrusions.

As a final consideration, the presence of small disconnected components with no one-to-one correspondence actually makes it difficult to use simple distance-based metrics that have both an ambiguous definition and an evident enhancement for even very small outliers. For these reasons, a further distance-based comparison does not give significant results and has not been reported.
In consideration of the evaluation metrics and visual inspection, it is evident that most of the masks generated by the different operators were compatible. However, notable differences in extension, geometry, and topology persist, which make each operator distinctly distinguishable from the others. These differences can compromise the reproducibility of the results and hinder further detailed analysis, for



(A) optimal agreement ($DSC = 0.90$, $RSD = -0.07$)

(B) modest agreement ($DSC = 0.66$, $RSD = -0.91$)

(C) poor agreement ($DSC = 0.58$, $RSD = -1.3$)

(D) very poor agreement ($DSC = 0.33$, $RSD = -1.9$)

FIGURE 3.4: Examples of masks differences with lesions of different degrees of complexity.

instance, discrepancies could appear in the extraction of radiomic features but also in monitoring the lesion size over time. In the end, in light of these observations, it becomes apparent that the potential differences and biases encountered in manual lesion segmentation could be effectively mitigated by leveraging automatic algorithms. By relying on automated processes, the inherent subjectivity and variability associated with manual segmentation might be minimised, thereby enhancing the reliability and reproducibility of the results. Thus, guided by these considerations, the idea of developing a customised automated pipeline emerged.

3.2 Literature review on segmentation techniques

A systematic literature review was conducted to examine the current landscape of breast lesion segmentation, focusing primarily on studies published from 2013 to 2023. PubMed (US National Library of Medicine, `http://ncbi.nlm.nih.gov/pubmed`) and Google Scholar (`http://scholar.google.it`) were utilised as primary databases, employing keywords such as 'breast MRI lesion segmentation' and related terms. The key trends are summarised here for conciseness, whereas a more detailed exploration is provided in Appendix A.

A significant portion of the identified studies lacked specificity regarding segmentation methodologies, incorporating it at various stages without detailed emphasis. Among these, manual segmentation methods were prevalent, followed by semi-automatic approaches such as pre-built software, level-set methods, and thresholding combined with manual refinements or convex hull algorithms. Region growing algorithms, notably employing the random walker, were also utilised in several cases. For automatic

segmentation, clustering techniques were commonly employed, with Fuzzy C-means being the most frequently utilised algorithm, followed by k-means and Gaussian Mixture Models. Some studies explored alternative methods such as Markov Random Fields and active contour models like Gradient Snake Vector Flow. In studies specifically focused on segmentation techniques, a diverse range of methods were observed. Overall, many studies divided segmentation methods into two stages: initially segmenting the entire breast tissue to narrow down the area of interest for lesion detection, followed by a specific focus on lesion segmentation. Whereas a minority of studies directly applied segmentation to the entire images. For breast segmentation, various techniques were employed, ranging from simple straight lines and marking points to more advanced probabilistic atlases and clustering methods such as fuzzy c-means (FCM). In recent years, the use of convolutional neural networks has gained traction, showing interesting performances in the segmentation of whole organs and tissues, including breast.
Regarding lesion segmentation, early approaches focused on clustering techniques, level sets, and active contour models. Some studies utilised time intensity curves for time series analysis to inform clustering algorithms as FCM, while others employed energy functionals and background distribution for active contour models based on level sets. Region growing methods and shape priors were also explored, along with techniques addressing the temporal dimension using independent component analysis.
A notable trend emerged with the rise of deep learning, particularly CNN based on the U-Net architecture, which saw increasing adoption for lesion segmentation. Studies utilised U-Net and its variations, often incorporating different inputs such as single time points, triple time points and multiple images from DCE-MRI series or other complementary modalities. Hierarchical CNNs, patch-based methods, and 3D models were also investigated, with U-Net consistently demonstrating promising results across various approaches. Furthermore, semi-supervised methods were employed to tackle the challenge of limited labelled data, showing promise in leveraging also unlabelled slices for improved segmentation accuracy.
A prevalent characteristic among the studies, with few exceptions, was the utilisation of small datasets, which inherently limits the generalisability of their findings. Additionally, it is noteworthy that the majority of studies focused on segmenting 2D slices rather than 3D volumes, likely due to datasets having low resolution along the cranio-caudal direction. Many studies employed 2D segmentation on individual slices and subsequently merged them to generate a final 3D segmented volume. Others applied 2D approaches to each plane separately and then reconstructed the final image. Furthermore, as highlighted by [27], the training and testing procedures often involved only the slices containing lesions, yielding promising results but operating under the assumption that the correct slices are already known and selected. However, a fully automated procedure should ideally be applicable to the entire image. Results obtained solely from lesion-containing slices may lack generalisation to generic volumes, as they might produce false positives in regions with similar enhancement. Unfortunately, the rationale behind this selection criterion is not consistently explained across studies, making it challenging to discern such cases. Moreover, there is variability among studies regarding the sequences of images used, and in some instances, different imaging techniques and parameters have been employed. The utilisation of fully 3D images and the incorporation of the temporal dimension in dynamic contrast-enhanced MRI have been explored to a limited extent, with only a few studies directly addressing 3D images rather than 2D slices. This preference for multiple 2D models trained across the three planes may stem from the considerable computational burden and resources required to train fully 3D or 4D models. Consequently, direct comparison of segmentation model performance across studies is challenging, and selecting an appropriate segmentation model based solely on the literature is not straightforward.

3.3 Model Development for Lesion Segmentation

After conducting a comprehensive literature review, the development of an effective lesion segmentation approach requires thoughtful consideration of both the dataset's nature and the available computational resources. A significant portion of prior studies focused on 2D approaches, employing the analysis of

MRI data slice by slice. While this methodology holds particular relevance for 2D MRI acquisitions that excite single slices at a time, its applicability also extends to high-resolution 3D images. From a computational cost perspective, the analysis of individual 2D slices proves less resource-intensive and yields a higher number of training elements. However, 3D models, in contrast to their 2D counterparts, capture spatial relationships and intricate features more holistically, providing a nuanced representation of lesions. This is especially advantageous in detecting smaller lesions, where the three-dimensional context becomes pivotal.
Other considerations should be made regarding the number of temporal points in the DCE sequence. Generally, the use of more than one point yields improved results, as indicated for instance by works of Galli et al. [27], Piantadosi et al.[76], Vidal et al. [106] and Zhu et al. [121]. The inclusion of multiple time point results useful especially when dealing with various lesion types, given the variation in enhancement curves across cases. However considering the specific case under study, focusing solely on malignant lesions, opting for a single significant point in the sequence might be a suitable choice. Since all images exhibit the same type of enhancement curve, this approach not only aligns with the objective of the study but also alleviates the computational burden.
Considering this perspective within the context of the 3D DCE-MRI dataset under examination, the study aims to develop a full pipeline capable of recognising and segmenting lesions starting from a single DCE-MRI image. To fully leverage the high resolution in all three dimensions, the preference is for the use of a 3D model. The segmentation approach will utilise a single time point, specifically chosen as the second subtracted image, which typically exhibits the most significant enhancement across all available cases. This time point also aligns with the one commonly used by radiologists for manual segmentation.

3.3.1 Model architecture

From the perspective of network architecture, excluding approaches that process volumes slice-wise, modification of the most used U-Net to 3D have been considered, as 3D U-Net [18] and V-Net [65], both leveraging volumetric convolutions, have emerged with similar promising performances. Inspired by the original U-Net [84], these architectures are designed as generalisations to 3D volumes. The U-shape structure is the hallmark of these architectures, as can be seen from figure 3.5, with the left side used for contraction to capture context, progressively reducing the input signal size and increasing the receptive field, while the right side is employed for expansion, allowing precise localisation and returning to the original input size.
The contraction path consists of multiple stages with the following structure:

1. Repeated application of convolutions each one followed by a non-linear activation unit;
2. Downsampling, doubling the number of channels, to increase the depth of feature maps (allowing to employ additional filters in subsequent convolutional layers to enhance the network's capacity to capture increasingly intricate patterns and structures within the data).

The expansion path includes:

1. Upsampling, halving the number of channels, to increase the spatial dimensions of the feature maps recovering higher spatial resolution;
2. Concatenation with feature maps from the corresponding stage of the contraction path, to combine high-resolution details with the abstract features learned in deeper layers;
3. Repeated application of convolutions each one followed by a non-linear activation unit.

The final output is obtained at the end of the expansion path through:

1. A final convolutional layer, reducing the number of output channels to the desired number of labels;
2. Conversion of final feature maps to probabilistic segmentations by applying an activation function.

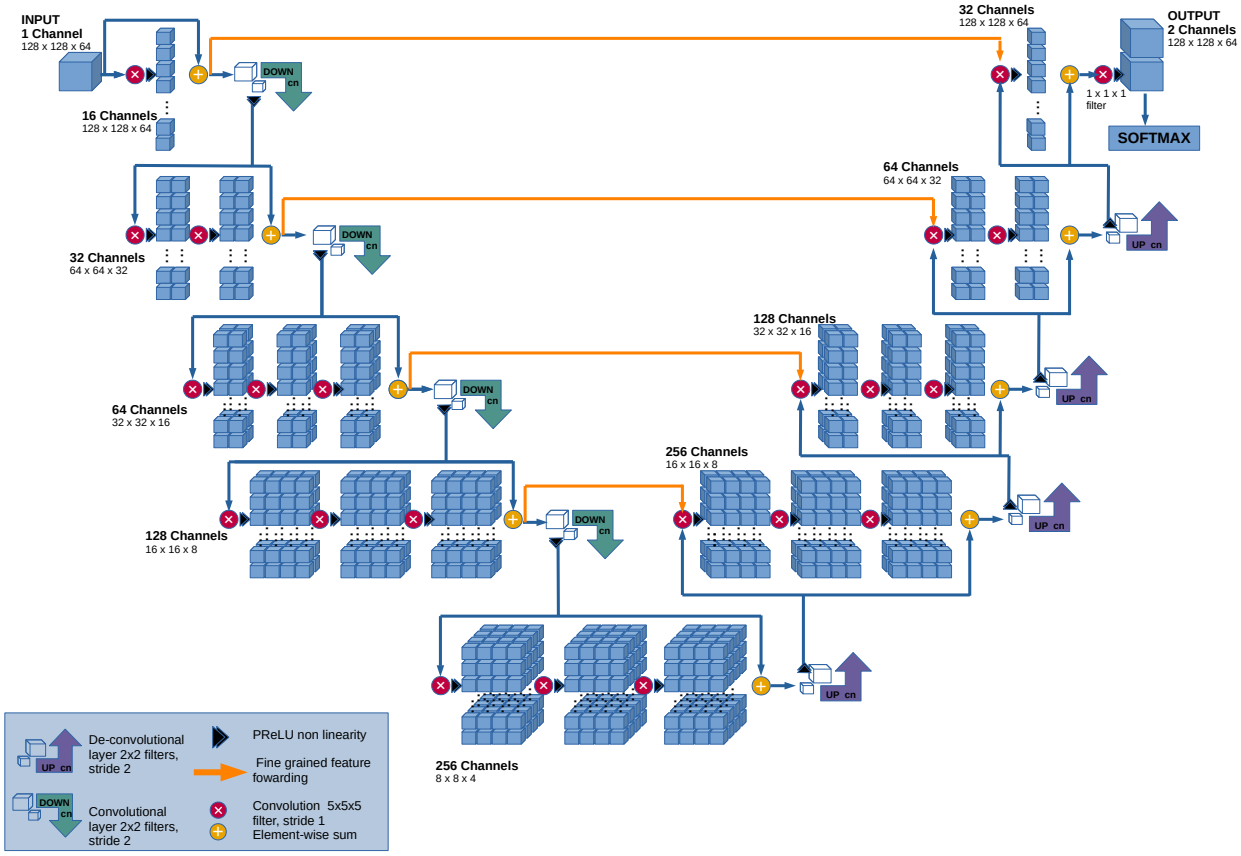


FIGURE 3.5: Schematic view of V-Net architecture. Adapted from [65].

The choice between 3D U-Net and V-Net requires delving into specific details of each model's layer structure. 3D U-Net employs four stages in the contraction path, utilising two subsequent convolutions of size 3x3x3 voxels in each. Following this, a copy of the last output is forwarded to the expansion stage, and downsampling is achieved through a 2x2x2 max-pooling layer leading to the subsequent stage. In the expansion path, upsampling is conducted using a 2x2x2 convolution with a stride of 2. After each upconvolution, the output is concatenated with the one from the contraction path, thereby preserving fine-grained details lost in the contraction path. This process spans four stages, with the first three consisting of two subsequent convolutions of size 3x3x3 voxels, and the last stage involving a single 1x1x1 convolution leading to the output. Batch normalisation and ReLU non-linearity are applied after each convolution in every stage.

V-Net, on the other hand, presents some differences, as depicted in Figure 3.5. Both downsampling and upsampling are performed with a 2x2x2 convolution and a stride of 2. As suggested by [94], substituting the max-pooling layer with a convolutional layer using an appropriate stride can preserve accuracy without any loss. This replacement provides an extra advantage of reducing memory usage during the training stage, as the storage of switches mapping the pooling output to the original input is no longer necessary for backpropagation. However, it's essential to note that the use of more convolutional layers

also increases the number of tunable model parameters and adds to its overall complexity. Another distinction is the utilisation of a parametric ReLU non-linearity (PReLU), enhancing accuracy and convergence with a learnable parameter for the slope. Additionally, the architecture includes a slightly more complex structure with five stages in both the left and right paths. In the contraction path, the first stage comprises a single 5x5x5 convolution, the second stage involves two subsequent 5x5x5 convolutions, and the remaining stages use triple 5x5x5 convolutions. Similarly, in the expansion path, the first two stages incorporate triple convolutions with a 5x5x5 kernel, followed by a stage with two 5x5x5 convolutions, another with a single 5x5x5 convolution, and finally, the last layer consists of a 1x1x1 convolution producing the final output channels. These channels are then converted into probabilities through a final softmax function. A distinctive feature of V-Net lies in the dual use of the input at each stage, which undergoes subsequent convolutional layers and non-linearities, and is simultaneously employed to learn a residual function by adding it to the output of the last layer of the stage before up or downconvolution. Similar to U-Net, the features extracted at the end of each contraction stage are forwarded to the right path to preserve fine details in the final prediction.
After evaluating the architectural nuances between V-Net and 3D U-Net, the distinctive features of V-Net led to its selection for our lesion segmentation task. It is worth noting that the field of medical image segmentation is dynamic and continuously evolving. In some studies, such as [110], minimal differences in performance were observed between these architectures, suggesting that further improvements, such as the incorporation of attention layers, could be explored for enhanced segmentation outcomes.

3.3.2 Segmentation pipeline

In designing an effective pipeline for lesion segmentation using the V-Net architecture, practical challenges related to diverse image resolutions and GPU limitations must be considered. The available dataset comprises images with varying resolutions, influenced by the Field of View (FOV) and individual patient characteristics. These images encompass not only the lesion region but also a full section of the torso, including other organs exhibiting similar enhancement to the lesion, introducing the potential for error. Given the computational expense of training a complex CNN, GPU acceleration is essential. However, GPU limitations necessitate careful selection of image sizes for efficient training.
To address these challenges, a two-stage approach was employed, consisting in a first model to segment the whole breast and a second model to segment lesions within the breast ROI.

Stage 1: breast segmentation model
To mitigate the analysis of undesired tissues and reduce image size, a model was developed to define the breast area. Raw breast segmentations, with ample borders, were used to train an initial V-Net architecture. Downsampling to a common lower resolution, compatible with available GPU capacity, was performed. As this stage focused on defining the breast area, downsampling did not compromise the required features, which are not dependent on fine-grained details. The resulting downsampled images and breast area segmentations were used to train a V-Net, producing a model that effectively restricts the area for further lesion detection. Post-processing involved morphological operations to obtain a mask with the original image resolution and smooth boundaries.

Stage 2: lesion segmentation model
Original images were cropped using the largest 3D bounding box containing the breast mask. Intensity values outside the breast mask were masked with a fixed background value. Despite a significant reduction in image size, it remained too large for direct GPU processing. To capture fine-grained features while preserving resolution, images were lightly downsampled to a common resolution and divided into smaller 3D patches. The training of a V-Net involved using these patches, containing both lesion and background areas, enhancing the network's capabilities and reducing the risk of overfitting.

The trained model is applicable to any single high-resolution volume patch. Lesion segmentation across the entire breast area is achieved by applying the model in an overlapping and tiled manner to ensure comprehensive information coverage.

Once the models have been trained, the final segmentation pipeline is applied as described in Figure 3.6, which involves the following steps:

1. heavy downsampling of the original image and application of the breast mask model;
2. postprocessing of the breast mask by upsampling to recover the original image resolution, followed by cropping and masking the area outside the breast mask;
3. moderate downsampling, if necessary, to reach unitary spacing, and division of the image into different sets of overlapping patches, then the lesion model is applied to all patches;
4. reconstruction of the final segmentation image using information from all patches by merging all the outputs;
5. eventual tuning of the segmentation threshold, if required, to refine the segmentation.

After defining the overall pipeline, the customised implementation details will be discussed in the next sections.

3.3.3 V-Net implementation and parameters

As discussed in Chapter 2, the training process of deep neural networks shares similarities with simpler supervised algorithms, involving the formulation of a cost function and utilising gradient descent to iteratively minimise it to find optimal parameters [63]. However, the presence of multiple layers in deep neural networks introduces a higher number of tunable parameters, escalating the computational complexity.

Output channels

In the original implementation of V-Net by Milletari et al.[65], the network takes as input a 3D image $\mathbf{x}$ of dimensions 128x128x64 voxels. At the output layer, it acts as a classifier with a softmax activation, providing two channels y^0 and y^1, defining respectively the probabilities of belonging to the background class 0 and to the class of interest 1. However, considering the simple binary classification task involving only discriminating voxels of interest (breast in the first stage and lesion in the second stage) from background voxels, the network can be simplified to have a single output channel passing through a sigmoid activation. This output map assigns to each voxel i the probability of belonging to the category of interest, denoted as $p(y_i) \equiv p(y_i = 1|\mathbf{x}_i; \mathbf{w})$, where $\mathbf{w}$ represents the weights learned by the model. The predicted segmentation is then derived by categorising voxels with $p(y_i) \geq p_{\text{thr}}$ as category of interest, and those with $p(y_i) < p_{\text{thr}}$ as belonging to the background class. This thresholding process serves as a binary classification criterion for determining the presence of the object of interest within the segmented output. In general, predicted masks are commonly obtained using a binarisation threshold of 0.5. However, this fixed threshold may not always be the optimal choice, particularly considering the characteristics of the model and the complexity of the data, especially in scenarios with high class imbalance and intricate background conditions. To address this variability, segmentation masks can be generated using a range of different thresholds. Evaluating the performance across these various thresholds allows for the determination of the most suitable threshold choice based on the specific characteristics of the dataset. This approach provides flexibility in adapting the binarisation threshold to the nuances of the data, ensuring a more robust and context-aware segmentation outcome.

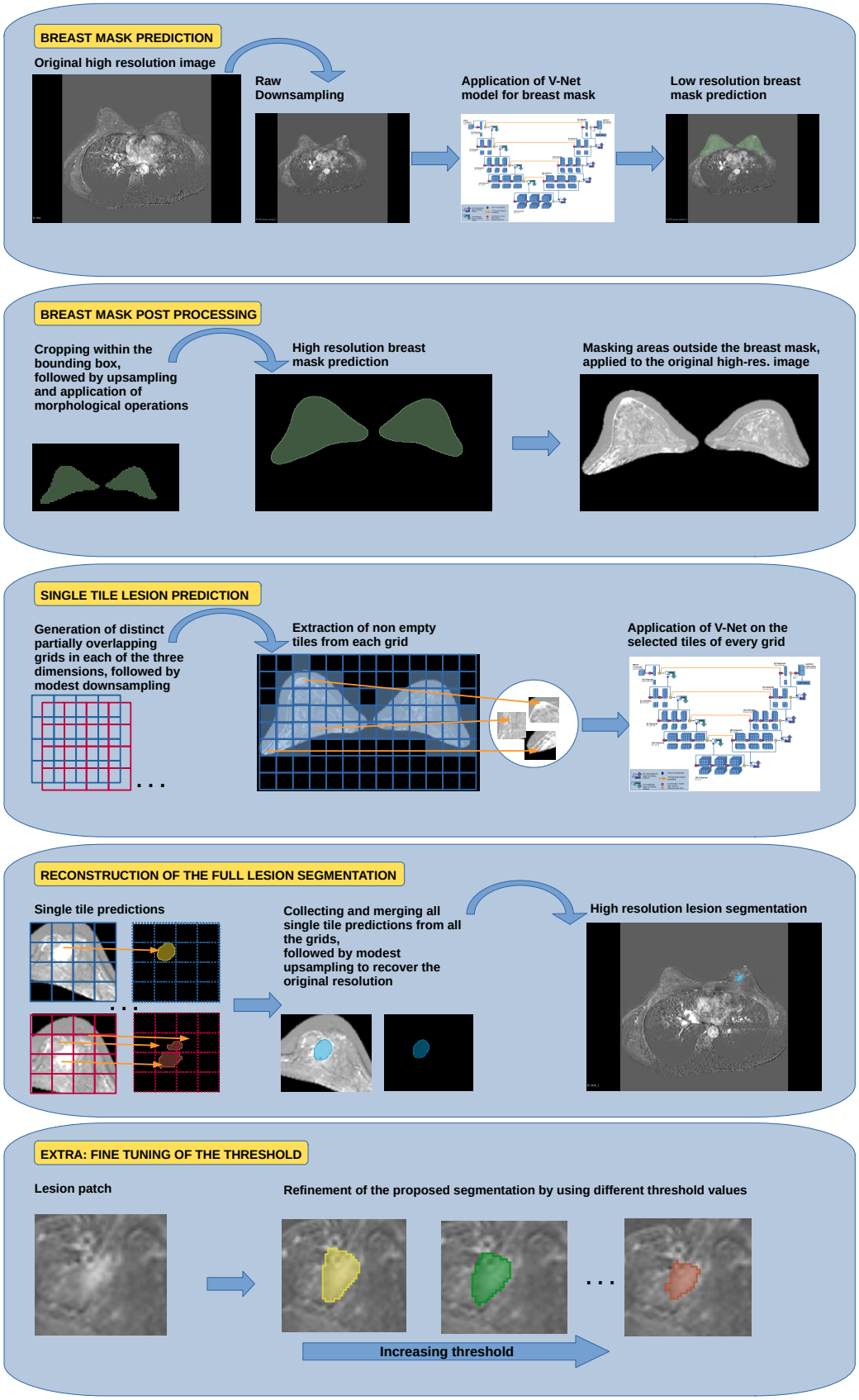


FIGURE 3.6: Segmentation pipeline

Cost function and evaluation metrics

Among the pivotal elements in the implementation of the neural network, the selection of both the loss function and performance metrics holds significance, as highlighted by [100]. It's crucial to note that while both the loss function and performance metrics contribute to performance evaluation, they serve slightly different purposes and possess distinct properties. The loss or cost function plays an active role in the model training phase, serving to optimise the model's parameters by quantifying the disparity between predicted and expected outputs. The primary objective during training is to minimise this cost function. However, it's important to acknowledge that these functions often have arbitrary scales, posing challenges in interpretation due to their dependence on specific model architectures and datasets. On the flip side, performance metrics come into play for evaluating the model post-training, offering insights into the model's generalisation capabilities and its accuracy in making predictions on new data. In contrast to loss functions, performance metrics provide a more immediately interpretable measure that remains independent of the model. They also facilitate comparisons between different configurations or models. A crucial distinction lies in the fact that during the training phase, the goal is to minimise the cost function to optimise parameters, while performance metrics are aimed at maximisation.
The selection of an appropriate loss function depends on the characteristics of the dataset and the segmentation requirements. According to findings by [5], three main classes of losses have been identified:

1. pixel-level losses: aiming to achieve high accuracy in classifying individual pixels;
2. region-level losses: focusing on the overall segmentation task, region-level loss functions are designed to maximise the overlap and alignment between the entire predicted mask and the ground truth;
3. boundary-level losses: tailored to emphasise the boundaries of objects, boundary-level loss functions are intended to separate overlapping objects more effectively.

Moreover, there is the option to combine losses from different classes, creating a so-called combo loss. This approach allows for simultaneously focusing on multiple aspects of the segmentation task, providing a more comprehensive optimisation strategy.
Among pixel-level losses, one of the most common for binary classification is the binary cross-entropy loss, which measures the difference between probability distributions of a given random variable. Also known as log loss, it is defined as the negative logarithm of the predicted probability for the target class:

$$\mathcal{C}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \tag{3.4}$$

where N is the total number of pixels (or voxels in the 3D case), $p(y_i)$ is the probability of pixel i belonging to class 1, $1 - p(y_i)$ is the probability of pixel i belonging to background class 0, and y_i is the ground truth label for pixel i. This loss tends towards zero when the target class matches the true label. To address highly unbalanced datasets, a variation of this loss involves assigning different weights to the terms for the two classes, resulting in a weighted binary cross-entropy:

$$\mathcal{C}_{wBCE} = -\frac{1}{N}\sum_{i=1}^{N} \alpha \cdot y_i \cdot \log(p(y_i)) + \beta \cdot (1 - y_i) \cdot \log(1 - p(y_i)) \tag{3.5}$$

where α and β are weights for class 1 and background class, respectively. By assigning higher weights to under-represented classes, for instance using inverse class frequency, the model focuses more on the minority class, thereby improving overall performance. Other versions of the cross-entropy are defined

by assigning increased weights to hard samples, such as the TopK Loss and the Focal Loss. These variations provide additional tools to fine-tune the loss function based on the specific challenges posed by the dataset and segmentation task.
On the contrary, region-level loss functions diverge from focusing on individual pixels, adopting a broader perspective that prioritises overall accuracy. These losses are instrumental in capturing the shape and layout of objects, particularly when global context outweighs pixel-level traits. One of the well-known region-based losses is the Dice loss, derived from the Dice coefficient, which quantifies the overlap between the prediction and target segmentation. The Dice loss is formulated as a variant of 1 minus a relaxed Dice coefficient, treating predictions as probability values rather than discrete binary values. This characteristic renders the loss functions differentiable and minimisable, making them valuable for unbalanced datasets with relatively small regions of interest. However, a common criticism is that, due to its non-convex nature, it might easily result in suboptimal outcomes [42]. For this reason, various modifications have been tried to make it more tractable. Several versions of the Dice loss exist, including the most generic Dice loss for C classes:

$$\mathcal{C}_{DSC} = 1 - \frac{1}{C}\sum_{c=0}^{C-1}\frac{2\cdot\sum_{i=1}^{N} y_i \cdot p(y_i^c)}{\sum_{i=1}^{N}[y_i^c + p(y_i^c)]} \tag{3.6}$$

the Dice Loss with squared denominator proposed by Milletari et al. [65] in the first use of V-Net to improve convergence:

$$\mathcal{C}_{mDSC} = 1 - \frac{1}{C}\sum_{c=0}^{C-1}\frac{2\cdot\sum_{i=1}^{N} y_i^c \cdot p(y_i^c)}{\sum_{i=1}^{N}[(y_i^c)^2 + (p(y_i^c))^2]} \tag{3.7}$$

and the generalised Dice loss [96]:

$$\mathcal{C}_{gDSC} = 1 - 2\cdot\frac{\sum_{c=0}^{1} w_c \sum_{i=1}^{N} y_i^c \cdot p(y_i^c)}{\sum_{c=0}^{1} w_c \sum_{i=1}^{N} y_i^c + p(y_i^c)} \tag{3.8}$$

where, $w_c, c = 0, 1$ are weights for the object and background classes, defined as inversely proportional to label frequencies to achieve proper balance. It's important to note that these losses may lack stability for small denominators ($y_i \sim p(y_i) \sim 0$), necessitating the addition of a small ϵ to both the denominator and numerator for loss stability. This adjustment ensures numerical stability, particularly when dealing with cases where probabilities are extremely small.
Additional losses incorporate the use of the Jaccard index or the Twersky index. The Twersky index, employing different weights for false positives and false negatives, is particularly valuable in medical applications where minimising false negatives is often more critical than reducing false positives, allowing for a better balance between precision and sensitivity [88]. Regarding boundary losses, their focus is on precisely defining the object's boundaries and distinguishing between different objects. While these losses excel at providing sharp boundaries, they come with challenges, including the difficulty of representing boundary points as differentiable functions and oversensitivity to outliers, as seen in metrics like the Hausdorff distance. A more intriguing approach involves the use of combo losses, which integrate beneficial elements from different loss categories, achieving a balance in diverse qualities. One commonly employed combo loss combines the Dice loss with the weighted cross-entropy. This approach addresses class imbalance by assigning higher weights to the minority class in the weighted binary cross-entropy (wBCE), while the Dice loss enhances the classification of small objects:

$$\mathcal{C}_{COMBO} = \alpha \cdot \mathcal{C}_{DSC} + (1-\alpha)\cdot\mathcal{C}_{wBCE} \tag{3.9}$$

where, the parameter α controls the weights of the Dice term with respect to the cross-entropy term, and the weights within the cross-entropy define the penalty for each class. Several variations, including combinations with less common focal losses, are also present in the literature. These combo losses offer a versatile approach, leveraging the strengths of different loss components to enhance the overall segmentation performance.
The original implementation of V-Net by Milletari et al. [65] compared the use of the dice loss presented in equation 3.7 with a multinomial cross-entropy loss. In their specific case of prostate segmentation, they found that the dice loss with a squared denominator yielded better results. However, it's essential to note that these findings may not be generalisable to all segmentation tasks. Recent studies on CNN, as outlined by [58], highlight the prevalence of BCE and Dice loss in various applications, with numerous variations developed. The optimal choice of a loss function often depends on specific dataset characteristics and model architectures. The findings of [58], comparing different loss functions on various datasets, reveal that for whole organ segmentation with slightly unbalanced datasets, both BCE and Dice losses yield similar results. However, losses designed for highly unbalanced datasets may result in oversegmentation. On the contrary, for highly unbalanced tumour segmentation, compound losses incorporating a Dice term demonstrate superior performance.
Considering the segmentation pipeline described, the first stage involves a slightly unbalanced problem, where breast tissue occupies a significant fraction of the entire image. In this scenario, the benefits of using a Dice or compound loss are marginal. The choice has leaned towards the well-known BCE loss due to its faster convergence. BCE amplifies the gradient significantly when predictions and labels are very different, and its higher penalisation of errors encourages the model to change parameters, avoiding local minima.
In the second stage of the segmentation pipeline, the whole image is divided into many small patches, some containing only the background class, some containing only the lesion class, and others containing both classes in a more balanced manner. Since many images contain only a single class, using the Dice loss and its derivatives is not well-defined for patches with only background or only lesion. Therefore, the choice again favours BCE loss for its numerical stability. To further improve stability, a small correction is applied by adding a small ϵ term to the arguments of the logarithms to reduce divergences:

$$\mathcal{C}_{\widehat{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i) + \epsilon) + (1 - y_i) \cdot \log(1 - p(y_i) + \epsilon) \tag{3.10}$$

This adjusted BCE loss ensures numerical stability, particularly when dealing with cases where probabilities are extremely small. Additionally as an experiment, training with the dice loss using only lesion patches was also conducted.
Regarding model evaluation, measures outlined in Chapter 2 have been utilised alongside the dice score and the relative size difference with reference to radiologists' ground truths. In the initial stage of breast segmentation, to gain a thorough understanding of the training phase, both the cost function and performance metrics were assessed across the entire process for both the training and test sets. However, in the subsequent stage, given the insignificance of patch-based metrics and the computational expense of real-time reconstruction, only the loss function has been monitored, and evaluation is conducted after the final reconstruction of the lesion from all the tiles.

Optmiser

After selecting the optimal loss function, the next crucial consideration is the choice of the optimiser algorithm, responsible for dynamically adjusting the model parameters to minimise the loss function. This decision should factor in the inherent characteristics of the neural network and the landscape of the loss function within the parameter space. Additionally, computational cost and available memory

are essential factors to consider.
Computationally, employing backpropagation on the entire training dataset can be prohibitively expensive. Computing the cost function and its gradient for the entire dataset becomes slow, especially when the dataset is too large to fit into the main memory of a single machine. To address this challenge, stochastic gradient methods are widely employed. These methods aim to minimise the cost function by computing gradients on a small subset of training samples (ranging from a single sample to a mini-batch of a few samples) and updating the parameters accordingly. One of the most common stochastic gradient methods is stochastic gradient descent (SGD), which maintains a fixed learning rate at each iteration. Although several modifications with adaptive learning rates have been developed, the simple SGD is often a good choice. According to findings by [35], it minimises training time while keeping a small generalisation error, preventing overfitting. In practice, SGD appears to be robust across different domains, outperforming adaptive methods that are more prone to memorising the training set.
From a convexity standpoint, in the case of CNN, the combination of linear transformations and non-linear functions generates non-convex cost functions in the parameter space (weights associated with each stage). This complexity means that finding the global minimum, as in simpler models, is not always possible. A comparison between SGD and adaptive optimisers [120] suggests that SGD is better at escaping sharp local minima, preferring flatter ones with higher generalisation capabilities on the test set. SGD benefits from anisotropic gradient noise, allowing it to escape local bad minima, whereas adaptive optimisers adjusting the learning rate for each coordinate may fail. While theoretically, the use of random initialisation optimisers does not guarantee convergence to the global minimum, recent findings for fully connected feedforward neural networks [17], [16], [7] show that local minima are located in a well-defined band delimited by the global minimum. The number of bad-quality local minima outside this band diminishes with the increase in the size of the network. This implies that for large networks, the probability of falling into bad-quality minima is significantly reduced. Moreover, falling into the global minimum often leads to overfitting the dataset. For deep multilayer networks, there are typically many good local minima with equivalent performances on the test set. Additionally, [69] found that for generic activations, when wide layers are present, the loss function surface becomes well-behaved with local minima very close to the global one. Considering all these factors, SGD has been chosen as the optimiser for both training stages.

Summary of parameters

V-Net was implemented in Python, utilising the PyTorch (version 2.0.1) package for tensor computation with enhanced GPU acceleration and an automatic differentiation engine tailored for neural networks. The workstation, running Ubuntu 20.04.1 LTS (64 bit), utilised an Intel® Xeon(R) CPU E5-2630 v2 at 2.60GHz and a NVIDIA GPU GM204GL [Quadro M5000] with 8GB of memory. Preprocessing and postprocessing of images are carried out using the Python package SimpleITK (version 1.2.4). While tensor operations are executed using PyTorch, routine operations on low-dimensional arrays and basic statistical operations are performed using the Python package NumPy (version 1.24.4).
Training parameters for the two segmentation stages are detailed in Table 3.2.
In the first stage of breast segmentation, the original input size of [128, 128, 64] was utilised, albeit with significantly lower resolution due to GPU compatibility constraints. The input size was determined by memory requirements, with a fixed spacing chosen to accommodate all images within the standard size without cropping any regions. To preserve information, empty regions were padded with zeros rather than cropping the images. Additionally, normalisation was applied to obtain intensity values in the range [0,1]. Batch sizes were reduced due to high memory requirements during training, limiting the number of images that could be loaded simultaneously. However, effective mini-batches were used by exploiting gradient accumulation. Readers are directed to Appendix B for additional practical details regarding the implementation. The learning rate was kept high for quick convergence

since the primary goal of this stage was raw segmentation. Although a learning rate schedule with progressively lower rates could achieve finer results, it was not implemented in this scope. The number of training epochs is determined by implementing early stopping before the model begins to overfit the training data. Given the relatively small dataset, reserving a portion for validation could significantly affect the model's performance. Hence, the approach suggested by [31] is adopted. The training set is initially partitioned into a training subset (80%) and a validation set (20%). This partition helps determine the optimal number of epochs to prevent overfitting. Once this optimal number is identified, the entire training dataset is utilised for the fixed number of epochs. Implementing early stopping not only prevents overfitting but also reduces the computational burden compared to alternative methods such as regularisation terms. Approximately 150 epochs were necessary to achieve sufficiently good results, and the entire procedure was then tested using 5-fold cross-validation. Post-processing involved upsampling to recover the original spacing and size. To obtain smoother images, binary dilation with a kernel of [3,3,3] voxels was performed, followed by Gaussian smoothing with a kernel of [5,5,5] voxels. Additionally, a check on the number and size of connected components was conducted. Typically, one or two components were identified, depending on whether the left and right breasts were merged. However, small noisy blobs sometimes appeared, so components with sizes less than 10% of the maximal component size were removed.

	Breast segmentation	**Lesion segmentation**				
		lesion only	lesion + 30% bg.	lesion +50% bg.	lesion +75% bg.	lesion +100% bg.
input size	(128, 128, 64)	patches of (32, 32, 32) extracted from the full image of (416, 256, 192)				
image spacing	(3.3, 3.3, 3.1) mm	(1,1,1) mm				
loss function	BCE	DSC	BCE			
optimiser	SGD (momentum=0.9)	SGD (Nesterov momentum)				
learning rate	0.01	0.01	0.001	0.005	0.005	0.005
batch size	16	16	64	128	128	128
epochs	150	150	100	100	100	70
pre-processing	min-max normalisation, 0 padding to fill standard size	min-max normalisation, 0 padding for values outside the breast mask				
augmentation	no	use of patches				
post-processing	upsampling, binary dilation [3,3,3], Gaussian smoothing [5,5,5], eventual removal of small disconnected components ($< 10\%$ $V_{max\ comp.}$)	no				
binarisation threshold	0.5	0.5	0.5	0.5	0.5	0.3-0.6

Table 3.2: Training parameters for the segmentation models

Regarding the second stage of lesion segmentation, similar preprocessing steps are undertaken. Intensity normalisation is applied, and standard spacing is fixed to unitary spacing. Values outside the breast mask are set to 0, and padding is applied when necessary. Due to memory constraints, various approaches were initially explored. Initially, raw spacing was attempted, but this led to the loss of fine-grained details and small lesions, rendering it impractical. Another approach involved the separation of the left and right breasts. However, with this strategy, aside from having one side with empty ground truth, the lesion volume fraction remains very small compared to the entire breast. Consequently, convergence is notably slow, even when utilising focal or Tversky losses. For these reasons,

those approaches were soon abandoned in favour of adopting a small patch approach to better balance resolution and computation time. The use of smaller patches improves the ratio of lesion volumes to the search area. Additionally, patches containing only a small margin of the lesion on the borders (less than 2-4% of the selected volume) were excluded during training to improve convergence and avoid information loss on borders. However, this exclusion does not pose any issues, as when the model is applied, the different grids of patches will cover the lesion in a more central area of a patch. After defining the patch size of [32,32,32], different experiments were conducted:

- training with only the patches containing the lesion using the Dice loss (using the improved version proposed by Milletari et al.[65]);
- training with all lesion patches and 30% of the background patches with BCE loss;
- training with all lesion patches and 50% of the background patches with BCE loss;
- training with all lesion patches and 75% of the background patches with BCE loss;
- training with all lesion patches and 100% of the background patches with BCE loss.

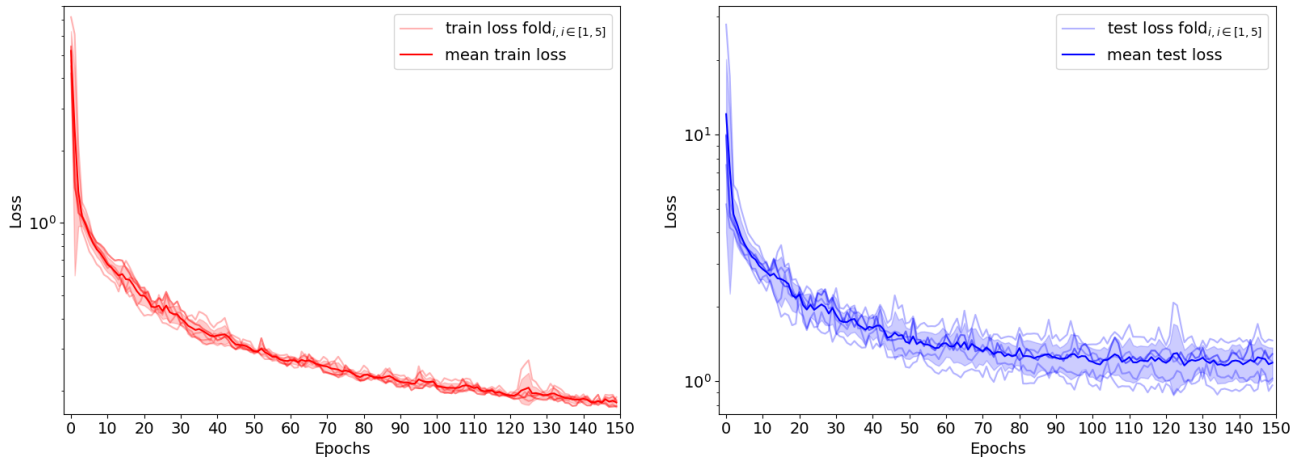
The learning rate, number of epochs, and batch size for gradient accumulation were determined by adapting to the sample size, using the same strategy employed for breast segmentation. At this stage, to manage computational times required for test various attempts, a single division into train and test sets was performed, with 104 images for the training set and 27 for the test set. Cases were randomly extracted, balancing ground truth from the two radiologists. Overall, the decision to use patches also mitigates the risk of overfitting, as from a small dataset of hundreds of images, one can obtain up to twenty thousand patches, serving as an intrinsic source of data augmentation without distorting the image. However, there is a risk of losing global context information, which must be considered in the reconstruction stage.
Finally, for the best-performing option among the various attempts, the effect of using different binarisation thresholds was investigated. This step is crucial for fine-tuning the segmentation results and optimising the model's performance.

3.4 Segmentation results

As follow we report the results of the first stage of segmentation and of the various experiments in the second stage.

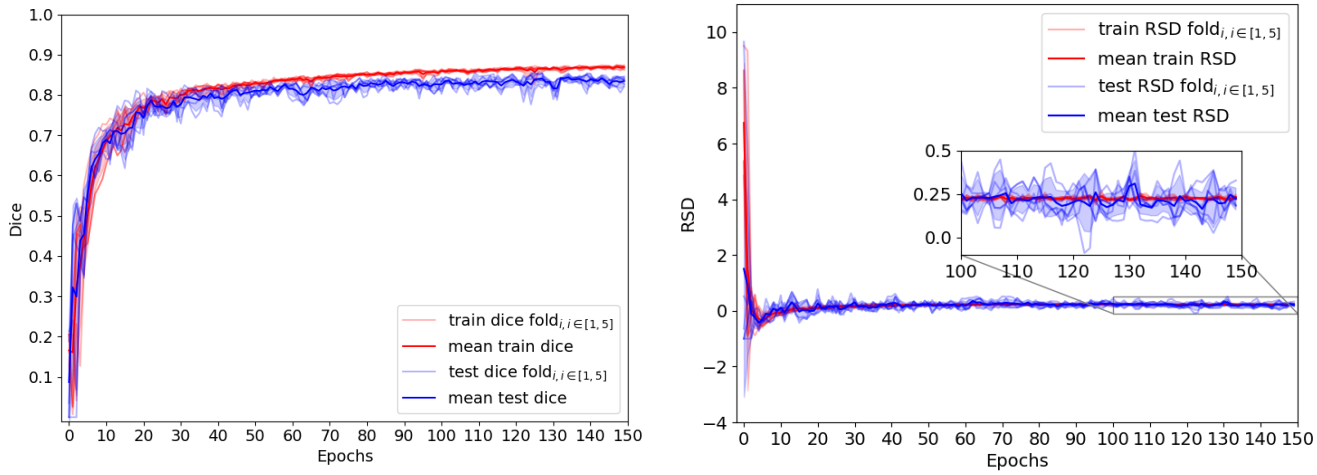
3.4.1 Breast segmentation

Regarding breast segmentation, the training process was conducted for 150 epochs, as determined by the early stopping procedure (training and validation loss were monitored, revealing that the validation loss ceased to decrease and began to plateau at approximately 150 epochs). Additionally, to ensure that no overfitting occurred, the loss function was examined in the final training stage. Figures 3.7 (A) and (B) depict the train and test losses, respectively, for the 5-fold cross-validation process.
Considering concrete performance measures, the Dice score was also evaluated during training, with results depicted in Figure 3.8 (A), while the relative size difference is presented in panel (B). Overall, the Dice score quickly rises above 0.8, indicating good performance on both the training and test data across all folds, suggesting that the model produces masks similar to those of the operators. Analysing the trend of the relative size difference in the later epochs reveals a systematic positive direction. At this stage, the positive trend is not concerning, as a larger breast mask that includes more tissues poses no issue. Conversely, the absence of tissues would have been problematic, as it could lead to missing potential lesions in subsequent stages. This emphasises the significance of evaluating not only the overall segmentation accuracy but also the impact on the representation of crucial anatomical details.



(A) Train Loss (B) Test Loss

FIGURE 3.7: In panels A and B, the darker solid line represents the mean value of the training and test losses, respectively, across five folds. Lighter lines depict the individual losses for each fold, while the shaded region illustrates the standard deviation.The test loss continues to decrease, albeit with a lower slope, indicating that no overfitting occurred. This observation suggests that further improvements could potentially be achieved with a smaller learning rate.

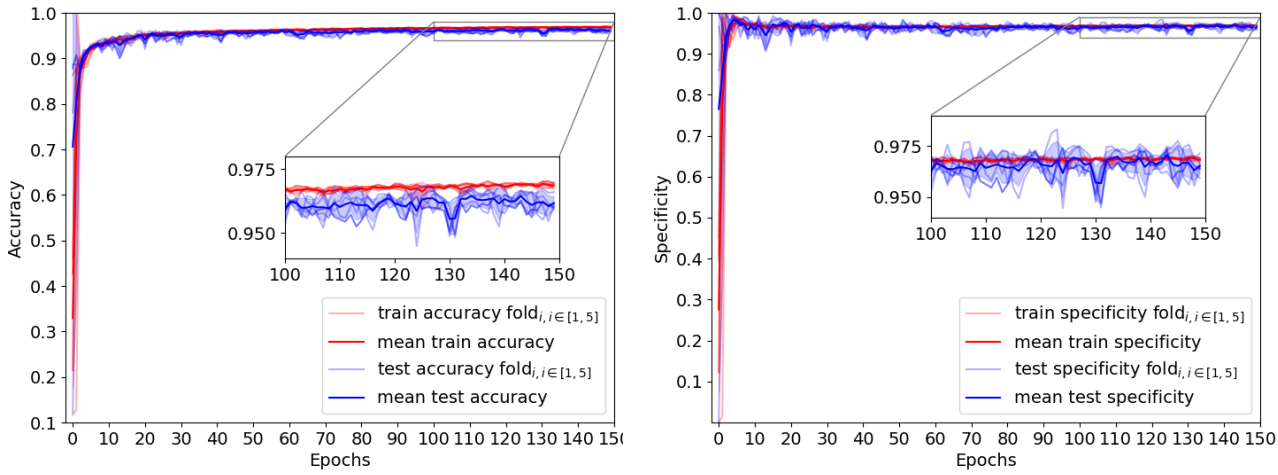


(A) Dice (B) RSD

FIGURE 3.8: In plots A and B, the solid darker red and blue lines represent the average values over five folds for training and test sets, respectively. Lighter lines depict individual fold values, while shadows denote the one standard deviation range. In the left plot (A) showcasing the Dice coefficient, a rapid increase is observed in both train and test Dice during the initial 10 epochs, reaching approximately 0.8. Subsequently, the increase slows down, and while the Dice stabilises in the test set. Fluctuations in the test set are higher, partly attributed to the smaller dataset. Moving to the right plot (B) with RSD, the metric fluctuates widely in the first 10 epochs, displaying outputs both larger and smaller than the ground truth, eventually stabilising within a more suitable range [-1, 1]. In the last 50 epochs, zoomed in, the RSD consistently becomes positive, around 0.25, for both train and test sets.

Also, accuracy and specificity were monitored to assess the model's ability to correctly identify positive and negative instances. Their trends are reported in Figure 3.9 (A) and (B) respectively. Overall, a

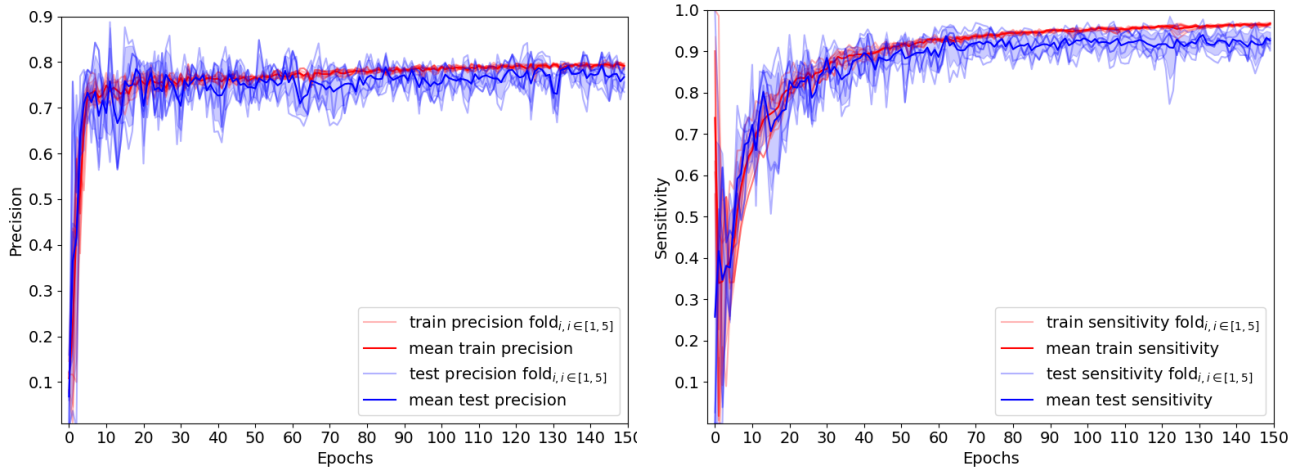
high specificity indicates the model's proficiency in correctly classifying non-target instances, which is particularly important to minimise false positives.



(A) Accuracy (B) Specificity

FIGURE 3.9: Plots A and B showcase classification metrics focusing on correct classifications, respectively accuracy and specificity. The solid darker lines represent the average values over five folds for training and test sets, while lighter lines indicate individual fold values. Shadows denote the one standard deviation range. One can notice that after about 10 epochs both accuracy and specificity reach quite high values overcoming 0.9. The zoomed plots show the details of the very last epochs, as far accuracy is concerned the train values seems to be still increasing and slightly higher than the test one, whereas the specificity is pretty similar in both the cases.

In addition, precision and sensitivity were analysed, as shown in Figure 3.10 (A) and (B). Precision reflects the false positive rate, which in this case does not reach very high values, indicating the presence of some false positives. This observation aligns with the positive trend observed in the relative size difference, indicating more abundant segmentations. However, achieving high sensitivity is more crucial to capture all the relevant context for further lesion analysis in the subsequent stage.



(A) Precision (B) Sensitivity

FIGURE 3.10: Plots A and B delve into precision and sensitivity, metrics essential for evaluating classification quality. As before the solid darker lines represent the average values over five folds for training and test sets, while lighter lines indicate individual fold values. One can notice that precision never reaches 0.8 meaning that false positives are still present both in training and test data. On the contrary sensitivity tends towards higher values above 0.9 enhancing the fact that false negatives are a little component.

Overall, the V-Net architecture facilitated a straightforward training procedure even with a small dataset. The final testing of the generalisation performance utilised the models obtained at the last epoch of each fold. Since this step is a simpler task in this case, the performance metric shows a nearly Gaussian distribution with no large tails at very low values, as observed for lesion segmentation by the operators. For each fold, mean performances[1] were evaluated and averaged to obtain a final cross-validated estimate. A summary of the results is provided in Table 3.3, showing the mean and its error along with the standard deviation across the different folds.

measure	**train**		**test**	
	mean $\pm\sigma_{mean}$	σ_s	mean $\pm\sigma_{mean}$	σ_s
Dice	0.868 ± 0.002	0.004	0.835 ± 0.005	0.01
accuracy	0.9686 ± 0.0005	0.001	0.962 ± 0.001	0.003
sensitivity	0.967 ± 0.001	0.002	0.927 ± 0.009	0.02
precision	0.791 ± 0.002	0.005	0.77 ± 0.01	0.02
specificity	0.9681 ± 0.0005	0.001	0.965 ± 0.002	0.004
RSD	0.232 ± 0.004	0.008	0.22 ± 0.03	0.06

TABLE 3.3: Evaluation metrics summary, featuring mean and standard deviation over 5 folds, obtained at the last epoch for both training and test sets

The most important metric overall is the Dice score, which is approximately 0.835 ± 0.005, accompanied by a positive relative size difference of 0.22 ± 0.03. In general, discrepancies between train and test results are quite small, indicating good generalisation.
To visually assess the efficacy of the automated breast mask segmentation using V-Net, we present three illustrative cases in Figures 3.11 and 3.12, providing an overview of the axial and sagittal planes, respectively. In each set of images, the sequence from left to right includes the second subtracted image (the original image provided in input at the model), the green segmentation representing the ground truth used in training, and the red mask depicting the output of the V-Net on the test set.
Case A: This example exhibits one of the less favourable outcomes, indicated by a lower Dice ($DSC < 0.7$). While it slightly underestimates the breast extension on the axial plane, a closer inspection of the sagittal plane reveals increased precision on the borders compared to the original manual segmentation.
Case B: Demonstrating an average outcome with a good Dice score ($DSC \sim 0.8$), the resulting mask tends to be slightly larger than the manually segmented case. This scenario is desirable as it mitigates the risk of losing vital information.
Case C: Marking an almost perfect agreement with $DSC > 0.9$, this case showcases a high level of similarity between the reference and the model output. Additionally, the automated segmentation appears smoother due to postprocessing. A subtle distinction lies in the fact that the original segmentation consists of two separate components, while the V-Net tends to merge them into a single piece.
In summary, the performance of V-Net in generating breast masks demonstrates great promise. Potential enhancements could arise from implementing data augmentation techniques, such as incorporating rotated images, contrast variations, or adding noise to increase the training sample size. However, given the primary focus of the study and the computationally-intensive nature of training a CNN with an extensive image dataset, only the original images were utilised.

[1] in this case, mean values are reported instead of median values since the data distribution showed that the two values were almost coincident.

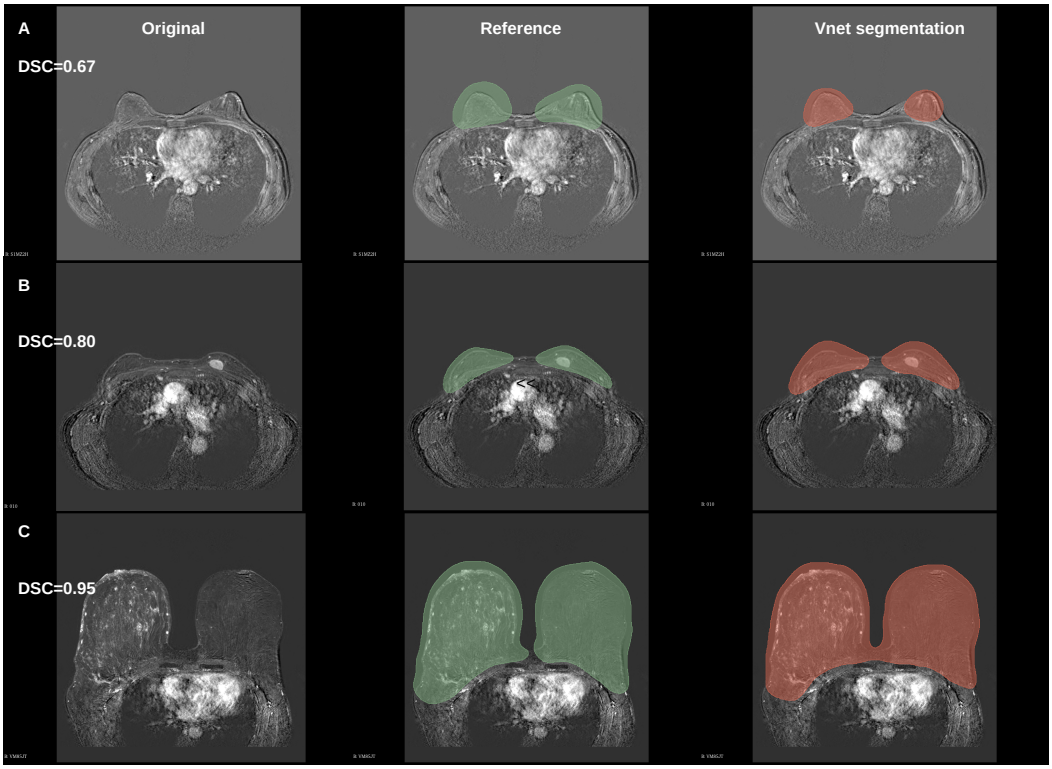


FIGURE 3.11: Examples of the obtained breast masks, axial view

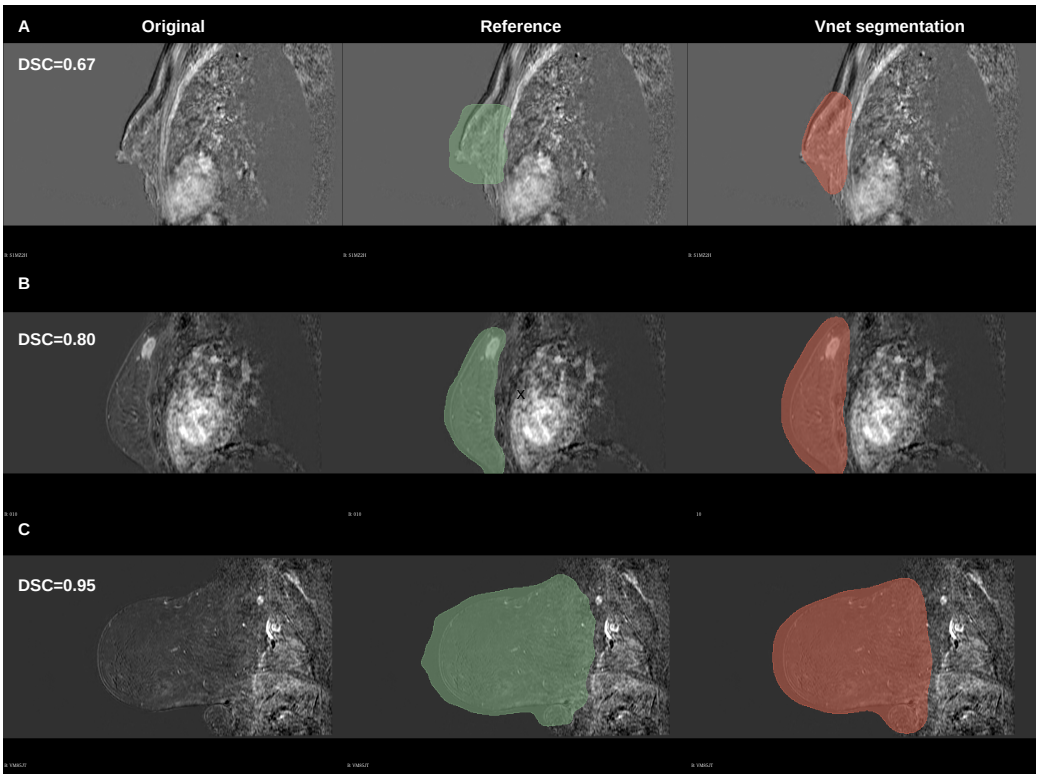
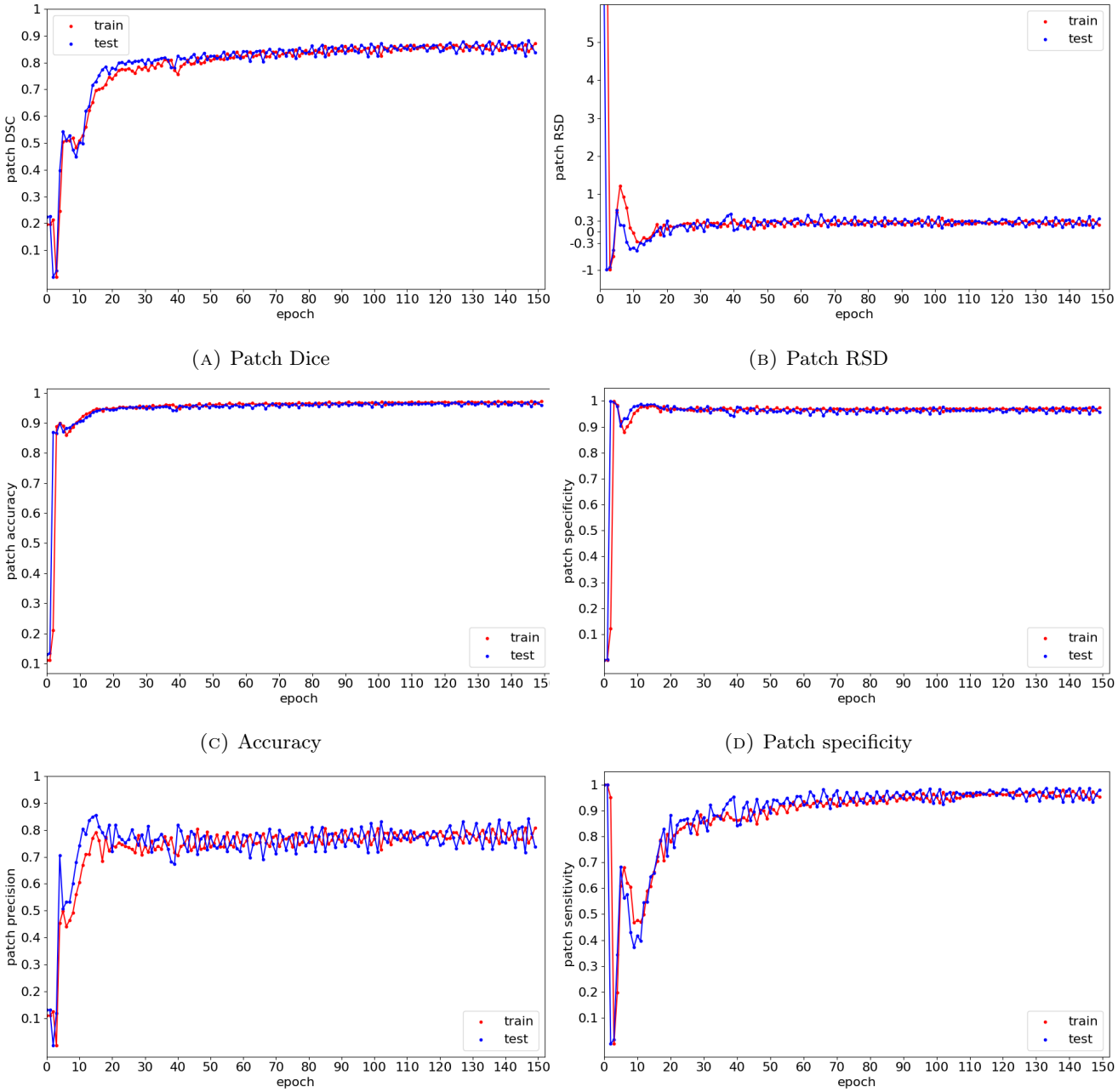


FIGURE 3.12: Examples of the obtained breast masks, saggital view.

3.4.2 Lesion segmentation results

The first attempt at lesion segmentation involved training using only lesion patches and the dice loss, a common approach observed in many studies presented in the literature review. Early stopping was applied as the model converged quickly due to the small dataset size. When applied to lesion patches, the model achieved high performance measures as reported in Figure 3.13, with a mean test Dice score of about 0.84, as well as high accuracy (~ 0.96), specificity (around 0.96), and sensitivity (~ 0.98). However, precision was slightly lower (~ 0.74), indicating the presence of false positives. Similarly, the relative size difference was around 0.35, further highlighting the impact of false positives.



(A) Patch Dice

(B) Patch RSD

(C) Accuracy

(D) Patch specificity

(E) Patch precision

(F) Patch sensitivity

FIGURE 3.13: Patch based metrics for the case with only lesion patches.

These promising results are valid only when the precise location of the lesion is known. When the model was applied to all patches, including those without any lesion, it produced a significant number of false positives, including parenchymal enhancements, the nipple area, lymph nodes, and tissue borders. Consequently, the model was unable to generalise to unseen tissues.

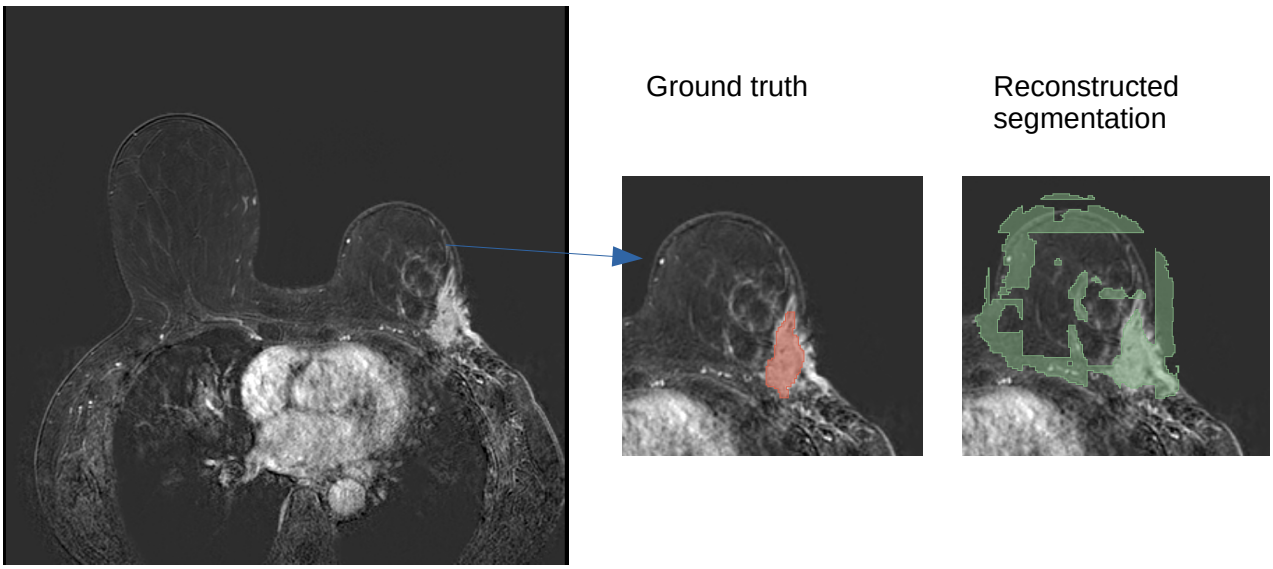


FIGURE 3.14: Example of a poor segmentation outcome. The ground truth provided by the radiologist is depicted in red, while the proposed segmentation is shown in green. Here the model incorrectly identifies breast borders and slightly enhanced parenchyma as lesions, resulting in a non-useful segmentation outcome, despite accurately capturing the main lesion.

Full reconstruction on all images was performed by applying the model to all patches, resulting in an excessive number of false positives and a low Dice score (around 0.05), precision ($<$0.03), and really high RSD ($>$80). An example illustrating the disparity between the ground truth and the main lesion is shown in Figure 3.14. Even with a high overlap, this model is not directly applicable to the whole image. A generalised version of the dice loss for background-only patches was also tested but proved to be unstable. One possible improvement for this approach could involve using a preceding network capable of classifying patches and selecting only those that are candidates for lesions. This would help refine the segmentation process and reduce false positives.

After this unsuccessful attempt, the focus shifted to approaches involving the utilisation of background information and selecting a loss function capable of handling background-only patches more effectively.
When employing the BCE loss function and incorporating background patches during training, more promising results were achieved. As observed previously, evaluating individual patches was not meaningful since the entire image needed to be reconstructed. Various attempts were made, including increasing the percentage of background patches. The main evaluation metrics on the test set after the whole reconstruction procedure are depicted in Figure 3.15 as a box plot due to the skewed distribution of the data.

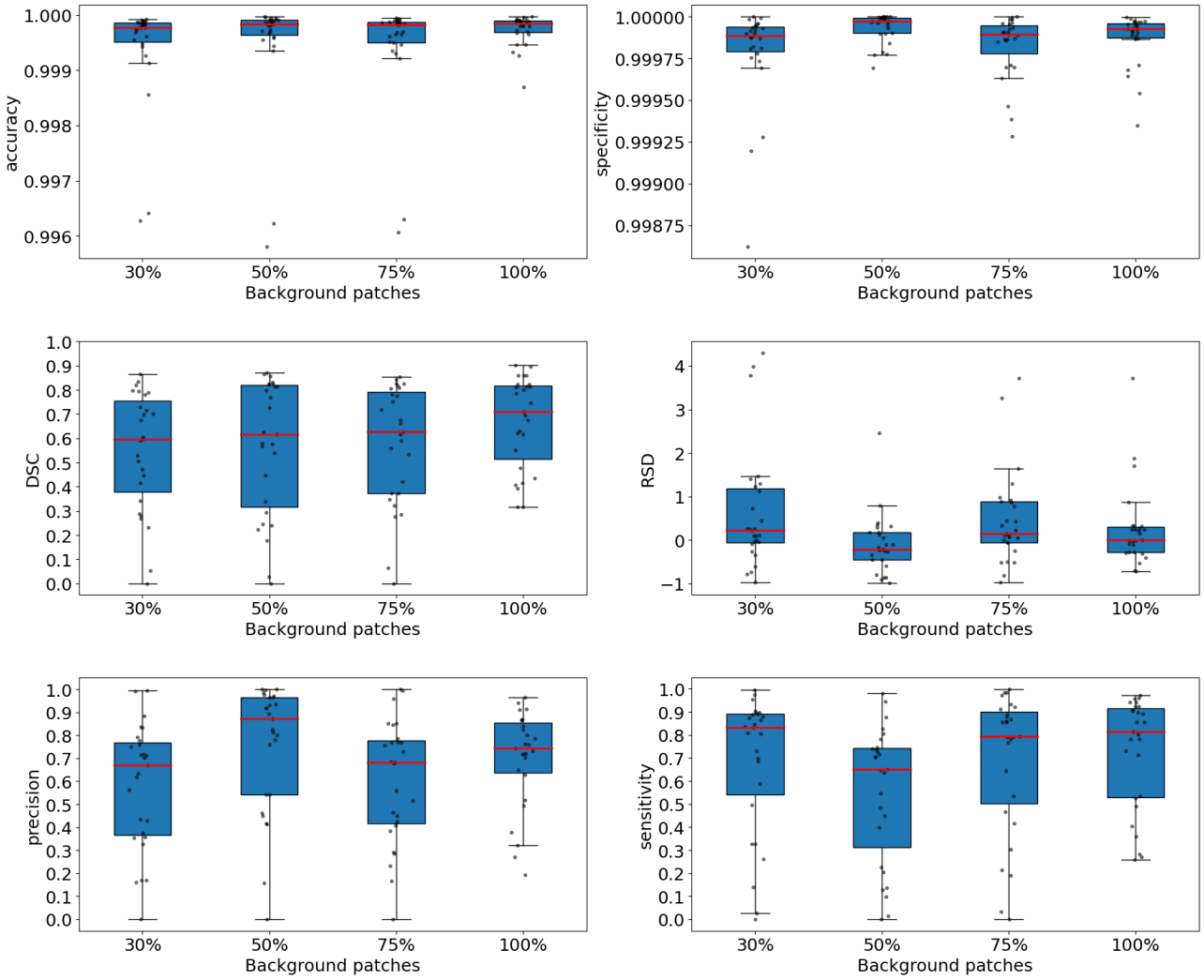


FIGURE 3.15: Box plot illustrating the main evaluation metrics on the test set, varying the percentage of background patches. The median values are highlighted in red, with the box representing the interquartile range and whiskers denoting the minimum and maximum values excluding outliers. Overall, there is a trend of improvement with the increase in background patches, leading to a better balance between false positives and false negatives.

Overall, the issues related to non meaningful false positive observed in previous attempts are no longer present, with the median dice score falling within an acceptable range between 0.6-0.7. Notably, when using a lower percentage of background patches (30%), all performance metrics are lower, particularly median precision and RSD, which include larger outlier values. Increasing the background percentage to 50% results in a significant improvement in precision, albeit at the expense of a corresponding decrease in sensitivity, leading to a similar overall dice score and a more contained RSD. With 75% of background patches, precision decreases while sensitivity increases, resulting in an overall positive trend in RSD. Finally, using 100% background patches yields the best performance on many fronts, with fewer outliers and a more balanced RSD around 0, thus achieving a better compromise between precision and sensitivity and obtaining a higher overall accuracy. An example of the proposed segmentation for the model with different training background patches is presented in Figure 3.16, demonstrating the best performance when using all background patches.

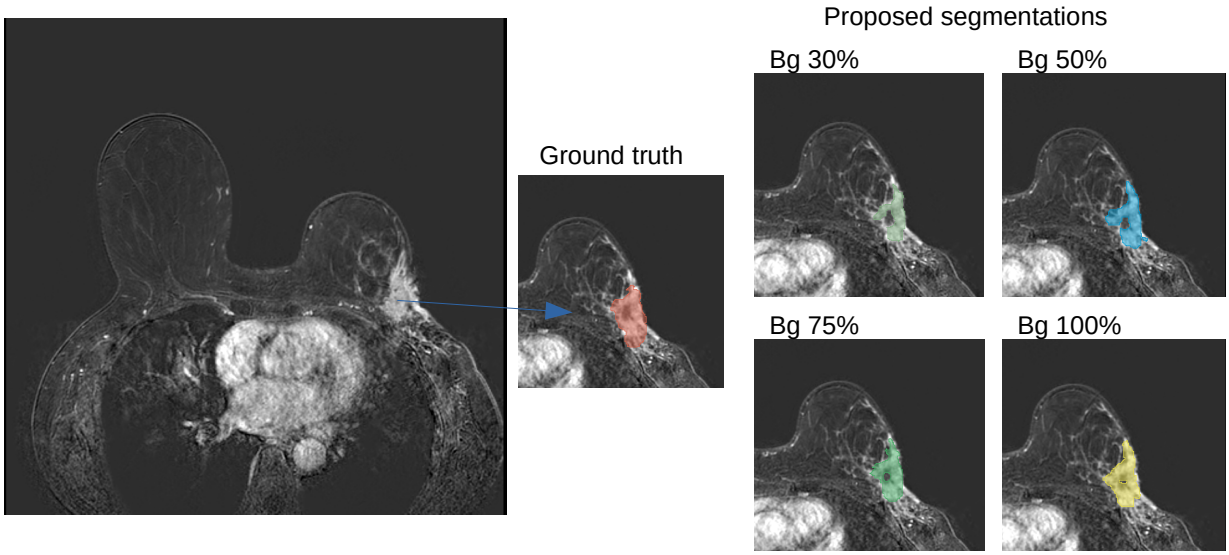
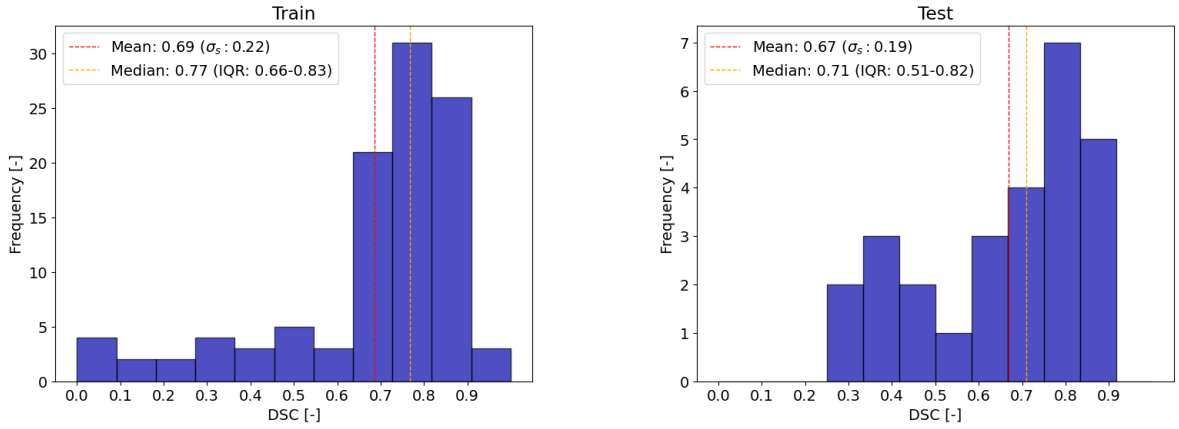


FIGURE 3.16: Example of proposed segmentation by the models obtained using different percentages of background patches.

The utilisation of 100% of background patches enables the utilisation of the entire information contained within each image, allowing for a higher number of training examples and enabling the CNN to fully exploit its potential. However, the use of multiple patches also increases training time accordingly, and the lesion information might be underweighted due to a high imbalance in patch numbers. Further improvements could be made in terms of data augmentation, particularly by increasing the number of lesion patches. While the full background use emerged as the better case, aside from mere segmentation metrics, one should take a closer look at data characteristics to fully understand the limitations and potentialities.



(A) Histogram of train dice for the case with the 100% of background patches.

(B) Histogram of test dice for the case with the 100% of background patches.

FIGURE 3.17

One can investigate the distribution of Dice scores on reconstructed images from both the train and test sets, as depicted respectively in Figure 3.17b (A) and (B). A main peak at high Dice values appears in both cases around 0.8, whereas a tail and small peaks at very low values are present. After a visual inspection of all the cases, multiple explanations for lower Dice scores were found. Generally, larger tumours with homogeneous appearances tend to yield better segmentations, achieving higher DSC values around 0.8. Conversely, in cases of fragmented and irregular tumours, some smaller portions

may be overlooked, particularly within inner darker areas where reduced enhancement is present. Moreover, lower Dice scores may occur when there is strong parenchymal enhancement due to different breast characteristics or strong field inhomogeneities. Such cases are quite rare in the dataset, so the model is not able to capture such information. Additionally, the use of patches loses the global context, whereas a network exploiting the full image could employ symmetry to check the bilateral breast. Additionally, lower scores may come from findings of other axillary lymphadenopathy, which is also present in very few cases of the dataset, as in the case depicted in Figure 3.18. However, this is a sign often associated with breast cancer, and perhaps a more extensive dataset could enable discrimination between them.

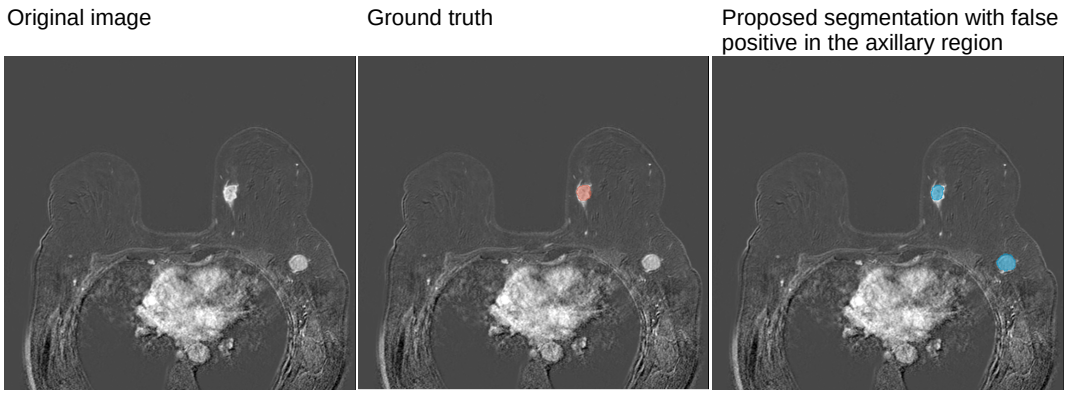


FIGURE 3.18: Example of segmentation with false positive findings in the axillary area.

Furthermore, lower scores and false positives may arise from the V-Net sometimes discovering suspicious areas that, after a crosscheck from the radiologist, were actually recognised as mislabelled tumours that were missed in the manual segmentation, as in the case depicted in Figure 3.19.

FIGURE 3.19: Example of a proposed segmentation discovering fragments missed by the radiologist

The findings align with those obtained in tumour segmentation across different organs, as demonstrated by Ma et al. [58], where DSC values for lesion segmentation exhibit a bimodal distribution, with modes near 0.8 and 0.1. Notably, results for individual lesion segmentation showcase a polarisation compared to the segmentation of entire organs or tissues. This dichotomy arises due to several factors, including the limited size of the training set and the diverse appearance, locations, shapes,

and sizes of tumours, as well as individual variations in background tissue characteristics. Overall, considering the non-straightforward task, results can still be considered good.
In the end, keeping the model trained with all the background patches as the preferred one, an investigation on the effect of the binarisation threshold was performed. A more nuanced evaluation was conducted using probability thresholds of 0.3, 0.4, 0.5, and 0.6, demonstrating discrete variability. Box plots for all the test metrics are presented in Figure 3.20.

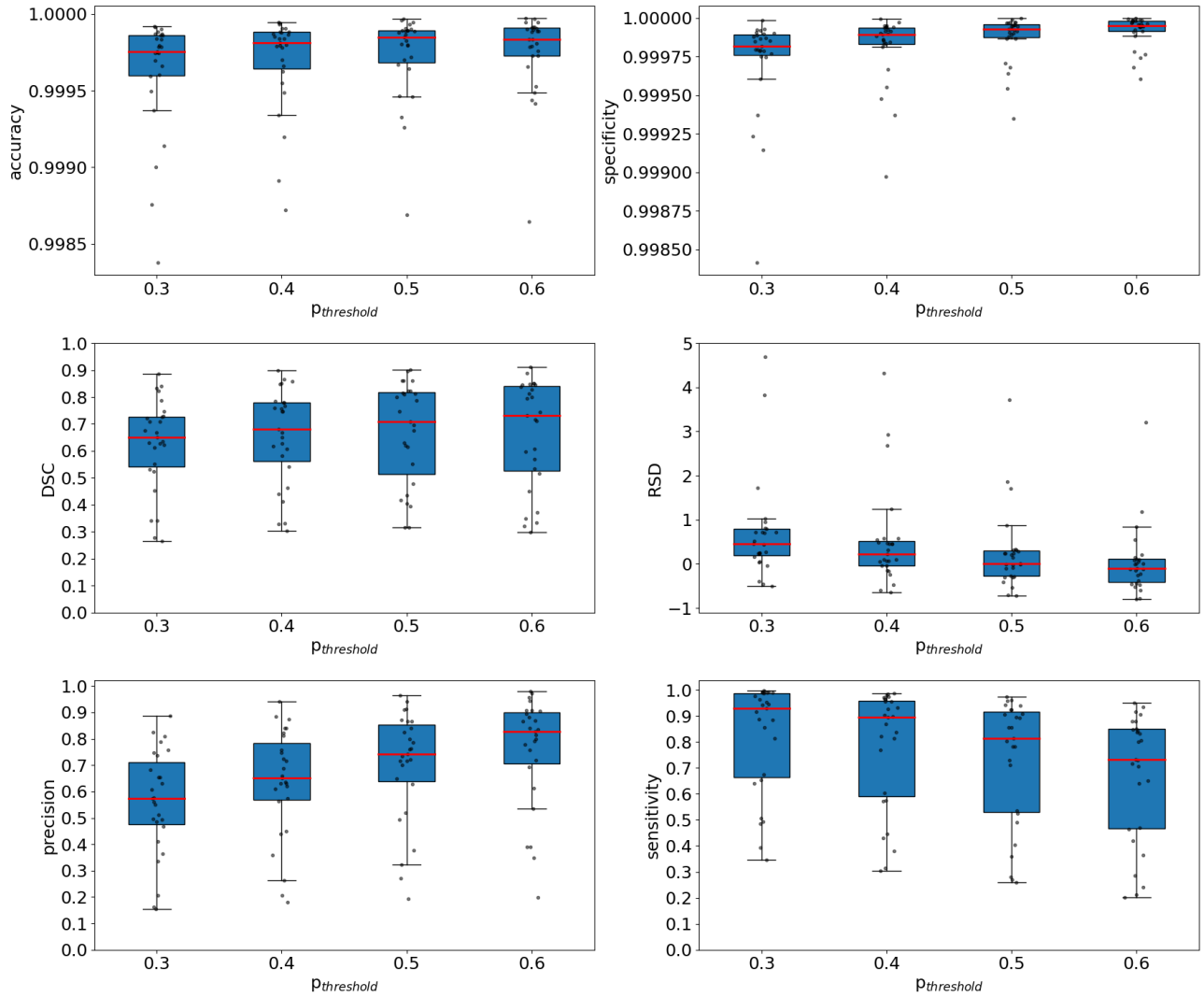


FIGURE 3.20: Box plot illustrating the main evaluation metrics on the test set, varying the binarisation threshold.

Concerning accuracy and specificity values are consistently high and increase with the increase in thresholding due to false positive reduction, but given the highly unbalanced dataset, accuracy alone provides limited information. Precision and recall offer more insights, with lower threshold values providing higher recall and lower precision, and vice versa for higher thresholds. An intermediate threshold in general gives balance between precision and recall. Dice score and relative size difference offered additional information. Lower thresholds resulted in lower Dice scores and a positively biased RSD, indicating overestimated segmentation volume or potential false positives. Higher thresholds showed a negative bias on the RSD, suggesting volume underestimation. However visual analysis suggested that a single fixed threshold might not universally suit all cases. Figure 3.21 illustrates

examples of segmentations for different thresholds. For cases with small, regular lesions, the threshold value minimally affected segmentation shape. Conversely, in cases with irregular shapes and less enhancement, a lower threshold was necessary to capture intricate details.

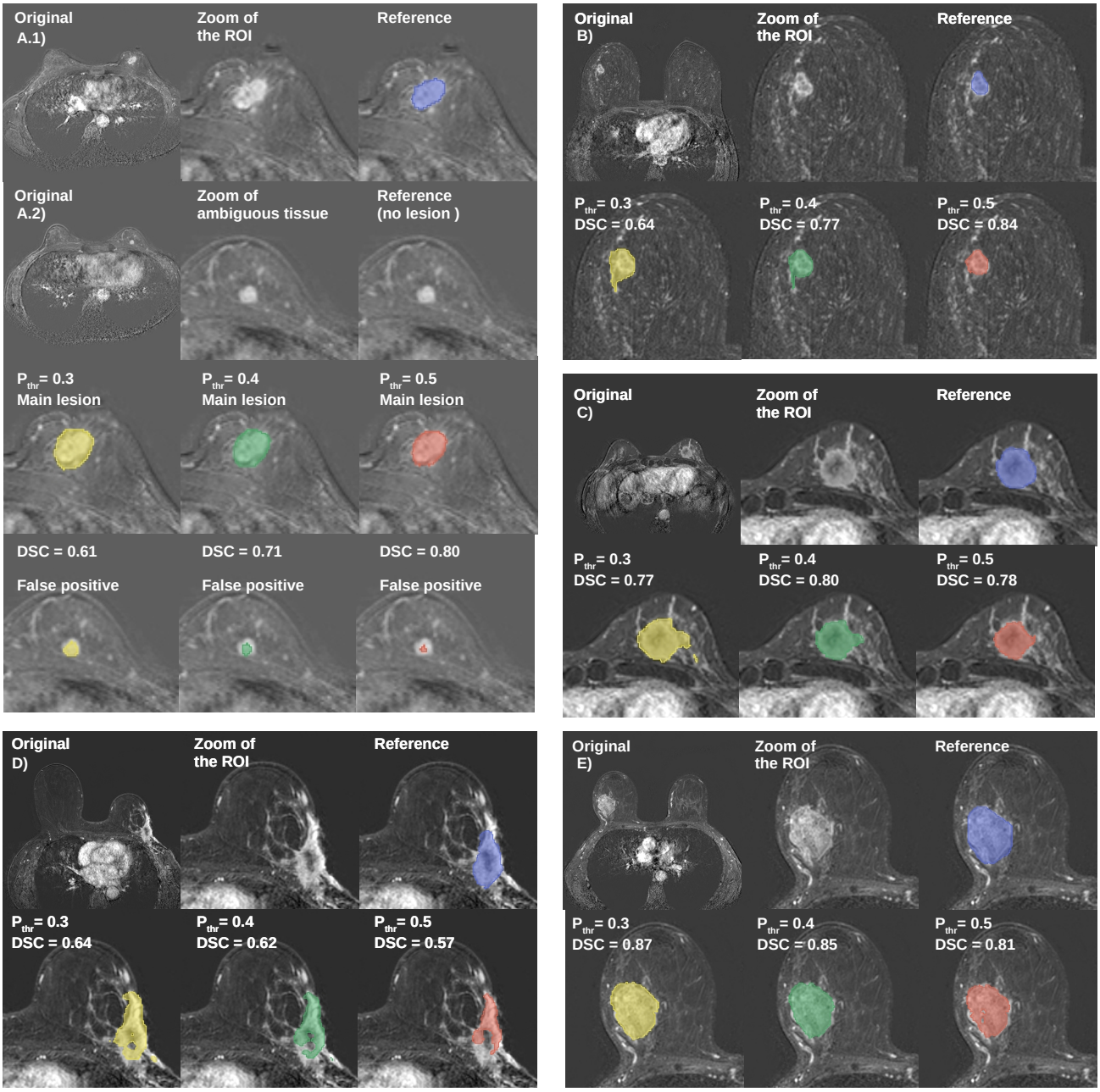


FIGURE 3.21: Examples of segmentation obtained using different binarisation thresholds.

To conclude, a summary of the main results of the successful attempts in terms of Dice score is presented in Table 3.4. The best model achieved a median Dice score of 0.73 (IQR = 0.53-0.84).
Comparing this with the evaluated intraobserver variability estimated previously i.e. median DSC 0.79 (IQR = 0.70-0.85), one can see that the result is slightly smaller, but still comparable, considering the

intrinsic limitations of the small sample. Comparison with results from other studies is challenging due to differences in datasets and evaluation procedures.

	train median DSC	IQR	test median DSC	IQR
bg30 (p_{thr}: 0.5)	0.65	0.42-0.77	0.60	0.39-0.76
bg50 (p_{thr}: 0.5)	0.61	0.33-0.78	0.62	0.32-0.82
bg75 (p_{thr}: 0.5)	0.72	0.46-0.80	0.63	0.37-0.79
bg100 (p_{thr}: 0.3)	0.75	0.62-0.80	0.65	0.54-0.73
bg100 (p_{thr}: 0.4)	0.76	0.66-0.82	0.68	0.56-0.78
bg100 (p_{thr}: 0.5)	0.77	0.66-0.83	0.71	0.51-0.82
bg100 (p_{thr}: 0.6)	0.76	0.62-0.84	0.73	0.53-0.84

TABLE 3.4: Summary of the performances in terms of Dice score

Nevertheless, the obtained results are satisfactory, with potential for improvement through increased dataset size, data augmentation, and architectural modifications. Additionally, exploring variations in model architecture and loss functions might further enhance performance, alongside the adoption of more powerful hardware capable of handling larger, high-resolution patches to capture more global information. The proposed automated segmentation achieved its primary goal of reducing human operator variability and providing a time-efficient segmentation. However, the results still require human supervision to overcome false positive detections and to fine-tune the threshold if required.
Indeed, both human segmentation and automatic segmentation introduce a degree of variability, whether from human experience or the choice of thresholding. It is crucial to acknowledge this variability in all subsequent stages involving feature extraction for predictive model building. This recognition ensures that the inherent uncertainties are appropriately considered and accounted for in the development and validation of predictive models, leading to more robust and reliable results.

Chapter 4

Development of the predictive model

Following the assessment of operator variability and the investigation into the automation of the segmentation stage, radiomic features are poised for extraction from the region of interest, ready for the development of a predictive model. Constructing such a model using radiomic data presents significant challenges. Image information can be enhanced through the application of filters prior to feature extraction, while different discretisation levels can yield a high number of features. Notably, these challenges stem from the limited availability of observations, primarily due to the absence of standardised clinical data collection for certain pathologies. When faced with a scenario where the number of observations is considerably smaller than the number of predictors, there is a notable escalation in the risk of overfitting and susceptibility to the curse of dimensionality. Consequently, implementing various strategies aimed at dimensionality reduction becomes imperative. Feature selection and feature engineering emerge as potential solutions; however, the existence of highly correlated predictors introduces the risk of multicollinearity with the outcome, potentially compromising the stability of any procedure. The absence of universally applicable standards for any given dataset underscores the need for the exploration and implementation of diverse approaches. Dimensionality reduction techniques manifest in a range of forms, encompassing both supervised and unsupervised methods. In either case, adopting a common framework that integrates model assessment and selection alongside these techniques is advisable. In an effort to address these challenges, this chapter will present a workflow for developing a robust predictive model.
Initially, feature robustness will be assessed and employed to select a suitable discretisation level, followed by a first stage of dimensionality reduction. Subsequently, filter statistical methods such as analysis of variance and correlation will be employed to further reduce features. Different models (logistic regression, Random Forests, and SVM) and further dimensionality reduction techniques (mRMR, LASSO, UDFS, PCA) will be used to predict the NAC outcome and will be tested and compared. Furthermore, the exploration will extend to integrating clinical variables with the selected radiomic features. For each combination of feature selection and classifier, models will be trained using features extracted from radiologists' ground truth. The performance of these models will be assessed and tested also using features obtained from the automatic segmentation provided by V-Net developed in the previous chapter.

4.1 Dataset

Recalling the study's objective to construct a supervised model for predicting the NAC outcome, this stage involved the same cohorts of women who underwent DCE-MRI prior to NAC administration, as presented in Chapter 3.1 for the segmentation analysis. Before NAC administration, women also underwent lesion biopsies to obtain histopathological information, including HER2 expression, ER, PR, and Ki-67 levels, to tailor therapy accordingly. The NAC regimen consisted of 4 cycles of anthracyclines and cyclophosphamide administered every three weeks, combined with sequential taxane administration once a week for 12 weeks. For patients positive for HER2, taxanes were administered in combination with Trastuzumab for one year. Subsequent to conservative surgery or mastectomy,

therapy outcomes were assessed via biopsy of the sentinel lymph node. Patients were classified as either pathological complete responders (pCR), corresponding to Pinder class 1, or non-complete responders, corresponding to all other Pinder classes outlined in Table 1.1. In addition to outcome information, clinical data prior to NAC administration were available, including patient age and semantic features defined by radiologists such as tumour grade, tumour margins (irregular or spiculated), and tumour type (multicentric, multifocal, or unifocal). In total, the study involved 127 cases (fewer than for segmentation, as outcome information was unavailable for some patients). A summary of the clinical characteristics as a function of the response to NAC is presented in Table 4.1. Predictive models can be constructed using radiomic information alone or in combination with clinical factors.

	pCR	**non-pCR**	**all**
Number	55	72	127
Age	51 (± 10)	55 (± 12)	53 (± 11)
Grade			
2	6	37	43
3	47	35	82
N/A	2	0	2
Margin			
irregular	36	37	73
spiculated	19	35	54
Type			
multicentric	7	18	25
multifocal	17	13	30
unifocal	31	41	72

TABLE 4.1: Clinical characteristics as a function of the response. Data are presented as the number of patients per category and as the mean with standard deviation for age.

4.2 Feature extraction

Once the images have been segmented, features can be extracted from the ROI. In this case, both dynamic and subtracted images were used, as using only the subtracted image could remove relevant information not directly linked to contrast agent uptake, especially concerning textural features, as evidenced in other works [67]. According to the most informative time point for malignant lesions, features are extracted from the third dynamic image and the second subtracted image.
Features are obtained using an open-source IBSI-compliant software package, PyRadiomics (version: v3.1.0), built on Python 3.8. According to IBSI guidelines [123], since MRI signals have no standard units, prior normalisation of grey levels is performed. Images are interpolated to reach a standard unitary spacing of [1,1,1] mm in all directions to be comparable with the resolution exploited in the segmentation stage. A basis spline interpolator is chosen for the images, whereas for the ROI, when necessary, a nearest neighbour interpolator is used to preserve label values' integrity.
Discretisation is applied in terms of a fixed number of bins as suggested by IBSI. Different numbers of bin counts (4, 8, 16, 32, 64, 128, 256) are tested and selected in a secondary stage.
A total of 105 radiomic features are extracted from each image, including 14 shape features (note that they are extracted only once, derived from only the unique lesion mask rather than grey level information), 18 first-order features, and 73 textural features. A summary of the extracted feature groups is presented in Figure 4.1. For a complete definition and implementation of all the features, the reader is referred to the IBSI documentation [123] and PyRadiomics documentation (`https://pyradiomics.readthedocs.io/en/v3.1.0/features.html`).

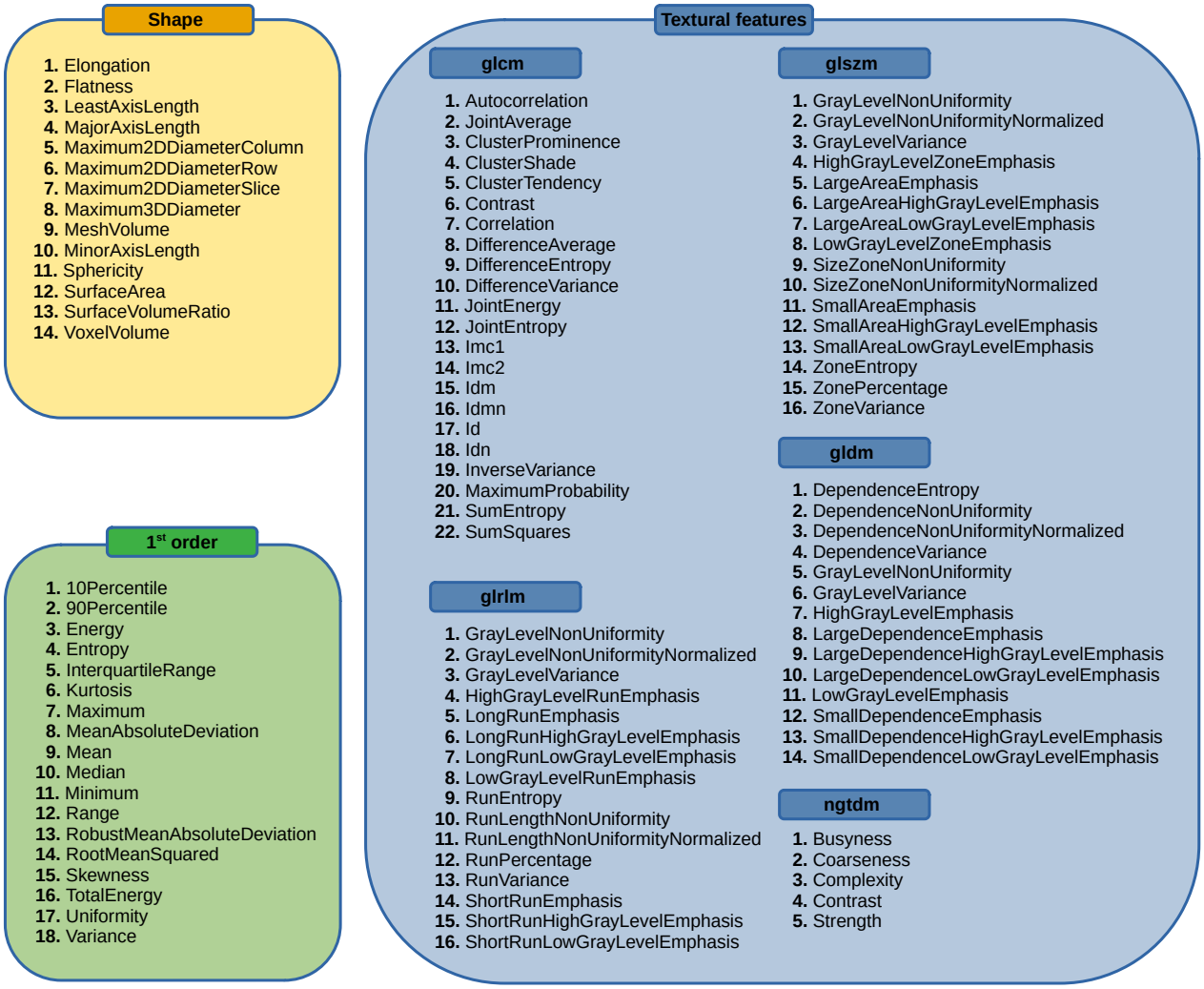


FIGURE 4.1: Summary of extracted features divided by group. Feature names are defined according to PyRadiomics available feature sets.

Features are extracted separately for both manual and automatic segmentation of the ROI.
In an attempt to enhance image information, features are extracted not only from the original image but also from derived images after applying various filters. A summary of the used filters and their parameters is presented in Table 4.2.
To enhance information on edges, first-order gradient filters are utilised to enhance rapid changes in intensity, or at higher orders, the Laplacian of Gaussian (LoG) filter is employed. The LoG filter is obtained from the convolution of a 3D Gaussian kernel to smooth the image and the Laplacian kernel to emphasise regions of rapid grey level changes. Different levels of variance can be chosen, with larger variances emphasising coarser textures with variations on larger scales, and smaller variances focusing on finer textures. Additionally, logarithm and exponential filters can be applied to change the dynamic range, enhancing low-intensity details and high-intensity details, respectively. Local binary patterns (LBP), which compare intensity values within a voxel's neighbourhood, are useful for capturing local texture information and structures. LBP analysis for 3D images requires sampling based on a spherical harmonic framework to preserve rotational invariant representation. Utilising different levels of spherical harmonics, as well as evaluating the kurtosis image, which quantifies the peakedness or

tails of the distribution of intensity values within the voxel's neighbourhood, can provide valuable information about patterns at various scales. Moreover, wavelet decomposition acting as high-pass (H) and low-pass (L) filters in each direction can be used to enhance high and low spatial frequency details, respectively.

n.	**filter**	**parameters**
f1	Laplacian of Gaussian 3D	$\sigma = 0.6$ mm
f2	Laplacian of Gaussian 3D	$\sigma = 1.1$ mm
f3	Laplacian of Gaussian 3D	$\sigma = 2.2$ mm
f4	Laplacian of Gaussian 3D	$\sigma = 3.3$ mm
f5	logarithm	-
f6	exponential	-
f7	gradient	-
f8	local binary pattern 3D	1 level
f9	local binary pattern 3D	2 levels
f10	local binary pattern 3D	kurtosis
f11	wavelet	LLH
f12	wavelet	LHL
f13	wavelet	LHH
f14	wavelet	HLL
f15	wavelet	HLH
f16	wavelet	HHL
f17	wavelet	HHH
f18	wavelet	LLL

TABLE 4.2: Summary of the applied filters.

4.3 Dimensionality reduction

Considering the use of multiple images from the sequence and the application of a given number of filters, the number of extracted features increases from hundreds to several thousands depending on the choice. In the present case, considering all the 18 applied filters and the use of both the third dynamic and second subtracted images, gives a total of 3472 features for each discretisation level choice, which is a much higher number with respect to the sample size of only 127 cases. This leads to incurring both the curse of dimensionality and a high risk of overfitting. As increasing the dataset size is not immediate, requiring years of patient monitoring for small centres, and the use of open-source data is not easy as different imaging and therapy procedures are applied, the only suitable choice is to perform dimensionality reduction. Concerning the optimal number of features, several thumb rules, relating a tenth or dozen of samples to each feature included in a model, are present; however, they lack true evidence. Different selection methods are used without common agreement and with high variability in the results. Some solid approaches to produce a significant radiomic signature are presented by[72] and [104]. In both cases, dimensionality reduction is presented as a multistage process involving the identification of unstable, irrelevant, and redundant features. A cascade pipeline, as described by [72], will be applied involving a first stage of assessing feature stability, followed by a second stage analysing feature variance, a third stage of correlation analysis, and a final selection stage. A visual exemplification is provided in Figure 4.2, showing different filter stages and the reduction of feature number with respect to the initial number. Alternative ways involve the use of feature engineering techniques that can be both performed on raw features or after prior filter stages.

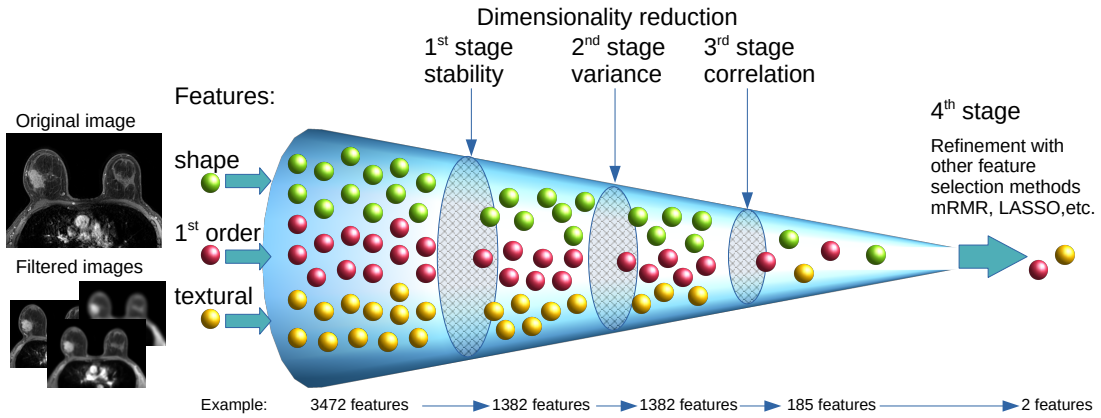


FIGURE 4.2: Feature selection with different filtering stages. (Image inspired by [72])

4.3.1 Feature stability

The first stage involves the assessment of feature stability and the exclusion of non-reproducible features. It's essential to consider that using both manual and automated delineation of the ROI introduces a certain degree of variability, as evidenced in Chapter 3. Features that exhibit high variability, whether due to different observers or the same observer at different times, or due to the choice of different segmentation thresholds, should not be included as they are likely not informative in a predictive model [104]. A fundamental principle for developing any reliable radiomic model is that it must be stable and consistent under the same conditions. Currently, there is no consensus on how to assess feature stability, reproducibility, or repeatability [11]. Different definitions of stability exist, involving temporal stability in a test-retest setting and spatial stability concerning the robustness of features to variation in segmentations [72]. Other methods to assess feature stability, as found by [11], involve stability to variations in different parameters, such as vendors, scanners, acquisition parameters, observers, and preprocessing parameters, but much remains to be investigated overall.
In any case, IBSI [123] recommends performing feature robustness assessment before performing feature selection. Moreover, it is important to note that robustness alone does not imply selection regarding discriminative power and predictive performance. Additionally, robustness might be dataset and disease-dependent. Overall, a robustness analysis can be regarded as a pre-selection step to achieve dimensionality reduction and produce more reliable results.
In the presented case, feature stability is confined to the context of spatial stability over variations in the segmented ROI. One approach to assessing the reliability of features is based on the analysis of variance, with one of the most common choices in the literature being the intraclass correlation (ICC) [111]. ICC serves as a generalisation of the Pearson correlation coefficient to a number of raters higher than two. To conduct ICC analysis, it is recommended to involve at least 30 heterogeneous samples and at least 3 different operators [47]. To address both requirements, accounting for inter-operator variability and variability in the choice of parameters for semi-automatic and automatic segmentation, as well as potential postprocessing effects such as Gaussian smoothing and morphological operations like dilation or erosion, a perturbative approach has been selected. To simulate different observers and parameters from any original segmentation, perturbations are applied to generate new ROIs:

- Binary dilation of the volume with 3 different kernel radii (1, 2, 3 mm) is performed to simulate ROI enlargements and inclusion of tiny parts of extra tissues or lesion borders. This accounts for systematic effects to both lower the binarisation threshold of the automated segmentation and to mimic systematic choices of the manual operation.

- Binary erosion with a kernel radius of 1 mm is conducted to simulate reduction in the ROI due to an increase in segmentation threshold or to account for more conservative operators.

The decision to use more dilation images than erosion cases is based on the fact that reducing the lesion area further diminishes the significance of intensity variation, as no different tissues are included. Moreover, some of the smaller lesions could be completely eroded, leaving no area to extract features. Conversely, dilation leads to higher variability, as different tissues might be included. All operations are performed using the Python library SimpleITK (version 2.2.1). After generating the new ROIs, features are once again extracted. In total, 5 raters are obtained, considering the original case, the three dilated masks, and the eroded one.

The choice to use perturbations of the ROI has been previously employed as an alternative to test-retest to estimate feature robustness by [64], [111], and [122]. This approach reduces the need to acquire multiple images from patients and to engage additional radiologists to segment the ROI, thus ensuring the generalisability of radiomic models.
As 5 raters are involved, accounting for systematic effects of dilation and erosion, an ICC two-way model based on consistency is chosen. Consistency is preferred over absolute agreement, as features are typically centred and scaled before further analysis or before being fed into machine learning models. In such cases, systematic differences that can be rectified through additive transformation approaches become irrelevant, and consistency measurements are preferred [62].
Assuming a sample of n subjects for which k raters measure the same feature based on their ROI segmentation, the measurements can be represented by an $n \times k$ matrix. Analysis of variance can then be performed by calculating all possible sum of squares and mean squares from this matrix. Each feature in the data matrix x_{ij} can be represented by a sum of five terms:

$$x_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij} \equiv \mu + r_i + c_j + \nu_{ij} \tag{4.1}$$

where μ is the mean value of the population of subjects, r_i is a term sampled from a zero-centred normal distribution with variance σ_r^2, and $\mu + r_i$ is the true value for the case i. c_j represents a systematic bias common to all measurements of rater j, sampled from a zero-centred normal distribution with variance σ_c^2, and rc_{ij} is an interaction term accounting for a bias effect not equal for all the subjects, whereas the term e_{ij} is a residual error for each case. As rc_{ij} and e_{ij} are both sampled from a zero-mean normal distribution, they can be described as a single term ν_{ij} with a new variance, the sum of the two variances $\sigma_\nu^2 = \sigma_{rc}^2 + \sigma_e^2$. Thus, the population intraclass correlation coefficient evaluating consistency defined for such a model is given by the variance of interest over the total error variance, neglecting the bias term [55]:

$$\rho_{2C} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_\nu^2} \tag{4.2}$$

Within this framework, the ICC(C,1)[1] along with their 95% confidence intervals are calculated.
According to [47], ICC values ≤ 0.5 indicate poor reliability, values in the range [0.5, 0.75] indicate moderate reliability, values in the range (0.75, 0.9] indicate good reliability, and above 0.9, an excellent reliability is obtained. With the idea of selecting features with reliability in the good to excellent range, in line with other performed studies, a cutoff of 0.75 was imposed on the lower bound of the CI, and only features with a population ICC > 0.75 according to an F-Test are considered sufficiently reliable. Feature stability is then assessed in terms of ICC of type ρ_{2C} for all the original and filtered images, considering various choices of bin count values. The assessment of feature stability is first used to select the optimal discretisation method, as suggested by [64]. The number of robust features is monitored

[1] or equivalently denoted as ICC(2,1) and ICC(3,1) according to standard notation in [47], with no difference in the calculation occurring in the evaluation by assuming both a random or fixed bias defined in ICC models of type 2 and 3.

as the number of bins varies, and the optimal bin number is chosen as the one that maximises the total number of robust features across all images and filters. Furthermore, once the discretisation is fixed, non-robust features are excluded from any further analysis.
The analysis is performed using the irr package (version 0.84.1) running on R (version 4.3.2).

4.3.2 Variance analysis

Secondly, it is necessary to remove features that are irrelevant and do not convey useful information, such as features with zero or nearly zero variance. These features are not useful for discriminating between classes in any predictor. Features presenting a unique value, known as zero variance predictors, are removed, as well as features that have few unique values relative to the number of samples, with a large ratio of frequency of the most common value to the second most common value, known as nearly zero variance predictors.
The cutoff for the frequency ratio of the most common value to the second most common value is set to 95/5, and the percentage of distinct values over the number of samples is set to 10, consistent with the values used in other studies, such as [64]. After this stage, features can be standardised to have a mean of 0 and a standard deviation of 1.
The analysis is performed using the R package caret (version 6.0-94).

4.3.3 Correlation analysis

As many ML algorithms still struggle to account for collinearity, a further selection of non-redundant variables is necessary. As suggested by [104], studying correlations among features and corresponding data visualisation are crucial. Subsequently, correlation heatmaps are constructed to identify correlation clusters, considering that many features have similar definitions or are derived from one another, making high correlation almost inevitable.
Following the approach recommended by [72], once correlated blocks of features are identified, dimensionality can be reduced by setting a threshold on the correlation coefficient. To address not only linear correlation but also any type of correlation, Spearman correlation coefficients are computed pairwise for all features. A threshold for feature elimination is established, and among features with pairwise correlations exceeding the threshold, the one that reduces the mean correlation with all remaining features is retained, similar to the method described by [109]. The threshold is set to 0.9 to remove only highly correlated variables while retaining those that may encode non-redundant information. Further selection, evaluating information and its relation with the outcome, is performed in subsequent stages. In this case, the analysis is conducted using the R packages caret (version 6.0-94) and corrplot (version 0.92) for visualisation.

4.3.4 Feature selection

Following the initial refinement process, which involved eliminating non-robust and non-reproducible features, as well as addressing zero variance and high correlation issues, the next step is to create a meaningful radiomic signature based on the remaining variables. It is crucial to note that despite the initial pruning, the number of features remains relatively high given the available dataset. Attempting to build a model with all of them would risk overfitting and succumb to data sparsity. Therefore, further feature selection is imperative. As outlined in Chapter 2, a variety of methods, both supervised and unsupervised, can be employed. Given that basic filter methods based on variance and correlation have already been applied, exploring alternative approaches becomes essential. Wrapper methods, although effective, tend to be computationally expensive and inefficient, and for this reason were excluded. Given the inherent variability and method dependence on the selection of features, some of the most widely used methods according to [116] are tested together with an unsupervised method:

- mRMR feature selection: this supervised filter method is chosen for its independence from specific models, providing a set of non-redundant features showing a relation with the outcome without accounting for the choice of the classifier.
- LASSO: offers an alternative perspective relying on penalised linear regression;
- UDFS: exploiting both discriminative analysis and penalised regression, assumes linear separability of the classes without requiring any knowledge of class labels [2].

Moreover, feature engineering using principal component analysis is also tested. This involves attempts to apply PCA to both the initial features that remained after stability assessment and the features that remained after filtering based on variance and correlations.
Notably, as observed in Chapter 2, both mRMR and UDFS, while adept at selecting informative features, do not inherently dictate the specific cardinality of the optimal feature set. This characteristic introduces an important dimension of variability into the analysis.
Unlike other selection methods such as LASSO, which automatically define the number of features to be retained, mRMR and UDFS provide flexibility in the selection process. They offer a wide range of potential feature subsets, and the choice of cardinality becomes an essential consideration. This flexibility is particularly pertinent when dealing with small datasets, where the optimal number of features may vary, impacting model performance and generalisability. Given the limited size of the dataset, our chosen approach involved selecting sets of features ranging from 1 to a maximum of 15 features (higher numbers are discarded to prevent overfitting risk). Models in the subsequent stages are trained with an increasing number of features to systematically assess the impact on performance, and the optimal number of features emerges from the best model.
In contrast, with LASSO, the optimal number of features is automatically determined as a result of the penalisation, although it may depend on the tuning of the penalisation parameter, showing minimal variability. Regarding PCA, the number of principal components to be used can be chosen to represent most of the variance in the dataset (e.g., at least 80%), or it can be tested in a similar manner to mRMR and UDFS, with an increasing number of components.
All methods are implemented in R using the following packages: mRMRe (version 2.1.2.1) for mRMR feature selection, glmnet (version 4.1-8) for LASSO, Rdimtools (version 1.1.2) for UDFS, and FactoMineR (version 2.9) for PCA.

4.4 Model development

Regarding the final stage of building a classifier with previously selected features, considering the task of performing binary classification on a small dataset, it's advisable to opt for simple classifiers to avoid overfitting. A systematic review of previous studies by [8] for pCR response classifiers and by [2] for general breast lesion classification revealed that commonly used classifiers include SVMs, decision tree-based classifiers, and simpler models such as logistic regression. In terms of performance, choosing and comparing models from literature can be challenging due to the significant dependence on dataset size. Some studies show that SVMs and decision trees achieve higher performance, while in others, simpler models outperform them, especially when dealing with smaller datasets that are more susceptible to overfitting by complex models with more parameters. To mitigate any bias from prior knowledge, three models are initially tested:

- Logistic regression: a straightforward choice known for its simplicity and interpretability.

[2] This method was tested by [64] in other classification tasks for malignant breast lesions in MRI, outperforming many of the other supervised and unsupervised selection methods.

- Support Vector Machines: recognised for their effectiveness in handling complex datasets and nonlinear relationships;
- Random Forests: a versatile ensemble method renowned for its robustness and ability to capture complex patterns in the data.

The analysis is implemented using the R packages glmnet (version 4.1-8) for logistic regression, e1071 (version 1.7-13) for SVMs, and randomForest (version 4.7-1.1) for RF.

4.4.1 Model assessment and selection

As various feature selection techniques and classifiers are utilised, robust performance evaluation is essential for both model selection, comparing the performances of different models to choose the best one, and model assessment, estimating the generalisation errors on new data. In scenarios where the dataset is insufficient for triple splitting into training, validation, and test sets, alternative strategies are required to estimate the extra-sample error, such as cross-validation. K-fold Cross-Validation is a commonly employed approach, involving the division of data into K similar-sized subsets. The model is trained on K-1 subsets and tested on one subset, repeated for all possible combinations. Aggregating the errors from each fold yields a comprehensive estimate of the prediction error [36]. In many cases, a leave-one-out cross-validation approach is adopted, setting the number of folds K equal to the number of samples. While this method guarantees low bias in the error estimates since almost all of the data is used in training, it suffers from high variance due to correlated error estimates on the test cases, and it is computationally intensive. Conversely, setting K to lower values generates more differentiated training folds and less correlated error estimates, resulting in lower variance but higher bias, overestimating the error on test data, due to using only a fraction of the data for training [36]. To strike a balance between bias and variance, a 20-fold cross-validation approach is adopted, utilising 95% of the data in the training fold and 5% for testing. However, a single stage of K-fold cross-validation may not provide a robust estimation of performance due to the randomness of fold generation. More robust and stable estimations are achieved through repeated cross-validation, where folds are generated multiple times using different initialisation[37]. Moreover, by repeating the entire process of partitioning and estimation multiple times, the variability is reduced, outperforming the non-repeated cross-validation estimates [46]. Therefore, it was decided to repeat the 20-fold cross-validation 10 times, each time with different initialisation using pseudo-random seeds in the fold splitting phase, resulting in a total of 200 estimates of model performance. This strategy ensures higher reliability in the evaluation process, especially given the relatively small size of the dataset.
Before moving on to evaluation metrics, it's crucial to clarify the actual object of evaluation. Since any step of supervised feature selection should be included in the cross-validation loop, evaluation encompasses both the feature selection method and the classifier. Therefore, the model comprises both feature selection and classifier. The correct procedure, as outlined by [92], [3], and [36], requires performing supervised feature selection within the cross-validation loop. This involves splitting the folds into training and test folds, conducting feature selection and model training on the training folds, and then testing on the test fold. Supervised feature selection should be performed on each training fold separately to prevent unfair advantages in prediction. Test data should be left out before supervised selection of variables is performed. It's essential to perform supervised feature selection on the training set only to prevent label information leakage from the data assigned to the test set, which could lead to overestimated performances. Similar considerations apply to hyperparameter tuning of both feature selection and classifiers. In contrast, unsupervised screening steps can also be performed prior to splitting since class labels are not involved [36]. Following these recommendations, the initial filtering stages of stability assessment, variance, and correlation analysis are conducted outside the cross-validation loop. On the other hand, all other stages of feature selection and hyperparameter

tuning are always included inside cross-validation for fairness.
For hyperparameter tuning, a nested approach is utilised, where the training fold of the outer cross-validation is further split into folds to conduct an internal cross-validation to select hyperparameters of both feature selection (e.g., the penalisation for LASSO) and classifiers (e.g., the number of trees for Random Forest, the kernel for SVM, etc.). Considering computational costs, a 5-fold cross-validation is employed in this case.
In summary, the entire process involved a nested cross-validation approach, comprising an outer loop of 20-fold cross-validation, repeated 10 times, and an inner loop of 5-fold cross-validation for tuning. Performances are estimated in terms of AUC. The model is fitted on K-1 folds, predictions are generated for the test fold, and used to generate the ROC curve and compute the AUC. The procedure is repeated for all the folds and the defined splitting schemes. A cross-validated AUC is then obtained as the mean $\overline{AUC}$ together with its standard error SE $= \sigma/\sqrt{M}$, where σ is the standard deviation, $M = K \times V$,K is the number of folds, and V is the number of repetitions [44]. Accordingly, 95% confidence intervals can be estimated as $\overline{AUC} \pm 1.96$ SE [50]. To perform model selection in terms of the number of parameters, the one-standard error rule is usually recommended, selecting the most parsimonious model that does not differ by more than one standard error from the model with the best performance [36].

4.4.2 Workflow

Features are extracted from both radiologist-delineated ROIs and automatic segmentations obtained with V-Net. The features extracted from the radiologist ROI serve as the gold standard and reference for the predictive model, while those obtained from automatic segmentation are retained for stability checks on the feasibility of using automatic segmentation in future studies with the same model.
For the gold standard features, various feature selection techniques and classifiers are trained separately using features obtained from the original image, each filter separately, and ultimately all filters combined. Model performances with radiomic features are assessed for each combination of classifier, feature selection method, and filter.
Additionally, in line with previous studies [67], the potential enhancement from incorporating the clinical variables shown in Table 4.1 is investigated. Clinical variables are preprocessed investigating correlations, continuous variables are centred and scaled, and categorical variables are encoded using dummy variables encoding. These processed variables are then combined with the selected radiomic features to train the classifier, and performances are re-evaluated.
The best models for each scenario are compared, and then tested using features obtained from automatic segmentations to assess model stability.

4.5 Results

Results from all stages of model development are presented as follows.

4.5.1 Stability and discretisation

Considering both the dynamic and subtracted images, along with all the 18 filters listed in Table 4.2, a total of 3742 features were extracted for each individual patient.
Stability assessment to morphological perturbations more than halved this initial number, resulting in different selections of robust features depending on the choice of bins' number. A summary of the number of stable features is presented in Table 4.3, delineating separately the estimates for the common shape features and all other features extracted from the dynamic and subtracted images. It is noteworthy that the subtracted images consistently offer a higher number of stable features overall.

bins' number	N of robust features			
	shape [/14]	**sdyn2** [/1729]	**dyn3** [/1729]	**all** [/3742]
4	13	548	482	1043
8	13	711	648	1372
16*	**13**	**720**	**649**	**1382**
32	13	691	621	1325
64	13	684	621	1318
128	13	650	607	1270
256	13	655	612	1280

TABLE 4.3: Number of robust features for different bin number values. The number of selected features for both the 3^{rd} dynamic image and the 2^{nd} subtracted image is slightly higher in the case of bc=16. In general one can notice that the number of robust features from the dynamic image is smaller with respect to the subtracted. Both the lowest and highest values of bin numbers lead to a decrease in the number of robust features.

In order to determine the optimal value for the bin number, individual images and filters were investigated. Tables 4.4 and 4.5 display the number of robust selected features extracted from the second subtracted (sdyn2) and third dynamic (dyn3) subtracted images, respectively, considering each specific filter. As shape features are not dependent on intensity, only the count of selected features among the remaining 91 features is reported.

bc	N of robust features, sdyn2									
	original [/91]	**f1** [/91]	**f2** [/91]	**f3** [/91]	**f4** [/91]	**f5** [/91]	**f6** [/91]	**f7** [/91]	**f8** [/91]	**f9** [/91]
4	26	26	33	46	46	26	29	54	34	19
8	34	44	44	47	46	20	41	54	36	18
16	36	46	40	43	41	23	47	53	34	15
32	36	41	38	42	39	23	48	55	30	12
64	33	42	38	41	43	23	46	55	24	12
128	33	38	34	42	43	23	46	50	24	12
256	34	41	38	40	42	22	40	50	24	13
bc	**-** -	**f10** [/91]	**f11** [/91]	**f12** [/91]	**f13** [/91]	**f14** [/91]	**f15** [/91]	**f16** [/91]	**f17** [/91]	**f18** [/91]
4	-	13	28	20	23	28	27	27	13	30
8	-	13	41	36	37	47	40	43	38	32
16	-	13	41	42	39	47	41	42	45	32
32	-	13	40	39	39	43	40	39	41	33
64	-	10	40	40	40	43	40	40	42	32
128	-	10	38	36	35	43	36	39	37	31
256	-	10	36	35	38	42	36	39	40	35

TABLE 4.4: Number of robust features for different filters and bin counts, in the case of the 2^{nd} subtracted image

bc	N of robust features, dyn3									
	original [/91]	**f1** [/91]	**f2** [/91]	**f3** [/91]	**f4** [/91]	**f5** [/91]	**f6** [/91]	**f7** [/91]	**f8** [/91]	**f9** [/91]
4	31	25	31	39	34	21	36	30	23	15
8	30	38	40	40	37	18	47	40	20	14
16	31	38	36	36	31	18	49	34	19	13
32	31	37	35	35	29	18	47	36	16	10
64	32	37	35	34	32	18	46	37	12	10
128	34	32	37	34	32	17	49	35	12	10
256	33	37	37	35	33	17	43	33	12	11
bc	**-**	**f10**	**f11**	**f12**	**f13**	**f14**	**f15**	**f16**	**f17**	**f18**
	-	[/91]	[/91]	[/91]	[/91]	[/91]	[/91]	[/91]	[/91]	[/91]
4	-	13	23	21	18	25	30	22	14	31
8	-	16	43	34	41	42	37	43	38	30
16	-	18	41	41	42	42	43	42	45	30
32	-	19	38	40	38	41	41	40	42	28
64	-	16	40	38	40	40	41	41	42	30
128	-	17	35	39	40	36	37	42	38	31
256	-	17	39	38	39	38	40	39	38	33

TABLE 4.5: number of robust features for different filters and bin counts, in the case of the 3^{rd} dynamic image

Observations from the analysis reveal that the application of different filters can have varying effects on the number of robust features. For example, the Laplacian of Gaussian filters (f2, f3) and the wavelets (f11-f18) tend to enhance the number of robust features.
Conversely, other filters, such as the local binary patterns (f9, f10) and the logarithm (f5), tend to decrease the number of robust features in both the subtracted and dynamic images. Given the variable nature of the number of stable features and the absence of a clear trend, and without prior knowledge of which image or filter is most relevant to the outcome, a generic approach was adopted in selecting the bin number. The optimal bin count was determined to be 16 by maximising the total number of robust features considering both original images and filters applied to the third dynamic and second subtracted image.
Once the bin number was determined, an examination of the surviving features for each class became necessary. Regarding shape features, nearly all of them remained stable, with the exception of the surface to volume ratio, which exhibited changes associated to the applied morphological perturbations. This resulted in a final count of shape features remaining almost unchanged, with 13 out of the initially computed 14 features. For other feature groups, which vary depending on the image and filter, a summary is presented in Table 4.6, detailing the number of selected features for each class and filter of the two images. It is evident that the glcm feature group and most higher-order feature groups were significantly reduced after perturbation, indicating that a radiomic model relying solely on these feature groups, without a selection based on stability, may struggle to handle minor perturbations of the ROI. Conversely, first-order histogram features appeared to be more stable.
The number of features selected from each filter varied, and it is important to note that simply assessing feature stability does not guarantee that filters with a higher number of features have variables that are significant for the outcome. Additionally, it's crucial to consider that selected features may still be highly correlated or stable due to having very few values. For this reason, it is necessary to proceed with the analysis by removing features with zero variance and correlated features.

image		**n. features per class**						**Tot**
		1st o. [/18]	**glcm** [/22]	**glrlm** [/16]	**glszm** [/16]	**gldm** [/14]	**ngtdm** [/5]	[/91]
dyn3	original	8	2	8	6	6	1	31
	f1	12	4	9	6	6	1	38
	f2	12	3	9	6	5	1	36
	f3	9	6	9	6	5	1	36
	f4	7	4	9	5	5	1	31
	f5	10	-	2	3	2	1	18
	f6	13	8	**12**	7	**7**	**2**	**49**
	f7	10	3	8	6	6	1	34
	f8	5	1	3	6	3	1	19
	f9	2	-	2	6	2	1	13
	f10	2	-	6	6	3	1	18
	f11	**14**	4	9	7	6	1	41
	f12	13	5	9	7	6	1	41
	f13	**14**	5	9	7	6	1	42
	f14	13	6	9	7	6	1	42
	f15	**14**	6	9	7	6	1	43
	f16	**14**	5	9	7	6	1	42
	f17	**14**	**9**	6	**8**	6	**2**	45
	f18	8	2	8	6	5	1	30
sdyn2	original	11	4	8	6	5	**2**	36
	f1	13	8	9	**8**	6	**2**	46
	f2	12	6	9	6	6	1	40
	f3	13	8	9	6	6	1	43
	f4	13	8	9	5	5	1	41
	f5	11	1	2	6	2	1	23
	f6	**15**	8	11	7	5	1	47
	f7	**15**	**11**	**13**	6	**7**	1	**53**
	f8	8	10	4	6	5	1	34
	f9	4	-	2	6	2	1	15
	f10	2	-	2	6	2	1	13
	f11	14	4	9	7	6	1	41
	f12	14	5	9	7	6	1	42
	f13	14	3	8	7	6	1	39
	f14	14	8	10	7	6	**2**	47
	f15	14	5	7	7	6	**2**	41
	f16	14	4	10	7	6	1	42
	f17	14	6	9	**8**	6	**2**	45
	f18	10	3	8	5	5	1	32

TABLE 4.6: Number of selected feature per class for each filter, one can see in bold the highest values of each class.

4.5.2 Variance and Correlation

Feature sets are analysed both separately, considering each single filter, and globally, considering a unique feature set. Zero and near-zero variance features were not present, so this stage did not produce a reduction in the feature number. In contrast, the presence of high correlation was observed. An example of a heatmap showing correlation among all the robust features for the dynamic image with filter 6 (the one that had the largest number of features) is presented in Figure 4.3. One can notice the presence of multiple clusters of highly correlated and anticorrelated feature groups, also between features of different groups.

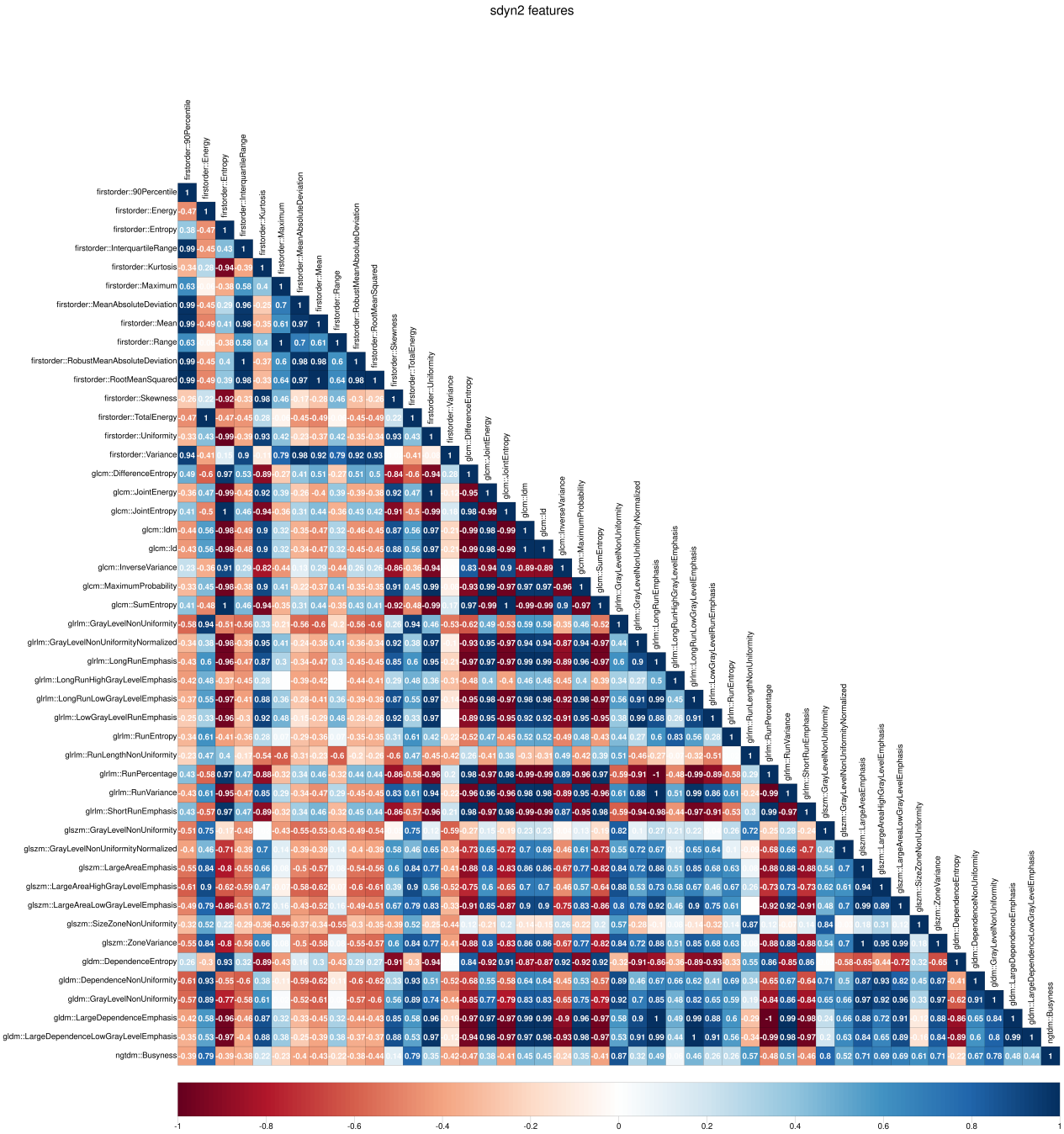


FIGURE 4.3: Correlogram of all the robust features extracted from the 2^{nd} subtracted image with filter f6.

After setting a threshold on the allowed correlation, the feature number further reduced, as can be seen from the summary in Table 4.7.

image	**original**	**f1**	**f2**	**f3**	**f4**	**f5**	**f6**	**f7**	**f8**	**f9**
n features	18	20	18	24	21	16	30	12	11	8
image	**f10**	**f11**	**f12**	**f13**	**f14**	**f15**	**f16**	**f17**	**f18**	**all**
n features	8	18	20	14	19	13	18	17	19	184

TABLE 4.7: Number of remaining features after removal of correlated clusters.

In this stage, clinical variables were also checked for correlations. A Fisher's exact test was performed on categorical variables concerning grade, margin, and type of the tumour. A summary of the results is presented in Table 4.8, showing that whereas type and grade, and type and margin are not related, independence cannot be fully assumed for margin and grade. However, as the degree of correlation does not necessarily imply redundancy, all the clinical variables are kept for the successive stages.

Fisher's exact test (H_0: independent variables)		
	grade	margin
grade	-	-
margin	**0.01***	-
type	0.10	0.65

TABLE 4.8: p-values from Fisher's exact test for categorical variables.

4.5.3 Feature selection and classifiers exploratory phase

After the former dimensionality reduction, different combinations of feature selection and classifiers were explored, first using only radiomic features.
Regarding the classifiers, the logistic model offered quite good performances on both the train and the test folds independently of the feature selection algorithm, with mean AUC values between 0.70 and 0.80. An example for the logistic model paired with mRMR feature selection is presented in Figure 4.4. Mean AUC and standard error are presented as a function of the number of selected features, considering each single filter separately, the original image, and the combination of all filters. In red, the maximum mean AUC is presented for each case. In general, a lower number of features gives higher results, and at some point, increasing the number of features does not improve the model performances. On the contrary, the other classifiers SVM and RF obtained very high performances on the train folds, but much lower results on the test folds, implying that such classifiers are too complex for the small dataset. The fine-tuning of the model hyperparameters within the nested fold probably led to overfitting. Examples are presented for SVM and RF combined with mRMR for comparison in Figure 4.5. One can notice that for SVM, the mean AUC is generally lower than 0.75 in most cases, and better performance is obtained only by including a much higher number of features compared to the logistic case. In the case of the RF classifier, AUC results are quite poor, staying in the range of 0.6-0.7, implying poor generalisation capabilities and suggesting stronger overfitting. Other combinations of RF and SVM are omitted for brevity; however, similar results were obtained when combined with other feature selection techniques, suggesting the need to limit the analysis to a simpler classifier given the nature of the small dataset.
Concerning dimensionality reduction techniques, all three feature selection methods obtained promising results in finding features containing discriminative power to classify the NAC outcome. Generally, the selected features differed, but performances with the same classifier were comparable in the best cases. A plot of the AUC obtained using LASSO and UDFS as feature selection methods is reported in Figure 4.6. One can notice that for some filters (f4, f5, f7, f9, f10), LASSO did not find any relevant features, whereas mRMR and UDFS still selected some features, but achieved overall lower performance compared to other filter choices. LASSO generally selects a reduced number of features, between 1 and 5, suggesting that there is no need to investigate much higher feature numbers. Overall, both LASSO and UDFS paired with the logistic classifier have mean AUC spanning between 0.7 and 0.8, similar to what was obtained with mRMR using the same classifier. Indeed, despite finding different feature sets and numbers, the discriminative capabilities are valid for all of them.

FIGURE 4.4: Mean AUC and standard error for the logistic model with mRMR feature selection as a function of the number of selected features for each filter, with the maximum AUC depicted in red in each case.

On the contrary, the application of PCA as a feature engineering technique was not suitable for discriminating components able to separate the outcome classes. In fact, both applying PCA as a prior filter before removing correlated variables and applying PCA within the fold resulted in components that, even if maximising variance, were not informative for discriminating the desired outcome. When PCA is applied after feature stability assessment, without filtering out the highly correlated variables, the new dimensions still remain correlated, as PCA addresses only linear correlations effectively. An example is presented in Figure 4.7a, showing a scatterplot of a couple of principal components that still exhibit correlation, suggesting the persistence of nonlinear correlations among features.To address this issue, PCA was also applied after the removal of highly correlated variables. However, no clear separation emerges among the two classes defined with these new variables. An example is presented in Figure 4.7b, where the classes appear randomly mixed. Performance by pairing PCA with logistic classifiers showed no discriminative capacity of the model, corresponding to an AUC of about 0.5 as presented in Figure 4.8, independently of the number of principal components in use.
Following this initial discovery phase, both the classifiers prone to overfitting, SVM and RF, and the feature engineering technique, PCA, were excluded from further analysis as they yielded non-significant results. Thus, the final analysis focused on the use of the logistic classifier and the three feature selection techniques: mRMR, LASSO, and UDFS.

FIGURE 4.5: Mean AUC for RF and SVM models

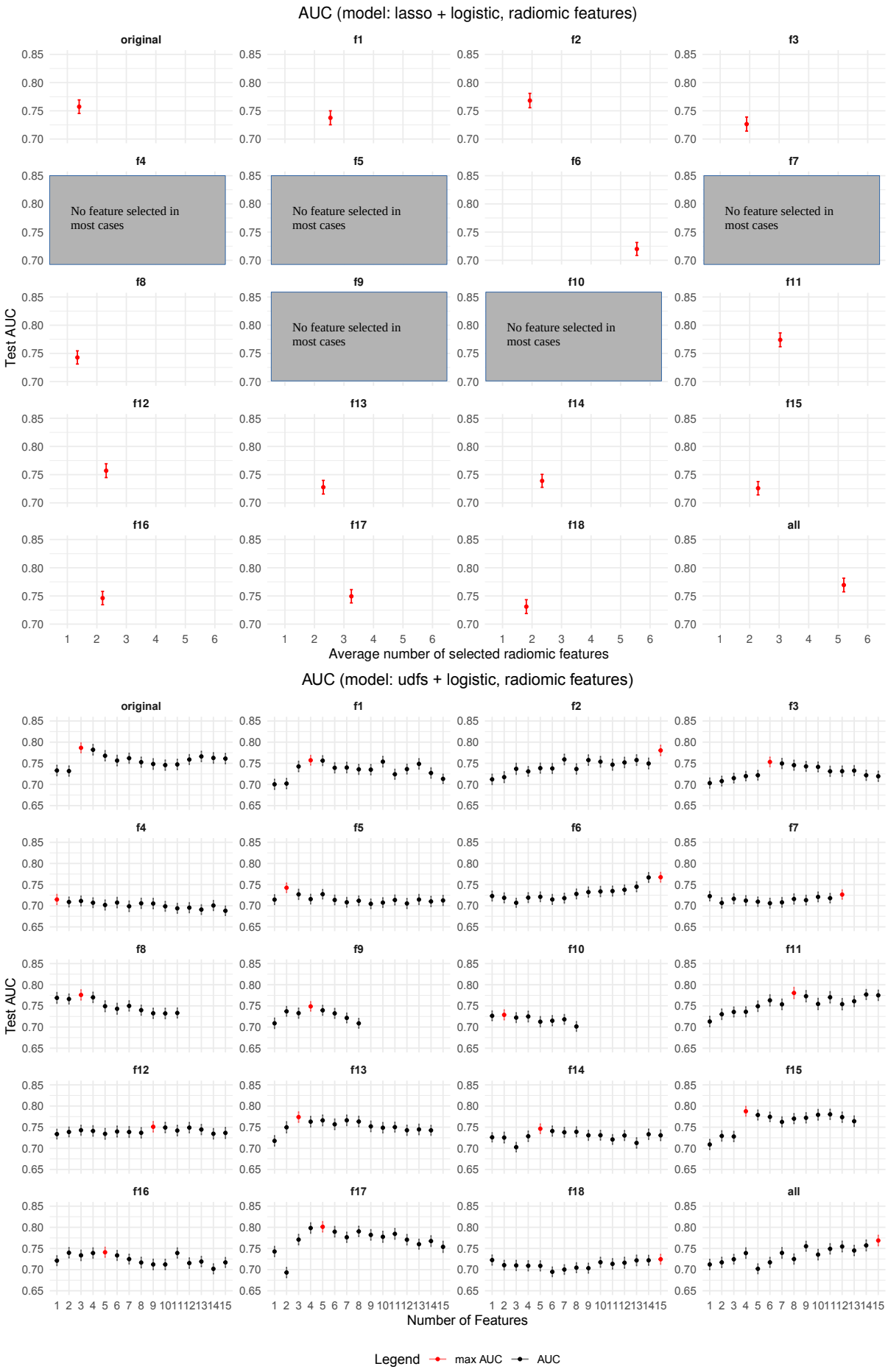
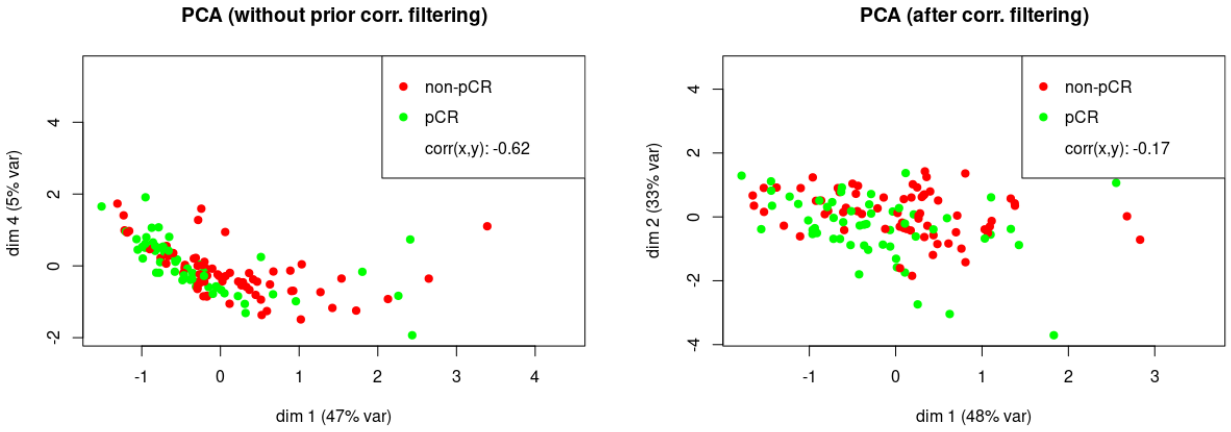


FIGURE 4.6: Mean AUC for LASSO and UDFS paired with logistic classifier models



(A) (B)

FIGURE 4.7: Examples of PCA applied before and after the removal of highly correlated variables are illustrated in plots A and B, respectively. In the first case, the presence of nonlinear correlation emerges and is not removed by PCA application. In the second case, no correlation emerges; however, the components do not allow a clear distinction of the classes of interest.

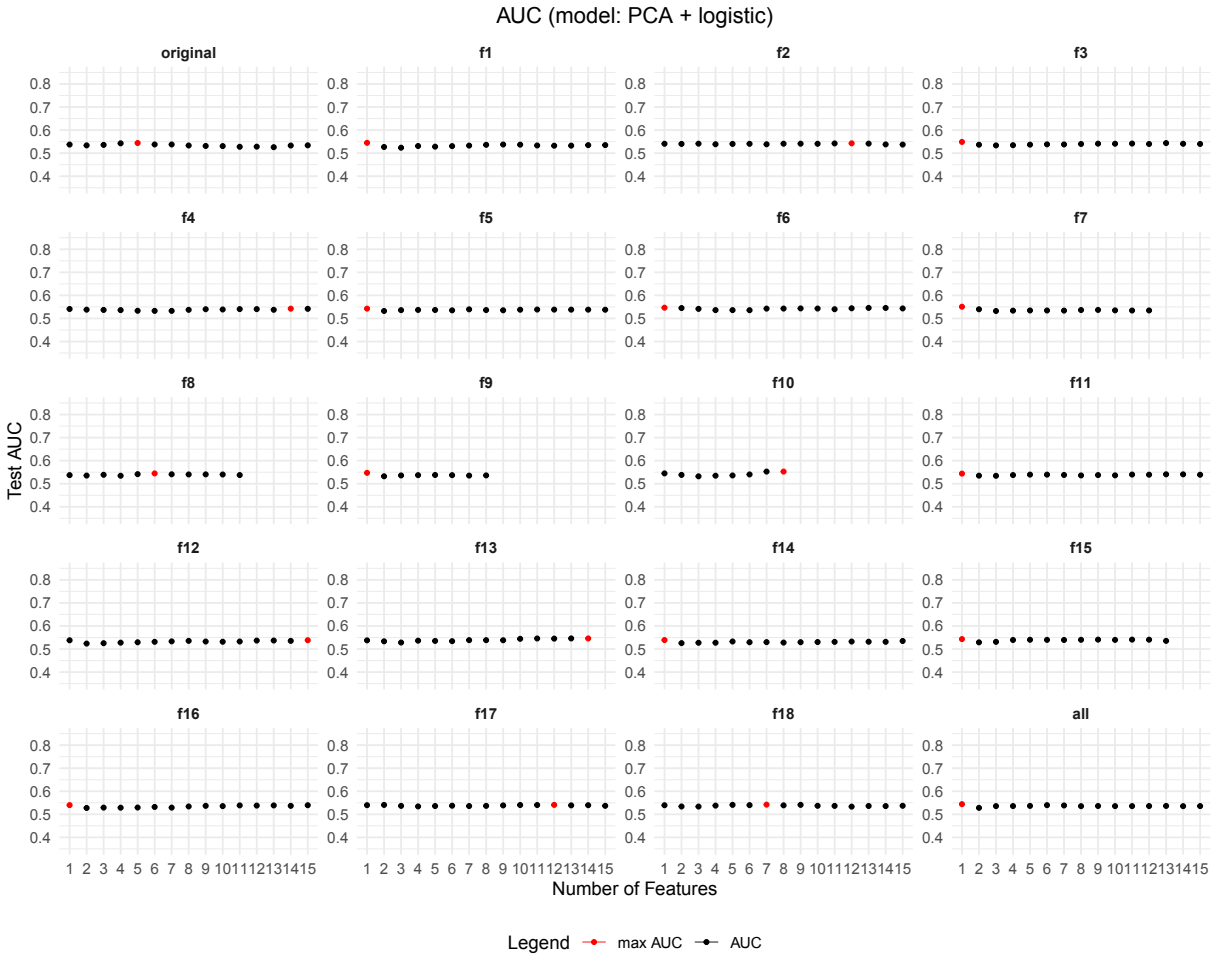


FIGURE 4.8: Mean AUC for logistic model and features obtained from PCA.

4.5.4 Inclusion of clinical variables and model selection

After the discovery phase, restricting the research to a single classifier, clinical variables were also taken into account to build the final classifier informed by both radiomic and clinical information. The inclusion of clinical variables showed a slight enhancement in the predictive power of the model compared to the inclusion of radiomic features alone. Results for mRMR, LASSO, and UDFS are depicted respectively in Figure 4.9 and 4.10.
Each combination of classifier and feature selection techniques, along with the number of selected features, was considered as a model. A search was conducted to sift through these combinations and select the best model within each feature selection method. AUC estimates from all folds and repetitions were compared using the Friedman test, followed by post-hoc pairwise Nemenyi-Friedman tests with Bonferroni correction on p-values. Observing separately each feature selection technique depicted in Figure 4.9 and 4.10, it becomes evident that after the inclusion of clinical variables, most of the highest AUC values are comparable within the error bars. Indeed, across all three feature selection methods, no significant differences between the highest AUC values were observed. Consequently, customised choices are necessary to determine the best model.

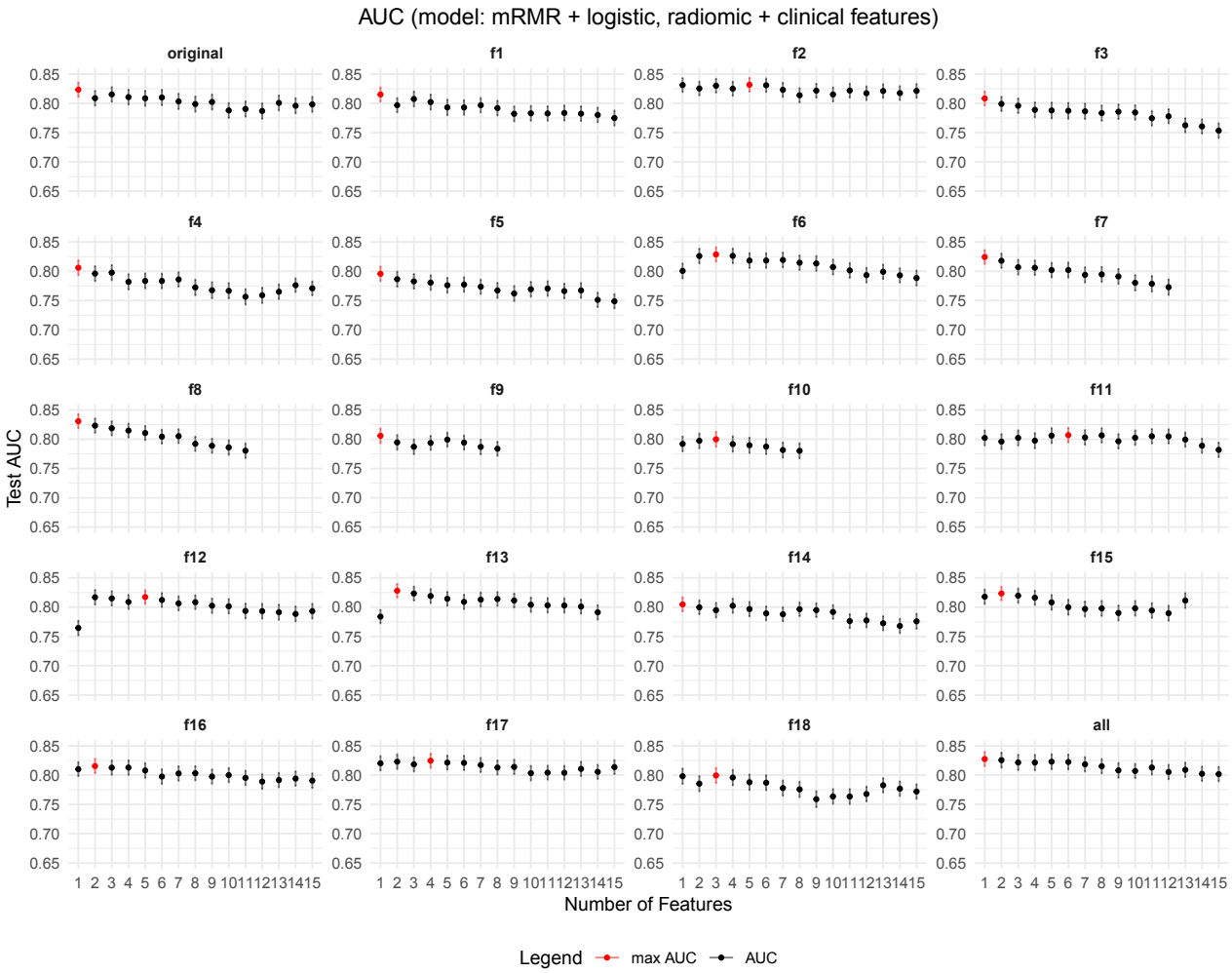


FIGURE 4.9: ,
Mean AUC and standard error for the logistic model with mRMR feature selection and inclusion of clinical features, as a function of the number of selected radiomic features.

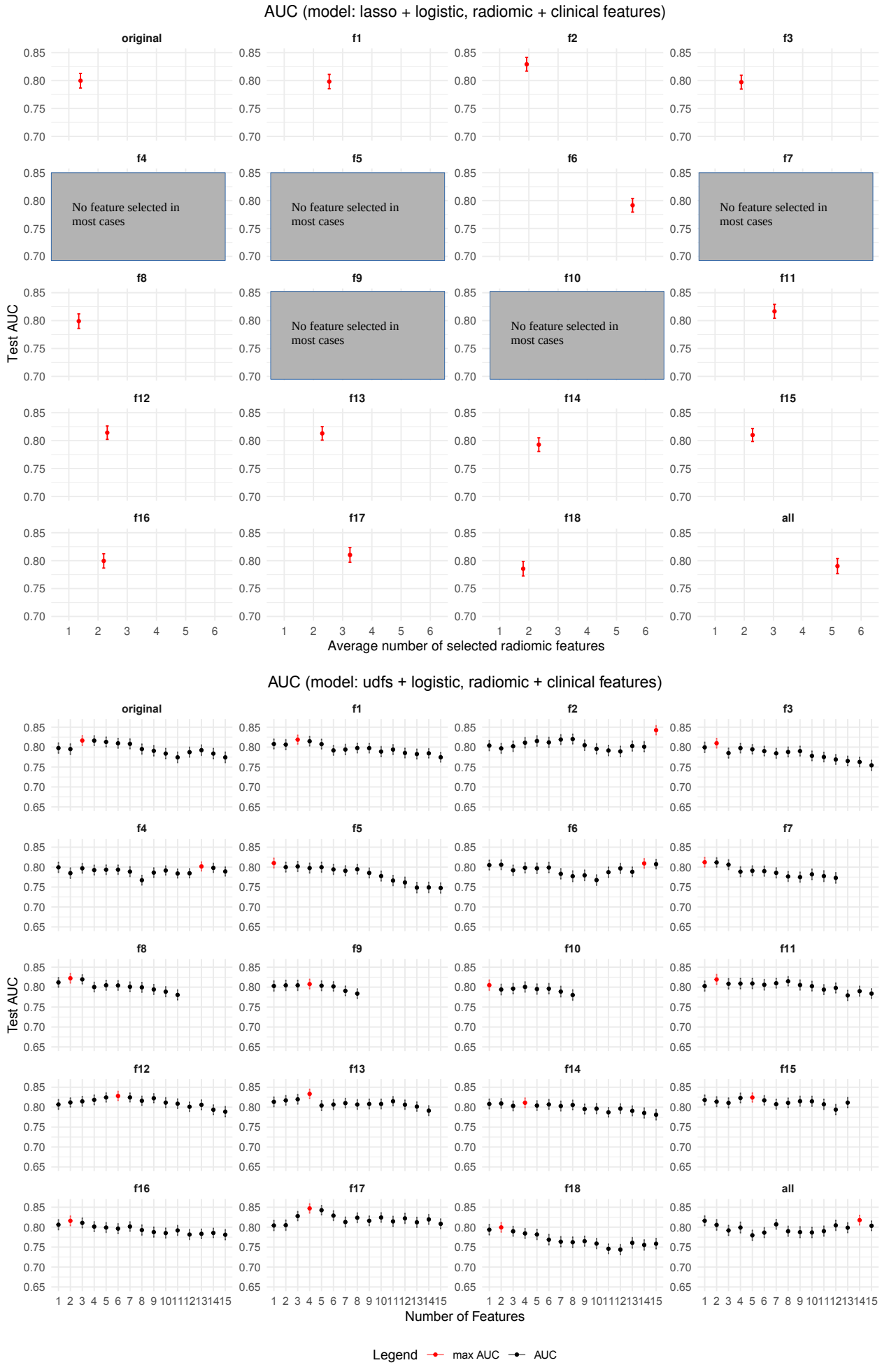


FIGURE 4.10: Mean AUC for the logistic model with LASSO and UDFS feature selection.

Typically, a conservative approach involves striking a balance between model performance and simplicity in terms of the number of parameters. Thus, for each feature selection category, a model was chosen based on this principle.

mRMR + logistic model

Considering models based on mRMR, the plot in Figure 4.9 illustrates that the highest levels of AUC are consistently achieved with the use of filter f2. The maximum AUC is observed to be 0.83 ± 0.01 (CI = 0.81-0.85), which is reached not only with filter f2 using 5 features, but also with 6, 3, and 1 features, covering an overlapping confidence interval. As a conservative choice, the model with the lower number of parameters is selected, ultimately choosing the model with only 1 feature.
The selected feature is determined based on the frequency of its selection within the cross-validation loop. A summary of selected features for filter f2 is provided in Figure 4.11, where the x-axis represents the number of features and the colour indicates the absolute frequency of selection. The most frequently selected feature is extracted from the third dynamic image and is the 'GrayLevelNonUniformityNormalized' of the glrlm textural feature group. This feature estimates the variability of grey-level intensity values in the image, although its interpretation may not be immediately straightforward due to the initial image filtering process.

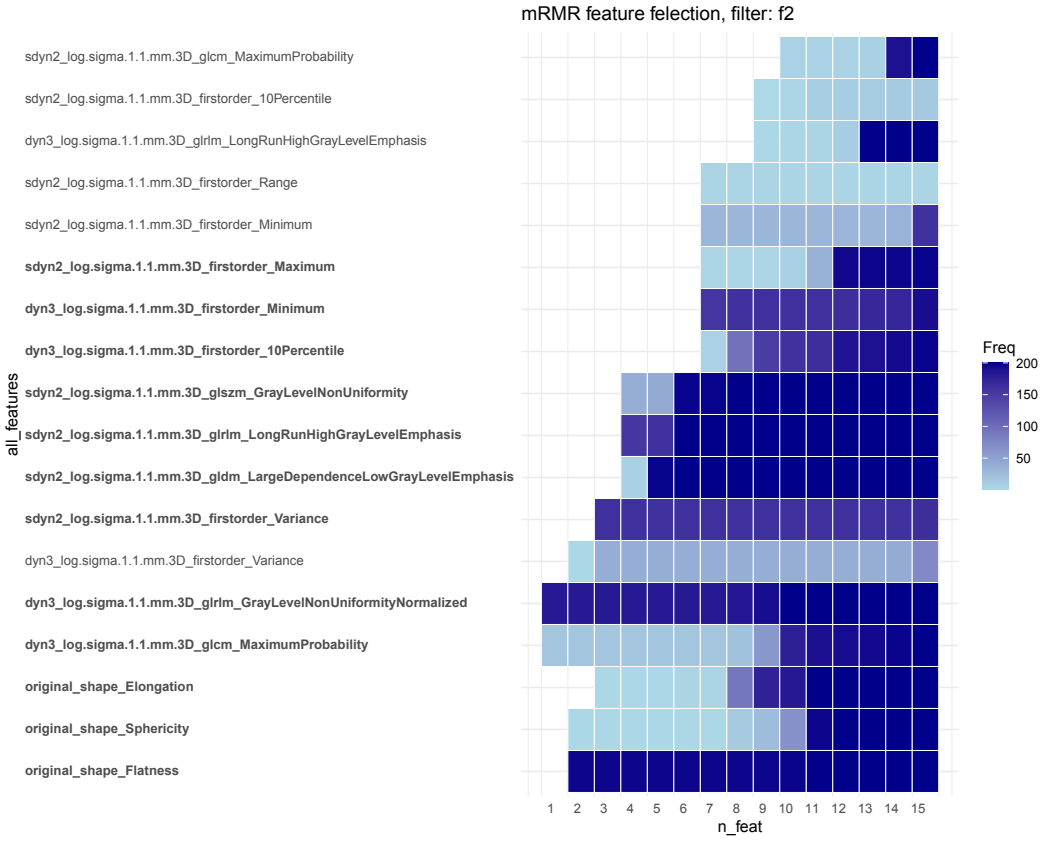


FIGURE 4.11: Heatmap showing the frequency of selected features using mRMR feature selection. The x-axis indicates the number of selected features, while colours denote the frequency of selection, with darker hues indicating higher selection frequency.

A box plot in Figure 4.12 illustrates the distribution of selected features for the two outcome classes in the dataset. Clear separation is observed among the median values of the two classes, with higher

feature values typically associated with non-pathological complete response cases and lower values for patients exhibiting complete response to neoadjuvant chemotherapy. This distinct separation indicates the discriminative power of the selected features in distinguishing between the two outcome groups.

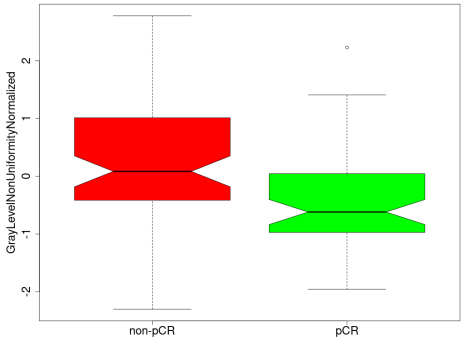


FIGURE 4.12: Distribution of selected feature for the two outcome classes in the dataset. Each box represents the interquartile range of feature values, with the median depicted by the horizontal line within the box

LASSO + logistic model

In the case of models based on LASSO, it is observed that both scenarios involving five of all filtered features and approximately two features from filter f2 yield comparable performances, with respective AUC values of 0.78 ± 0.01 (CI = 0.75-0.79) and 0.78 ± 0.01 (CI = 0.75-0.80). As before, the preference is given to the model with fewer variables for simplicity. The frequency of feature selection can be examined in Figure 4.13, where the most frequently selected features are once again derived from the

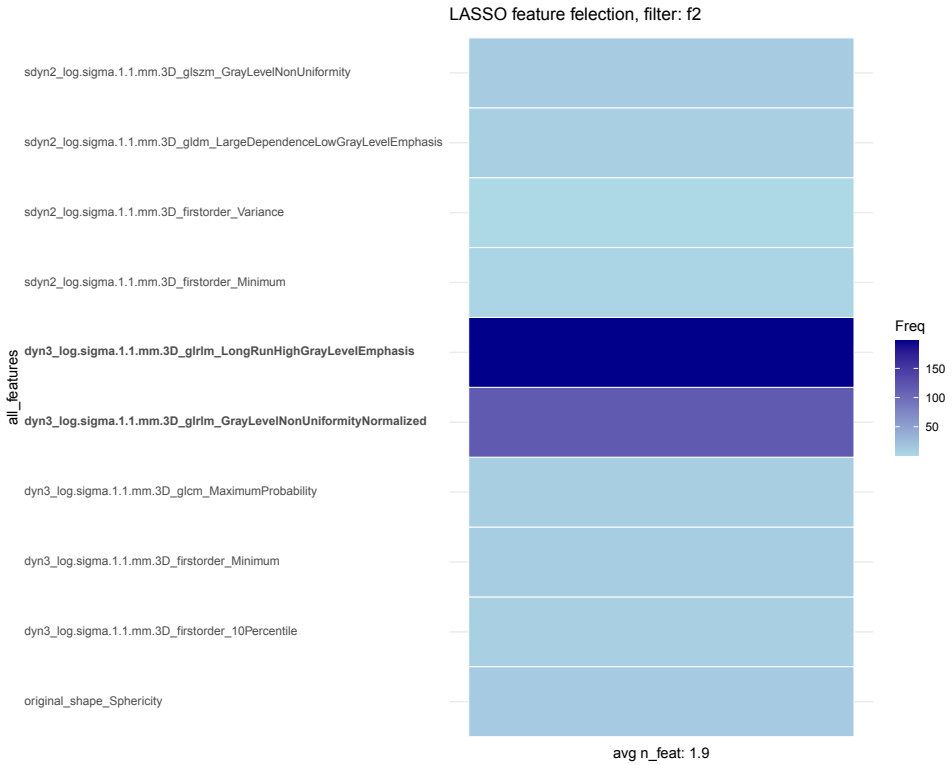


FIGURE 4.13: Heatmap showing the frequency of selected features using LASSO feature selection. The mean number of selected features is reported below.

third dynamic image and belong to the glrlm group. Specifically, 'GrayLevelNonUniformityNormalized' is again identified as one of the most selected features. Additionally, a different feature, 'LongRunHighGrayLevelEmphasis,' emerges prominently. This feature measures the joint distribution of long run lengths and high grey-level values, contributing to the discriminative power of the model. A scatter plot illustrating the two selected features, with NAC classes represented by different colors, is depicted in Figure 4.14. It is observed that the two features exhibit a modest correlation.

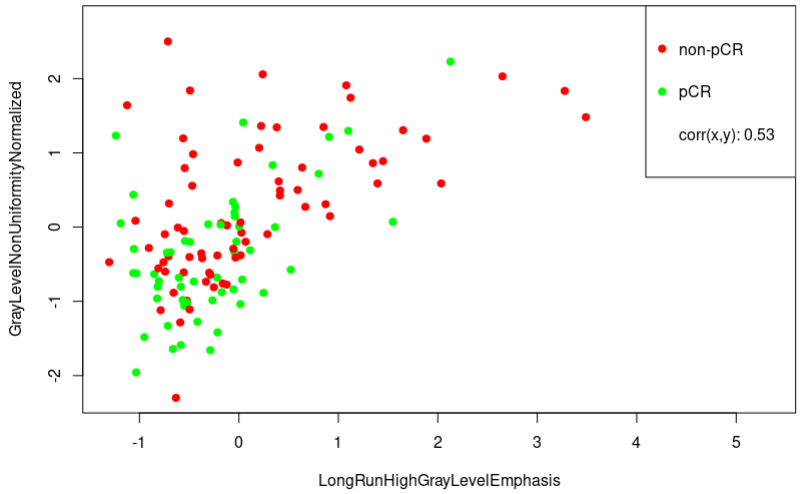


FIGURE 4.14: Scatterplot of features selected with LASSO.

UDFS + logistic model

In the case of UDFS, the filter with the highest performance is the wavelet f17, where 4 and 5 features yield similar AUC values, respectively 0.85 ± 0.01 (CI = 0.83-0.87) and 0.84 ± 0.01 (CI = 0.82-0.87). Opting for model simplicity, the version with only 4 features is chosen.
The selected features originate from both the second dynamic and third subtracted images. One can see a summary of the frequency of selected features in Figure 4.15.
Specifically, two first-order features, 'Maximum', are present in both the dynamic and subtracted images. Additionally, the first-order feature 'Kurtosis' from the subtracted image and the glszm feature 'GrayLevelVariance', representing the variance of grey levels in the zone, are included. As the images are filtered, direct interpretation of these features is not immediately apparent. A view of the selected features is presented as scatterplots in Figure 4.16 and 4.17.
One can observe that UDFS feature selection, compared to mRMR, allows selected features to exhibit more correlation. However, correlation does not always imply redundancy. In this case, the use of two similar features, such as the maximum of both the dynamic and subtracted images, enhances the model's predictive capability to some extent. However, despite the selection of first-order features, their interpretation should be done cautiously, as they originate from filtered images and may not carry the same significance on original images. Moreover, limitations inherent in the small dataset size prevent well-defined patterns from emerging within the scatterplots.

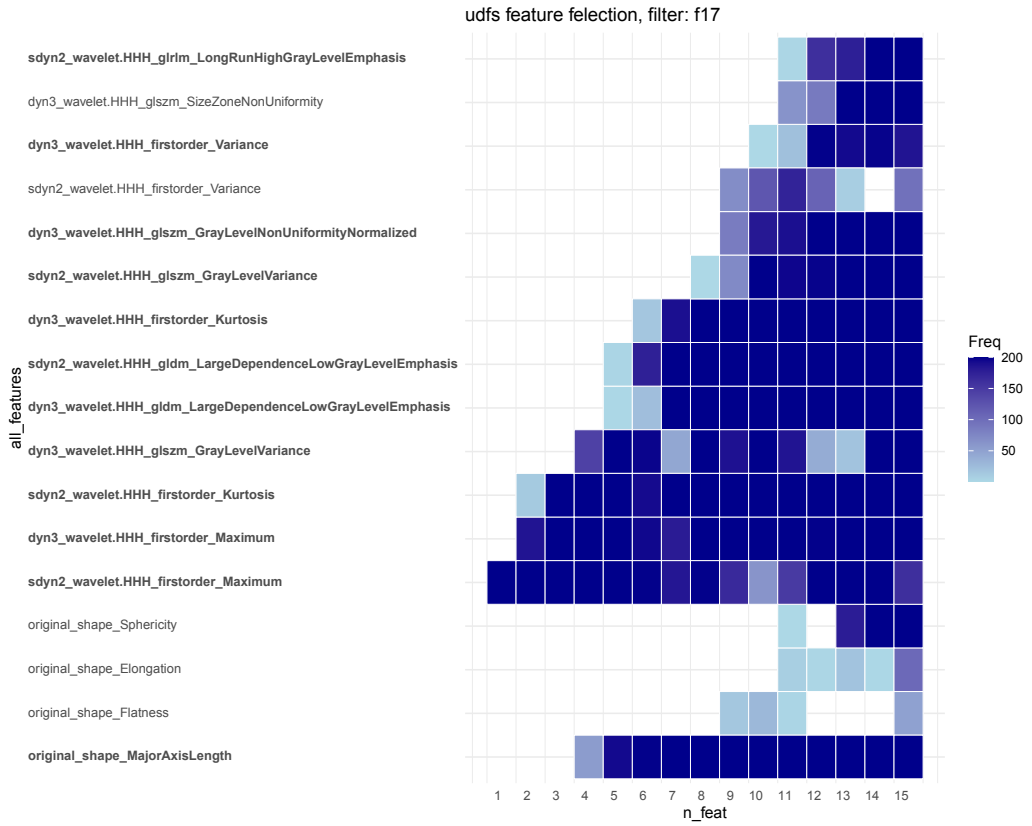


FIGURE 4.15: Frequency of selected features using UDFS.

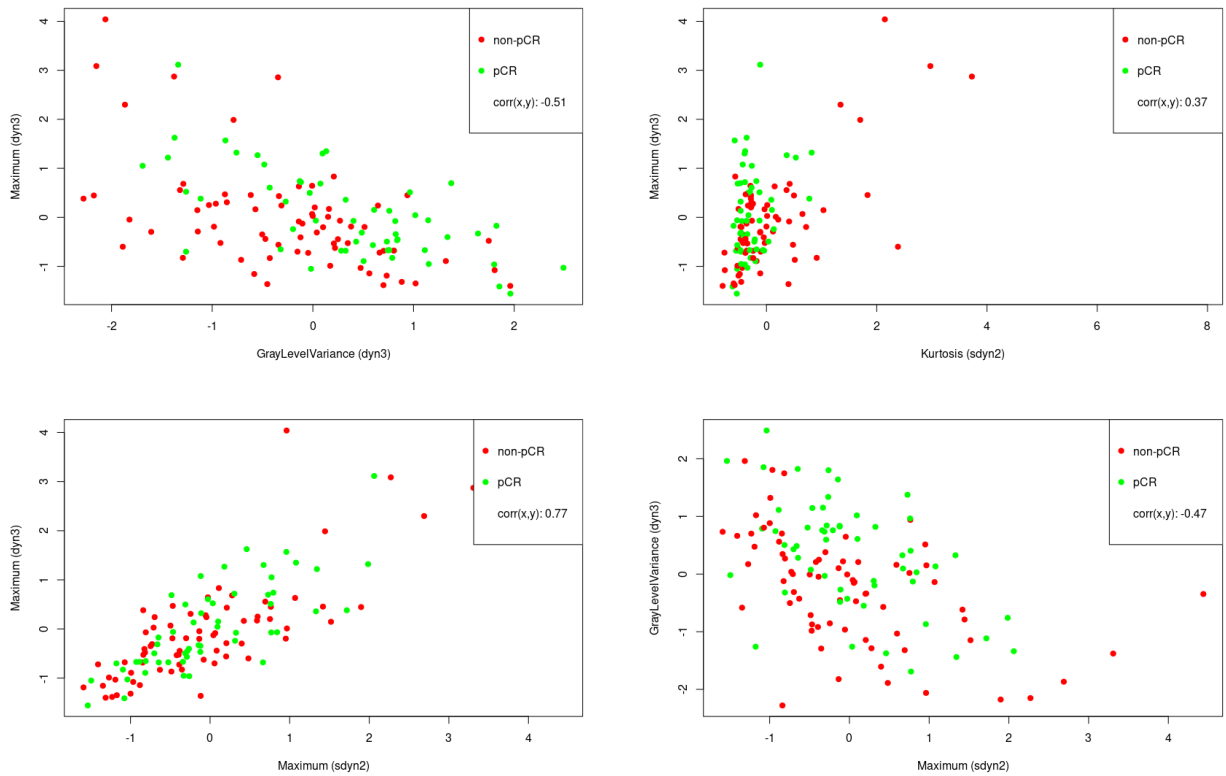


FIGURE 4.16: Scatterplots of features selected with UDFS part 1.

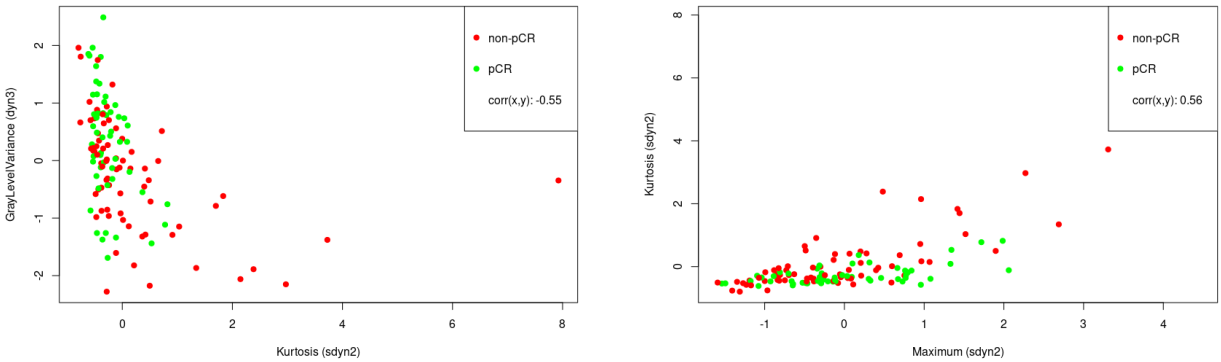


FIGURE 4.17: Scatterplots of features selected with UDFS part 2.

Selection of the best model and assessment on automatic segmentations

A summary of the performance of the chosen models for each selection method is presented in Table 4.9, showing comparable performances especially for mRMR and UDFS models. While the model with features selected by LASSO exhibits slightly lower performance, but still demonstrating good predictive capabilities.

model (manual seg.)	**n. radiomic features**	**$\overline{\text{AUC}}$**	**SE**	**CI (95%)**
mRMR+ logistic (radiomic + clinical)	1	0.83	0.01	0.81-0.85
LASSO + logistic (radiomic + clinical)	2	0.78	0.01	0.75-0.80
UDFS + logistic (radiomic + clinical)	4	0.85	0.01	0.83-0.87

TABLE 4.9: Summary of mean AUC in the best models, using manual segmentation

Before selecting the best model, stability must also be taken into account. Thus, the classifiers trained on manual data are tested again using radiomic features obtained from the automatic segmentations. However, the clinical variables remain the same as fixed by the operators. Results in terms of AUC are reported in Table 4.10.

model (automatic seg.)	**n. radiomic features**	**$\overline{\text{AUC}}$**	**SE**	**CI (95%)**
mRMR+ logistic (radiomic + clinical)	1	0.80	0.04	0.72-0.88
LASSO + logistic (radiomic + clinical)	2	0.80	0.04	0.72-0.88
UDFS + logistic (radiomic + clinical)	4	0.81	0.04	0.73-0.88

TABLE 4.10: Summary of mean AUC in the best models, using automatic segmentation

One can notice that despite having a larger variability, values are still comparable with the manual ones. It is important to note that clinical variables did not change, so the only source of instabilities

in the models comes from the included radiomic features, which in this case did not show significant differences. Considering the first set of models as the gold standard, based on the radiologist segmentation, one can observe that differences between the feature selections offered by mRMR and UDFS are not significant. However, the number of selected features is quite different. The feature selection by mRMR finds only one relevant radiomic feature, whereas UDFS requires four features to reach similar results. Additionally, it's notable that the single feature chosen by mRMR is also one of the two selected with LASSO, and in general, both mRMR and LASSO select features from the same filter. This suggests greater reliability in using that filtered feature group.
Overall, accounting for the most parsimonious and reliable model, the choice falls indeed to the mRMR feature selection with a logistic classifier. This model achieves good performances with features of both manually segmented and automatically segmented ROIs.

In conclusion, the adopted framework, integrating stability assessment, filtering of non-relevant and correlated variables, along with focused feature selection, has facilitated the development of a robust predictive model employing a simple logistic classifier. The evaluation of the model using features extracted from automatic segmentation has demonstrated promising performances, indicating the potential utility of both the model and automated segmentation in future studies.
The abundance of radiomic features necessitates cautious analysis to mitigate the risk of overfitting. The exploration of various techniques has enabled the tailoring of the radiomic pipeline to the constraints of the small available dataset, providing bespoke solutions. However, it is crucial to acknowledge the limitations, particularly the absence of external validation, which restricts the generalisation performance to the internal dataset. Future efforts should prioritise external validation to ascertain the model's robustness across diverse datasets and ensure its applicability in broader clinical settings.
In summary, this approach offers a methodical framework for leveraging radiomic features in predictive modelling, laying the groundwork for personalised treatment strategies in breast cancer care. Despite the challenges, the findings underscore the potential of radiomics in enhancing prognostic accuracy and guiding clinical decision-making.

Chapter 5

Conclusion and future directions

The objective of the study was to develop a predictive model for patients affected by malignant breast cancer to aid in the prior assessment of the usefulness of neoadjuvant chemotherapy regime. A dataset of DCE-MR volumes acquired before NAC administration was utilised to mine information linked to future therapy outcomes. Radiomics and machine learning techniques were integrated to build a full radiomic pipeline, encompassing ROI extraction, feature extraction, feature selection, and training of a binary classifier to predict the final outcome.

The development of a CNN for automatic segmentation enhances the efficiency and reproducibility of the process for further studies. Overall, the performance achieved in lesion segmentation reached a median Dice score of 0.73, with an interquartile range of 0.53-0.84. An assessment of performance variability between two radiologists who performed manual segmentation yielded a median Dice score of 0.79, with an interquartile range of 0.70-0.85. While automatic segmentation demonstrated slightly lower compatibility with manual segmentation compared to the agreement between two human operators, the difference was not substantial. The primary limitations of the segmentation pipeline, contributing to the presence of low Dice scores, primarily stemmed from false positives, including the additional detection of enhanced axillary lymph nodes and enhanced parenchyma. Further improvements to overcome these limitations involve increasing the dataset size to include more cases presenting such tissues, aiding the model in generalising better. Enhancements in the model architecture, such as incorporating attention layers or modifying the loss function, along with data augmentation, may offer additional assistance. Another limitation arose from hardware constraints, necessitating the use of small patches to preserve high resolution while sacrificing a global view of the image. Improvements in workstation hardware, such as the integration of a GPU capable of handling larger images or attempting to amalgamate information from all patches within a unique CNN at multiple scales, could enhance the capture of global information. This could prove beneficial in reducing false positives in axillary locations or exploiting symmetry in the contralateral breast to reduce false positive detections in breast parenchyma. Additionally, the entire procedure was cross-validated solely on an internal dataset. To comprehensively assess segmentation performance and generalisation capability, validation on an external dataset is essential. While a publicly available DCE-MRI dataset was published by [118], they only provided lesion slices, rendering it unsuitable for validating the entire procedure, as the complete image is required. Generally, the availability of public databases is limited, with most cases utilising only 2D acquisitions, making external validation challenging.

Various models and feature selection techniques were explored. While more complex models like SVM and RF exhibited a tendency to overfit, simpler logistic models proved adept at circumventing this issue. Feature engineering methods like PCA did not significantly contribute, generating non-discriminative new features. Different feature selection methods, including mRMR, LASSO, and UDFS, yielded feature subsets with comparable performances when paired with the logistic classifier. The incorporation of clinical data obtained prior to NAC further enhanced the predictive model's performance, allowing a more comprehensive assessment of patient-specific factors. The best model, utilising mRMR feature selection, along with the inclusion of clinical variables and a logistic classifier, yielded a mean AUC of 0.83 ± 0.01 with manual segmentation and 0.80 ± 0.04 with automatic segmentation. Results indicated

comparable performance, implying negligible differences in segmentation for the prediction task and the feasibility of using automated segmentation in future studies. Overall, the selection pipeline proved robust to ROI variation. However, limitations stemming from the dataset size and the absence of external validation persisted. Future investigations could involve deeper assessments of feature stability, encompassing a broader range of ROI perturbations and acquisition parameter dependencies. Customised discretisation for different feature groups and exploration of additional unsupervised feature selection techniques may further enhance model generalisation capabilities. Data augmentation methods for binary classifiers, could also be employed to mitigate overfitting. Furthermore, another limitation is that the radiomic features utilised were limited to traditional hand-crafted features, whereas emerging perspectives involve leveraging features extracted using deep learning algorithms. An intriguing frontier lies in exploring deep learning classifiers, which rely on convolutional neural network feature maps rather than individual features. However, the feasibility of this approach is constrained by the requirement for larger datasets to effectively train CNNs. Additionally, it's worth noting that all the data originated from a single centre and images are acquired from a single scanner. To enhance reliability and generalisability, the utilisation of multicentre data would be preferable. This would enable the validation and verification of findings across diverse clinical settings, contributing to a more robust understanding of the predictive capabilities of radiomic models in breast cancer prognosis.

Overall, despite the mentioned limitations, the results obtained are promising, indicating the potential for the model to aid clinicians in making informed decisions regarding the suitability of NAC for individual patients. Ultimately, the primary aim and significance of this research lie in developing a model capable of predicting the outcome of neoadjuvant chemotherapy. This model facilitates the administration of increasingly personalised NAC tailored to individual patients. By doing so, it helps mitigate overtreatment for patients who may respond well to milder therapies, while also minimising the adverse effects associated with NAC when it is not beneficial. Further validation and prospective studies will be needed to assess the clinical utility of this integrated approach and its impact on patient care. The developed pipeline provides a robust framework that could be applied in various other contexts beyond breast cancer, demonstrating its potential for broader applicability in medical imaging analysis, promoting the integration of machine learning models based on radiomics and precision imaging into clinical practice.

Appendix A

Literature review on segmentation techniques

A systematic literature review was conducted to explore the state of the art in breast lesion segmentation. The focus of this research was primarily restricted to the past decade (2013-2023) to emphasise the recent proliferation of deep learning techniques in this field. PubMed (US National Library of Medicine, http://ncbi.nlm.nih.gov/pubmed) served as the reference database for this review. To ensure a robust selection process, research papers were screened using keywords such as 'breast MRI lesion segmentation,' along with related synonyms and abbreviations.
The bibliographic search on PubMed yielded 314 results. After excluding 27 papers that were either unavailable or not published in English, and an additional 138 papers deemed off-topic for involving other imaging techniques or organs, or were general reviews, 149 papers remained. Out of these, 109 were found to lack specificity about segmentation but included it at some stage of the work.
Among those more generic studies manual segmentation methods were predominant (47 out of 109), followed by various semiautomatic techniques, including prebuilt software (6 cases), level set method (2 cases), Chan-Vese algorithm extension (1 case), manual segmentation combined with thresholding or thresholding followed by manual refinements (12 cases), and thresholding followed by the convex hull algorithm (2 cases). Region growing algorithms (using the random walker) were applied in 10 cases. Shifting focus to automatic segmentation, a substantial portion of the remaining works (27 cases) relied on clustering techniques. Fuzzy C-means was the most commonly used clustering algorithm, with a minority employing k-means and very few cases exploiting Gaussian Mixture Models. In addition to clustering, less common alternatives included the utilisation of Markov Random Fields and active contour models such as Gradient Snake Vector Flow. Out of the remaining 40 papers, various segmentation techniques were observed. The main findings are presented below, organised chronologically based on publication date.
Jayender et al.[43] proposed a novel approach to segment breast lesions in DCE-MRI by utilising time series analysis based on the modelling of linear dynamic systems (LDS) and Fuzzy C-means clustering (FCM). Time intensity curves were employed to generate an observation vector, and to estimate the parameters of the LDS, which represented the most dominant dynamics of the system. These parameters were then fed into the FCM algorithm, which classified the voxels as tumorous or healthy. Segmentation results were evaluated by comparing them with both manual segmentations by radiologists and segmentations provided by commercial software. The study demonstrated promising results in terms of Dice score (0.72-0.77). However, it is noteworthy that the dataset was relatively small, consisting of only 24 patients, each with diagnostic DCE-MRI and T2-weighted images, as well as sagittal post-contrast images.
A different method was proposed by Hong et al. [38], exploiting a joint framework able to both perform lesion segmentation and registration in a breast DCE-MRI dataset. Their method exploited an unified energy functional containing contribution from both segmentation and registration energy. Despite showing nearly optimal results with a Dice score higher than 0.9, the evaluation of the method was performed on a really scarce dataset constituted by only 6 subjects.
Chang et al.'s [13] work does not directly involve lesion segmentation but provides a simple yet efficient method for segmenting breast tissue using only three lines in an automated manner. Their approach is

based on automatically determining three marking points. The first point is obtained by intersecting the middle vertical line, which intersects the midline of the skin anterior to the sternum, with the chest wall. The other two points are obtained by intersecting fixed vertical lines, located to the left and right of the midline crossing the sternum, with the horizontal line passing through the first point. Straight lines with fixed angles to the horizontal line are then drawn from the lateral points. This preliminary step of segmenting breast tissue can serve as a useful preprocessing step before lesion segmentation, aiding in the localisation and delineation of breast regions of interest.

Sim et al.[91] present a system for defining the breast ROI and identifying and classifying breast lesions in DCE-MRI. While the primary focus of the work appears to be on classification rather than segmentation, it also offers insights into the segmentation process. The pipeline involves defining the breast region followed by lesion identification. The first stage includes thresholding to remove values below 75% of the maximum signal intensity, followed by the utilisation of both horizontal and vertical projection profiles to determine the vertical and horizontal breast boundaries, respectively. After overlapping the segmented horizontal and vertical images, Sobel edge filtering is applied to obtain the final breast ROI. Subsequently, a discrete Fourier transform is employed to enhance the images, followed by a kinetic approach based on contrast uptake and washout characteristics applied to one pre-contrast and five post-contrast images to generate a composite image. Lesions not bounded in three dimensions are then rejected. However, the overall performance was reported only for the classification task.

McClymont et al.[61] developed a method for segmenting breast lesions based on mean-shift clustering and the utilisation of graph cuts on a region adjacency graph. The approach involved a training set comprising 35 subjects and two separate test sets containing 93 and 9 lesions, respectively, acquired through various modalities including T1 and T2 weighted axial images, and 3D DCE-MRI images. Preprocessing involved the creation of vector-valued voxels obtained by stacking values from each modality to provide input to the model. As a preliminary step, an initial 3D mask of the whole breast tissue was obtained using the Otsu thresholding method, followed by opening and hole filling. Subsequently, a 3D mask of the chest wall was computed using Hayton's algorithm. The final mask of breast tissue was derived from the intersection of the voxel anterior to the chest wall and the initial mask, thereby reducing computation time and minimising false positive detection in the chest cavity. Lesion segmentation was then applied within the restricted breast ROI. This included mean-shift clustering to group similar voxels in the multidimensional feature space, followed by the construction of a region adjacency graph over the obtained clusters. Vertices identifying suspicious tissue were determined by comparing the difference between post- and pre-contrast signals, and segmentation was performed using graph cuts. Blood vessels and too small candidates were rejected, while among the remaining candidates, only those with the highest post-contrast enhancement were retained. Segmentation performance was evaluated against the ground truth manual segmentation of an experienced radiographer. The two test sets corresponded to the same scanner and a different scanner compared to the training set. The median Dice scores for the two test sets were 0.76 and 0.75, with interquartile ranges of 0.17 and 0.16, respectively. Overall, the segmentation method demonstrated efficacy in fully automatic segmentation using multimodal MRI data.

Liu et al.[57] proposed a method for segmenting lesions in breast DCE-MRI using a background distribution active contour model based on level sets. The approach involved an initial manual selection of the ROI containing the entire lesion with enhancement. Subsequently, the active contour model was applied within the selected ROI. The model, based on physiopathological information, was tuned by utilising background information to generate thresholds derived from the background intensity distribution, which is homogeneous in contrast to the complex distributions of lesions. The dataset comprised 42 DCE-MRI images acquired axially. Segmentation results were compared with manual segmentations by two radiologists, demonstrating good performance with a mean Dice score above 0.79.

Levman et al.[52] proposed a semi-automatic model for segmenting lesions in breast DCE-MRI. Their

approach involves a manual initial step, where the operator defines an ellipse in the image containing a suspicious lesion, followed by another ellipse encompassing normal tissue. Subsequently, segmentation is formulated as a supervised learning problem with a tunable parameter. However, no ground truth was used to evaluate segmentation efficacy, and method performance was estimated only for subsequent specific classification tasks.

Gubern et al.[32] focus more on classification rather than segmentation; however, they proposed a method that includes a lesion segmentation stage. In the first stage, the breast is automatically segmented using three probabilistic atlases to account for individual variability. Then, the first post-contrast image in DCE-MRI data is used to detect suspicious voxels in the defined breast ROI. A random forest classifier was trained on features based on relative signal enhancement and blob structure measures extracted from labelled voxel data. Lesion candidates are then obtained by identifying the local maxima of voxel likelihood, which are subsequently used as input seeds for a segmentation software.

Hu et al.[39] propose an alternative segmentation method for lesions in breast DCE-MRI based on the construction of high-dimensional images, where multiple time points are considered, providing an $n \times T$ image where n is the number of pixels and T is the number of time points, preserving the full information of the time intensity curve. The image dimensionality is then reduced using Laplacian Eigenmaps, preserving the essence of the manifold. A 3D feature image with a more evident lesion area is obtained, and to the latter, a standard k-means clustering technique is applied to segment the lesion. Automatic segmentation results were compared to manual ones made by radiologists, showing high specificity (0.98 ± 0.01) and sensitivity (0.90 ± 0.04). However, the considered dataset was relatively small, with only 36 cases.

Dalmics et al.[20] study is not solely focused on segmentation; however, it provides insights into semi-automated lesion segmentation in MRI by employing a multiseed smart opening algorithm. Their algorithm is based on the user selection of multiple seed points in the lesion and the use of 3D region growing operations, followed by morphological operations such as erosion and dilation with optimally tuned strengths. The regions of interest identified by each seed are then merged to form a common ground truth. Experiments involving the shifting and removal of some of the initial seeds showed segmentation with optimal compatibility in terms of Dice score (about 0.85 ± 0.19 and 0.88 ± 0.16, respectively). Their dataset included 325 patients, with a total of 395 lesions, suggesting considerable reproducibility of semi-automatic segmentation with different operators placing the initial seeds.

Ertas et al.[24] focus on the segmentation of the whole breast region. The breast tissue was segmented using bias-corrected fuzzy c-means clustering. The database consisted of 82 patients with variability in breast tissues (fatty, fibroglandular, and dense) who underwent MRI in different centres equipped with various scanners. For each patient, T1-weighted images acquired in the coronal plane were utilised. The clustering is applied slice by slice, followed by refinements including 2D hole filling, search for 2D connectivity, and removal of small objects. The breast area is then defined as the largest connected component. Slices are merged to form a 3D volume, and 3D morphological opening followed by closing are applied to reduce over-segmented and under-segmented cases. Then, they separate the left and right breast using the line in the axial plane crossing the midsternum, ensuring that this could be useful for further lesion segmentation. Overall, the performance of their model was good, obtaining a relative overlap (Jaccard index) > 0.77 in all cases. However, they remarked that simply defining the breast volume as the tissue between the breast chest wall and breast air boundary is ambiguous, especially regarding the extent of lateral and superior regions, which still lack an accepted definition. This leaves the choice to the operator to include or exclude the axillary tail of the breast.

Fusco et al.[26] focus more on lesion classification rather than segmentation. However, they detailed an automated segmentation procedure for breast DCE-MRI. Their process includes three steps: first, they extract a breast mask using Otsu thresholding on the image created considering a parametric map of the sum of voxel-wise intensity differences. Second, they perform hole filling and removal of leaks

using morphological erosion and dilation. Finally, the third step involves extracting suspicious voxels from the ROI by imposing two conditions: selecting voxels for which the maximum of normalised TIC is greater than 0.3, and requiring that the maximum intensity is reached before the end of the scan time. These conditions ensure the selection of significant contrast uptake and include plateau and washout patterns. However, comparison between manual segmentation and automatic segmentation was provided only in terms of classification performance, which limited the interpretation of the results to the specific task.

Gui et al. [33] provide a general segmentation technique for compact objects that show application with different imaging techniques and organs, including MR images and breast lesions. Their model is based on the assumption that the shape is a prior knowledge for segmentation, since many lesions and organs show compact shapes that resemble circles, ellipses, and variations. Such prior shapes can serve as constraints in the segmentation process. They propose an isoperimetric constraint based on the ratio of perimeter to enclosed area of an active contour, acting as a regularisation term in a level set segmentation scheme. This way, the active contour model's boundaries maintain a compact shape resembling the required organs or lesions in the medical images. However, metrics of segmentation performance were not reported, showing only visualisations of the results, with a focus on datasets for other organs and techniques.

The work of Chen et al.[15] is not focused on segmentation but provides insights into their procedure, which includes a first step of breast segmentation using a chest model with three body landmarks. The template was co-registered to each subject's chest region, and then the chest wall muscle was identified using an edge detection algorithm and curve fitting. The two breasts were then separated at the sternum. Next, the ROI was enhanced using an unsharp filter based on the inverse of the Laplacian filter, and the tumour was segmented using fuzzy c-means clustering.

Wang et al.[108] focus on a generalisation of the level set method to segment breast lesions in 3D DCE-MRI images. Their method includes the addition of a 3D shape-weighted value according to changes in evolution, aiming to improve both the contour to be closer to actual tumour margins and to eliminate background noise. They compare both 2D and 3D models on 3D images, highlighting that the use of 3D techniques is superior since the upper and lower slices of images are connected. However, most of their results pertain only to blurred computer-simulated images, and only three cases were presented for real breast MRI data.

Kumar et al.[48] propose a modification of FCM clustering based on superpixels to segment suspicious lesions in multiple organs from images acquired using various techniques, including breast MRI. Their approach is based on dividing the image into superpixels and partitioning those superpixels into a certain number of fuzzy clusters, incorporating the use of spatial information by exploiting neighbouring and similar superpixels. However, they provided evaluation for only three cases of breast MRI.

Illan et al.[40] delineate a segmentation method for challenging non-mass-enhancing lesions of the breast for DCE-MRI. Their method exploits independent component analysis (ICA), assuming that each voxel of the DCE-MRI image contains a signal representative of the contributions of enhancement kinetics of different contributing tissues. A set of time signals is then analysed as a blind source separation problem, expressing different dynamic behaviours as a linear combination of a reduced set of sources. These sources can then be used as features to perform classification. ICA aims to maximise the statistical independence of the set of sources by working on temporal curves at the voxel level. Each voxel intensity can be considered in terms of a linear decomposition of temporal sources, whose coefficients are the scores. These scores are then used as features to perform tissue classification. High scores are related to lesions, whereas low scores correspond to normal tissues. The analysis of other components can also be used to discriminate between benign and malignant cases. As preprocessing, they performed registration to the pre-contrast volume and applied 3D Gaussian smoothing. They also reduced the size of the images by finding the middle chest point and discarding all the content of the image after that point, reducing heart and other organ signals. The dataset consisted of a small

set of 16 patients who underwent coronal T1-weighted DCE-MRI. The obtained segmentations were compared with the manual ones provided by the radiologists, obtaining a maximum Dice score of about 0.53, considering the challenging task and the small dataset.

Chen et al.[14] developed a semi-automatic, time-saving contouring method to segment tumours in 3D breast MRI. Their procedure is based on the independent use of 2D level set segmentation in each plane (transverse, coronal, and sagittal), which can be intersected or merged to create a full 3D contour. The choice of 2D segmentation aims to overcome the time-consuming nature of 3D level set methods. Images were preprocessed by applying anisotropic diffusion, Wiener filtering, and greyscale adjustment to denoise while preserving fundamental boundary information. The operator is then required to select a rectangular ROI from the transverse plane on a 2D slice, delineating the box containing tumour boundaries in three slices: the first, the last, and the middle, to capture the vanishing and maximum diameter of the tumor. Then, the 2D level set method is applied slice-wise, using a 2D continuous function to describe tumour boundaries, achieving relatively fine resolution. The 2D contours are then used to produce two final 3D segmentations: one obtained by merging and the other by intersection. The study was tested on 20 images, including axial T1 and T2 weighted images and 3D images with contrast agent, providing a mean Dice similarity index on the transverse plane of about 0.92 for the best case obtained by merging all 2D results, while 0.86 for the intersection case. However, they also reported that in 2 out of the 20 analysed cases, the procedure proposed defective segmentation due to the preprocessing procedure.

Spuhler et al.[95] focused on breast lesion segmentation for DCE-MRI with deep learning using the U-Net architecture. Their database consisted of 263 training subjects and 54 test subjects, each patient having a sequence of one pre-contrast image and four post-contrast images acquired in the sagittal plane. Segmentation performance was assessed by comparison with ground truth provided by three different radiologists. The U-Net employed two input channels: one for the first pre-contrast image and one for the first post-contrast image, aiming to capture the same information as the radiologists performing manual segmentation. To augment the data, they performed small random rotations (up to 5°), and the network was fed with individual 2D slices. After selecting the optimal segmentation threshold from the training set, comparison with the radiologists yielded a mean Dice score on the test set in the range of 0.61 to 0.71, comparable to the inter-operator mean variability (ranging from 0.61 to 0.68). Furthermore, they evaluated the impact of their automated segmentation by assessing the performance of a radiomic model to predict lymph node metastases based on features extracted from the provided segmentation. They found no significant differences in the AUC, indicating that the U-Net approach for segmentation achieved predictive power comparable to that of the radiologists for that specific task.

Vogl et al.[107] proposed a multiparametric segmentation approach for breast lesions, incorporating 3D DCE-MRI, diffusion-weighted imaging, and ^{18}F-fluorodeoxyglucose-PET. Their algorithm involved extracting local textural, intensity-based, and kinetic image features and training a random forests classifier for voxel-wise segmentation based on these features. A preprocessing step restricted the region to the breast area using a region-growing algorithm. Intensity features were derived from changes in contrast over time in DCE-MRI and other imaging techniques. Their dataset comprised only 34 patients, with the classifier trained using 1000 randomly selected samples per class and per patient, and validation performed on a new patient using leave-one-out cross-validation. They reported a mean Dice score of about 0.58 using only DCE images, which improved with the addition of features from other techniques to 0.67. This suggests that segmentation performance significantly improves with the inclusion of features from diffusion-weighted images as well.

Verburg et al. [105] proposed various methods for performing chest wall segmentation, particularly in cases of extremely dense breasts. Although not directly related to lesion segmentation, breast wall segmentation is often a crucial preprocessing step for subsequent analysis of breast MRI images. They tested two automated methods: one knowledge-based and the other based on deep learning with a

CNN. In the knowledge-based method, a rectangular ROI was automatically defined, containing an area 5 cm posterior to the intermamillary cleft and 1 cm anterior to the breast tissue. Background and foreground voxels were separated using Otsu thresholding, and landmarks were placed at the most anterior tissues in the image data and at the most posterior air-tissue boundary. The volume was then cropped to this ROI. The knowledge-based method aimed to find the curve dividing the chest wall and breast in each image through two steps. The first involved creating a cost image that assigned high cost in locations where the chest wall was unlikely to be present, whereas low cost was set to the edge contrast zone. The second step used a path-finding algorithm to track paths accumulating the smallest sum of cost values, finally finding the breast wall. The second method utilised a dilated CNN with increasing spacing between kernel elements. The CNN was directly applied to 2D slices along the three principal axes to obtain 3D probability volumes. These three components were averaged and thresholded at P=0.5 to obtain the final binary prediction delimiting the region anterior and posterior to the chest wall. Their dataset comprised 115 T1-weighted MR images of extremely dense breasts, with 79 images used for testing. Both models performed well with no significant difference, holding a mean Dice score for the knowledge-based method of 0.982 ± 0.006 and for the deep learning model of 0.984 ± 0.008. They also reported that from a practical standpoint, the deep learning approach ran faster and provided better performance in terms of false negative samples.

In the work of Li et al. [54], although the main focus is not on segmentation, the authors provided a customised semi-automatic lesion segmentation method for DCE-MRI images. Specifically, they implemented a regional growth algorithm, wherein the radiologist places an initial seed on the lesion, and the region growth criteria involve searching for connected pixels within a certain threshold defined by the Otsu method, without the need for fine-tuning the threshold by the radiologist. However, as segmentation is not the primary topic of the work, performance metrics were reported for only a couple of cases, with a mean Dice score near 0.89.

In the work of Piantadosi et al. [77], the focus is on segmenting 3D DCE-MRI of breast parenchyma from the background of hair and other tissues such as the pectoral muscle, chest wall, and the heart. This segmentation stage serves as preprocessing prior to lesion detection, reducing noisy tissue from other organs and also the computational burden. Their proposal involves the use of a 2D CNN approach, such as the U-Net, on 3D images by training three models, one for each plane (coronal, sagittal, and transverse), exploiting specific advantages of each. The sagittal plane is useful to enhance armpit cavities and better reject sternum and heart tissues; the transverse plane helps detect the pectoral muscle, whereas the coronal plane is better for detecting the breast air boundary. The U-Net consists of a U-shaped path, involving a contraction path progressively decreasing the spatial size while increasing the feature size, and an expansive path doing the opposite until returning to the original size. To reconstruct the final image, different approaches have been tried, including merging the predictions using a pixel-wise AND operator, intersecting the results using a pixel-wise OR, employing a majority voting, and a weighted majority voting from the three predictors. Training and evaluation were conducted independently on two separate datasets, one with 42 subjects and the other with 88 subjects. Results were provided for both cases, obtaining a median Dice score above 0.95 for all the reconstruction procedures, but showing slightly higher performances for the weighted majority voting approach. Furthermore, they evaluated the use of this approach when trained with one dataset and tested on the other, observing that when the model is trained on a dataset with small resolution and then applied on a dataset with larger resolution, the approach might fail (obtaining a Dice score of about 0.55 ± 0.08) due to the poor adaptation of the first layer of the CNN to different resolutions, whereas the opposite approach seems to still be working (offering a Dice score of 0.91 ± 0.03).

In the work of Çetinel et al. [12], while the primary focus is not solely on segmentation, they present a two-step approach for lesion segmentation in DCE-MRI. The first step involves segmenting the breast region, which includes adaptive noise reduction followed by local adaptive thresholding and connected

component labelling to eliminate undesired areas, resulting in the extraction of the breast region. Subsequently, a horizontal projection is obtained and used to determine a cutting line to separate the breast area. The breasts are then separated into left and right, retaining only the side containing lesions. The second step focuses on true lesion segmentation, which begins with Otsu thresholding with 32 classes. The result of this thresholding is further refined using Markov Random Field methods. However, no segmentation performance metrics were reported to compare with a manual ground truth.

In the work of Min et al.[66], the focus is on breast lesion segmentation and characterization in breast MRI data. Their system begins with preprocessing, segmenting the breast region using Otsu thresholding and Hayton's algorithm from the T1 weighted images. The breast volume is then split into left and right using the middle line in the breast mask, allowing for the processing of one breast at a time. Next, region candidates for lesions are generated using 3D multiscale morphological sifting, which employs linear structuring elements to obtain lesion-like patterns at three different scales (original, 1/2, and 1/4). Subsequently, multilevel Otsu thresholding and size thresholding are performed. Features regarding intensity, texture, morphology, and kinetics are extracted from region candidates using all available sequences, including T1, T2 weighted, and DCE-MRI. This approach generates multiple candidates, which are then classified into lesion or normal tissues using the extracted features via the random under-sampling boost technique based on boosted decision trees, enabling treatment of class imbalance simultaneously. After classification, some of the detected lesions show significant overlap and can be fused to generate a final segmentation contour. Their dataset comprises 117 cases, each including a T1 weighted image, a T2 weighted image, and a sequence of five DCE-MRI scans. They used 58 cases for training and 59 for testing, with ground truth segmentation provided by expert radiologists. The method achieved an average Dice similarity score of 0.72 $\pm$0.15 and a median of 0.74.

In the work of Zhang et al. [117], the focus is on the segmentation of breast lesions in DCE-MRI datasets using a mask region-based CNN approach. The input comprises the pre-contrast image (used to identify the chest area) and the subtracted images of the diseased breast and the contralateral normal breast. Various predefined shapes and distributions of bounding boxes are used to identify abnormalities in the images, with the analysis conducted on 2D slices. These bounding boxes are then ranked and classified to determine if they contain lesions or not, after which a segmentation network is employed to determine tumour boundaries. The pretrained ResNet-101 serves as the backbone for the CNN. For training, 241 cases with fat-saturated sequences were used, while testing was conducted on 98 cases with non-fat-saturated sequences. Ground truth was obtained by fuzzy C-means clustering on the second subtracted image, following the selection of a square ROI corresponding to the maximum intensity projection and the application of an unsharp filter to enhance tumour boundaries. The output of the CNN was compared to the ground truth, revealing that small tumours were sometimes challenging to detect, leading to false negatives, while strong parenchymal enhancements occasionally caused false positives. For true positive cases, the reported range of Dice scores was 0.31-0.97, with a mean of 0.79. Nevertheless, the model proved robust for both fat-saturated and non-fat-saturated images.

In the work of Vidal et al.[106], a deep learning method for breast lesion segmentation in DCE-MRI is proposed, leveraging a modified U-Net architecture. The approach involves training three 3D patch-based U-Net models, each taking different combinations of temporal sequences from the DCE-MRI dataset, effectively utilising a full 4D dataset. The dataset consists of 46 patients, each undergoing 3D DCE-MRI, providing one pre-contrast image and between four to six post-contrast T1-weighted images. Preprocessing steps include generating a breast mask via landmark detection to eliminate noise from other enhanced organs, intensity normalisation, and extraction of balanced patches. Patch extraction ensures handling of class imbalance by extracting an equal number of patches representing the lesion minority class and the background class, up to 2000 patches for each. The three U-Net models differ in their input data: the first model takes inputs based on three time points (pre-contrast, second post-contrast, and last post-contrast), the second model uses the full time series (pre-contrast

and four post-contrast volumes), while the last model incorporates pre-contrast and last post-contrast volumes along with a customised volume obtained by computing the standard deviation over the time series at each voxel to better encode time-intensity variations. Performance evaluation is conducted using 5-fold cross-validation on the 46 cases, with ground truth masks provided by radiologists. False positives are attributed to MRI volumes including secondary lesions and other confounding findings not corresponding to malignant lesions, such as axillary lymphadenopathy and vessel enhancements. The ensemble of the three models is evaluated using both majority voting and the union of the three models' output, resulting in an overall mean Dice score of about 0.68, but a higher mean Dice of about 0.80 when considering the main lesion only. Testing the impact of using isotropic voxel size suggests that maintaining the original image resolution yields better results. Segmentation results are compared to those obtained with FCM, showing significantly higher performance for the U-Net ensemble. Overall, the study demonstrates that fully automatic methods without requiring user selection of the ROI can handle breast lesion segmentation, though the task remains challenging with small datasets.
In the work of Galli et al.[27], a three-time-points (3TP) framework is proposed for segmenting breast lesions in DCE-MRI. The method aims to fully utilise contrast agent kinetic information by employing three significant time points of the DCE sequence as inputs for a modified 2U-Net, which processes the images slice by slice. The procedure begins with preprocessing steps, including generating a breast mask to remove extraneous tissues and performing motion correction to align slices across different acquisition times. The breast mask is obtained using a multiplanar 2D U-Net, trained separately on each anatomical plane, with the outputs merged using a weighted voxel-wise combination. From the masked image, the 3TP slices are extracted, creating 3-channel images by stacking slices corresponding to specific time points: the pre-contrast image, the post-contrast image after 2 minutes, and the post-contrast image after 6 minutes. These 3-channel slices are then divided into left and right breasts and provided as input to the U-Net, which performs slice-wise segmentation. During training, a balancing of slices containing only lesions and slices containing healthy tissue is conducted. Performance evaluation is carried out using 10-fold cross-validation on a dataset of 33 patients with T1-weighted DCE-MRI images, resulting in a median Dice score of approximately 0.70 considering both lesion and healthy slices in the computation. The study highlights the importance of evaluating performance on slices containing both lesions and healthy tissue to avoid overestimating performance, as previous studies often focused solely on lesion-containing slices. Additionally, the authors compared their customised U-Net with the original basic U-Net and U-Net++, finding slightly lower median performances with the original models (median Dice scores of approximately 0.67 and 0.65, respectively).
In the work of Qin et al.[80], the aim is to enhance the U-Net model for breast lesion segmentation in DCE MRI data. The proposed approach involves a two-step procedure consisting of breast mask segmentation followed by lesion segmentation. For breast segmentation, a 2D U-Net is employed to delineate the region of interest (ROI) for lesion detection. Prior to training the lesion model, extensive data augmentation techniques are applied, including mirroring, scaling, and elastic deformations. An enhanced U-Net model is introduced, which combines a dense residual module with dilated convolution and recurrent attention modules. This modified architecture aims to improve segmentation accuracy. The dataset comprises 160 cases of T2-weighted MRI images, with 128 cases used for training and 32 for testing. The outputs are reconstructed by merging the 2D slices to obtain a final 3D reconstruction. During the breast segmentation stage, a mean Dice coefficient of approximately 0.92 is achieved, while the lesion segmentation stage reports a mean Dice score of around 0.78. These results outperform experiments conducted with standard U-Net, attention U-Net, residual U-Net, dense U-Net, and V-Net models, which yielded a Dice coefficient of about 0.76 on the same dataset.
In the work of Pandey et al.[71], an unsupervised method is proposed for lesion segmentation in DCE-MRI. The methodology involves multiple stages: 1. image subtraction and registration, to enhance lesion visibility and alignment for further processing; 2. phase-preserved denoising to reduce noise while maintaining edge and boundary details; 3. adaptive wiener filtering to preserve image features

crucial for accurate segmentation; 4. graph-based segmentation, where maximum flow and minimum cut problems are formulated in the continuous domain. This technique leverages continuous max flow methods to delineate lesion boundaries; 5. post-processing, undesirable regions are removed using morphological operations such as erosion and dilation. The dataset used in the study comprises 23 cases, and the performance of the proposed method is evaluated against manually segmented ground truth. The obtained mean Dice coefficient is approximately 0.92, indicating high segmentation accuracy compared to the manual annotations.

In the work of Yang et al.[112], although the primary focus is not on segmentation, the authors utilised Fast R-CNN for breast lesion segmentation in MRI. They employed a dataset consisting of 858 patients and reported an overall volume operator error, calculated as 1- Jaccard index of approximately 10%.

In the work of Peng et al.[74], a hierarchical 2D breast tumour segmentation model for 2D breast MRI is proposed. The procedure involves a two-stage process: first, breast segmentation using U-Net, and second, lesion segmentation using a customised lesion morphology aware network with a novel loss function designed to maximise inter-class distance, thereby reducing false positives and false negatives. The first stage employs a 2D U-Net to produce a 2D output mask for breast segmentation. The second stage utilises a modified approach based on adaptive deformable convolution, which adjusts the receptive field according to the lesion morphology using a learnable dilation factor. The lesion morphology aware network is constructed based on the structure of DeepLab V3+ with the ResNet-101 backbone. Their method is evaluated on a dataset comprising 590 patients who underwent T1-weighted contrast-enhanced MRI. During training and evaluation, only 2D slices containing more than 5 pixels of tumour are considered, resulting in a total of 5025 2D slices, with 20% reserved for testing against radiologist ground truth. The breast segmentation stage achieves a mean Dice score of approximately 0.96, while the lesion segmentation stage obtains a Dice score of 0.87 ± 0.14. Comparative evaluations with U-Net, DeepLab V3+, and hierarchical CNN show lower performance (mean Dice scores ranging from 0.83 to 0.85). Additionally, the automated segmentation method sometimes detects lesions missed by radiologists, which are subsequently confirmed upon review. However, it's important to note that these results are based solely on 2D slices already containing lesions.

The work of Zhu et al.[121] involves a stage of breast lesion segmentation for multiparametric MRI. Specifically, for the segmentation of DCE MRI images, they employed the V-Net architecture, constructing 3D images by tiling 2D slices from three time points of the sequence. Their dataset comprises 2823 patients, and the segmentation results were compared to manual segmentation, achieving a mean Dice score of 0.86.

The study of Rahimpour et al. [81], focuses on segmenting breast lesions in 3D DCE-MRI using a visual ensemble of CNNs. Preprocessing involves bias field gain correction, isotropic voxel size resampling, cropping to cover only the breast area and armpit, and image normalisation. Three models based on the U-Net architecture were trained. The first model used only the first post-contrast image as input, while the second and third models utilised both the first post-contrast and first subtracted images, employing either an image-level or feature-level fusion strategy. Each model produces a segmentation output, and a visual choice can be made from the radiologist. Their dataset comprises 141 patients who underwent T1-weighted DCE-MRI, with 111 used for training and 30 for testing, from various institutes and MRI scanners. The test cases were segmented by two radiologists, with a mean Dice score of 0.78 ± 10. Individually, the three models achieved Dice scores of 0.73 ± 0.23, 0.75 ± 20, and 0.70 ± 0.26, respectively. Considering the best visual choice from the radiologists, a mean Dice score of 0.79 ± 0.16 was obtained. Additionally, averaging and majority voting among the three predictions were evaluated, yielding a lower Dice score of about 0.75 ± 0.22.

In the work of Liu et al.[56], although the main focus is not on segmentation, some results regarding breast lesion segmentation in DCE-MRI are presented. They utilised a 2D-DenseU-Net based on the DenseNet-121 architecture. Their dataset comprised 82 cases, and the automatic segmentation results were compared to ground truth provided by radiologists, yielding a mean Dice score on the whole

dataset of approximately 0.72. Furthermore, for 30 cases, the variability across two different radiologists was estimated, resulting in a mean Dice score of about 0.97.
In the work of Zhang et al.[115], a breast lesion segmentation stage based on the nnU-Net model is introduced. The nnU-Net takes as input only the first post-contrast DCE image. Their dataset comprised 196 training cases and 50 test cases. The segmentation performance was evaluated by comparing it with manual segmentations provided by radiologists, resulting in a mean Dice score on the test set of approximately 0.83. Additionally, the segmentation results from different radiologists were compared, yielding a Dice score of about 0.83 on the training dataset.
In the work of Ru et al.[85], the focus is on breast tumour segmentation in various medical image types, including DCE-MRI. Their proposed model involves a neural ordinary differential equation (ODE) with an attention module to capture spatial and channel features. Unlike CNN models, the neural ODE can be treated as a continuous process and solved by traditional ODE solvers. The U-Net architecture is modified to include ODE blocks, reducing the number of trainable parameters, resulting in an architecture called Attention U-Node (Att-U-Node). Different datasets were used to test the model in both 2D and 3D contexts. The Att-U-Node 3D model was tested on 10 patients with 3D DCE-MRI sequences acquired, achieving a Dice score of about 0.7 through 3-fold cross-validation. Other models were also tested, with lower performances observed (U-Net3D and V-Net with Dice scores of about 0.65, U-NetR with 0.63, nnU-Net 0.61-0.63 depending on the application of post-processing, and U-Node3D 0.63). On the other hand, the 2D model Att-U-Node was tested on a dataset of 150 patients who underwent DCE-MRI, achieving a Dice score of about 0.70 through 3-fold cross-validation. Additionally, other 2D models were tested (2D U-Net, AttU-Net, and U-Net++ with Dice scores about 0.64, TransAttU-Net 0.66, U-Node 0.69). Their methods were also tested on a US dataset, showing promising results and suggesting model robustness and adaptability.
In the work of Sun et al.[97], a breast lesion segmentation subnet for MRI is employed as part of their model, which includes three stream decoders simultaneously identifying the lesion location, boundaries, and overall segmentation. The model utilises ResNet-34 as the building block and incorporates a location stream to focus feature extraction within the lesion location, preventing the loss of small lesions. An upsampling layer is then employed to aid in boundary refinement. The segmentation decoder utilises multiple scales of receptive fields to adapt to lesions of varying sizes using U-Net. To address class imbalance problems, data augmentation techniques such as rotation and flipping are applied. Their dataset consists of 248 patients divided into 70% training and 30% test sets, with ground truth provided by radiologists for each patient. The model achieved a mean Dice score of 0.70 on the test data. Additionally, various segmentation methods were tested, with inferior results observed for U-Net (Dice 0.61), U-Net+ (0.63), U-Net+ with data augmentation (0.65), AttU-Net (0.63), CE-Net (0.67), W-Net (0.60), and Shape Attentive U-Net (0.66). Other multitask methods also yielded lower results, ranging from 0.54 to 0.63 in Dice score. Furthermore, the utility of the three streams was evaluated, showing a slight decrease in performance when one or more stages were removed (Dice of 0.69 when only the segmentation stage was used).
In the work of Si et al.[90], the focus is on breast lesion segmentation in T2-weighted DCE-MRI. Their model utilizes multilevel thresholding based on Kapur's entropy, incorporating Gorilla Troops Optimization and Rotational Opposition-Based Learning. The dataset used comprises 100 slices from 20 patients, all with T2-weighted sagittal DCE-MRIs. Performance evaluation was conducted by comparing the model's results with ground truth provided by radiologists, achieving a mean Dice score of approximately 0.92.
In the work of Zhou et al.[119], breast lesion segmentation on DCE-MRI data is addressed using an improved U-Net model incorporating an attention module and edge feature extraction modules. Preprocessing steps included cropping the layer containing the lesion, normalisation, and augmentation techniques such as flipping, rotation, scaling, and shifting. Their dataset comprised 1110 patients for training, 476 cases for internal testing, and an additional 174 cases for external testing collected from

other centres. Performance evaluation was conducted by comparing the model's results with ground truth provided by radiologists, yielding a mean Dice score of 0.83 in the internal test set and 0.82 in the external set. Furthermore, the performance of other state-of-the-art methods was evaluated, with reported lower performances including SegNet, U-Net++, U-Net, Att-U-Net, CS-Net, and Multistage Hierarchical Learning, all yielding Dice scores ranging from 0.69 to 0.77.

A supplementary literature review was conducted using Google Scholar (http://scholar.google.it) to identify any additional publications not captured in the initial search on PubMed. While many of the findings overlapped with those from PubMed, only the key trends are presented here for conciseness, as they align closely with the performance range observed in the previously mentioned studies. As already found during the first half of the decade under review, the majority of studies focused on manual segmentation methods, as well as techniques such as region growing and interactive region growing. Some researchers also explored approaches involving region growing or spectral embedding preceding active contour models. Additionally, thresholding and multilevel thresholding techniques were widely utilised. Some studies experimented with a process involving thresholding followed by lesion candidate classification using common classifiers such as SVM. Clustering methods, including FCM and k-means, were also prevalent, with efforts made to enhance their performance by integrating spatial information or employing various optimisation algorithms. Furthermore, there were instances of semi-supervised methods being utilised to handle a reduced number of labelled slices. Additionally, rare cases of marker-controlled watershed methods were identified in the literature.
With the advent of deep learning, convolutional neural networks have gained significant popularity, particularly with the widespread adoption of architectures like U-Net. U-Net variations have emerged, including patch-based approaches, extensions to 3D U-Net, nnU-Net, utilization of residual blocks in Res U-Net, incorporation of multiple levels within the U-Net architecture, and integration with ConvLSTM structures to capture temporal information. In addition to U-Net, other hierarchical CNN architectures, recurrent neural networks , artificial neural networks, CNNs with multiple streams, multiscale strategies, hybrid 2D and 3D approaches, and deep neural networks with shape priors have also been explored. This trend reflects a growing interest in leveraging deep learning for segmentation tasks, particularly in extending these techniques to address challenges in 3D imaging datasets. Combining the findings from both avenues of research, it becomes evident that the prevailing trend in recent years is the adoption of convolutional neural networks (CNNs) derived from the U-Net architecture. A common approach observed across many studies involves the generation of a breast mask as a preprocessing step to narrow down the search area for lesion detection and segmentation. This strategy aims to improve the efficiency and accuracy of the segmentation process by focusing on relevant regions of interest within the breast tissue.

Appendix B

Notes on algorithm implementation and optimisation

Training large convolutional neural networks, such as the chosen V-Net, is inherently resource-intensive, demanding specialised GPU hardware and the application of various optimisation techniques to maximise performance. The significant computational load imposed by volumetric convolutions in the selected architecture renders CPU training impractical due to its slow processing speed. Consequently, the preferred approach is to train the model on a GPU, leveraging its parallel processing capabilities to accelerate computations. However, this approach is not without limitations, primarily stemming from the finite GPU memory, which serves as a bottleneck constraining both input size and model complexity. To manage the computational demands of training large CNNs, various strategies were explored and implemented:

- Limiting input size to smaller 3D images using patch-based approaches.
- Loading batches containing only one image at a time onto the GPU instead of storing the entire dataset in memory. This was achieved using the PyTorch DataLoader class, which provides an iterable over the dataset and enables the placement of data tensors in pinned memory for faster transfer to CUDA-enabled GPUs.
- Utilising the SGD optimiser (without momentum) instead of other optimisers, as it requires less memory by storing only gradients and not other parameters in the state memory.
- Implementing gradient accumulation to handle virtual larger batch sizes than the GPU's memory would allow. This approach enables the use of larger batch sizes, which can improve convergence and accuracy, particularly on workstations with small GPUs. Gradient accumulation involves modifying the final step of the training process. Typically, the training process includes loading the input batch, predicting the batch label, computing the loss function, and performing a backward pass to compute gradients and update weights after every batch. With gradient accumulation, gradients are stored, accumulating new gradients with each processed batch, and weights are updated only after processing several accumulation batches. Additionally, when resetting gradients and all tracking parameters, they are set to 'None' instead of being set to zero to better spare memory. For a quick implementation example in PyTorch, please refer to `https://kozodoi.me/blog/20210219/gradient-accumulation`.

Among these strategies, the most effective combination was using the DataLoader with batches containing a single image (or patch when higher image resolution was required) and leveraging gradient accumulation to achieve larger batch sizes. Interestingly, the impact of storing optimiser states was not as significant. Consequently, using improved SGD with momentum terms was feasible, as the stored parameters could still be managed within the available memory. Incorporating SGD with momentum and Nesterov momentum significantly accelerated the training phases. For a comprehensive overview of various gradient descent optimisation algorithms, refer to [86].

Bibliography

[1] AIRTUM and AIOM. *I numeri del cancro in Italia 2023*. Intermedia editore, 2023.

[2] Luisa Altabella et al. "Machine learning for multi-parametric breast MRI: radiomics-based approaches for lesion classification". In: *Physics in Medicine & Biology* (2022).

[3] Christophe Ambroise and Geoffrey J McLachlan. "Selection bias in gene extraction on the basis of microarray gene-expression data". In: *Proceedings of the national academy of sciences* 99.10 (2002), pp. 6562–6566.

[4] Ali Abbasian Ardakani et al. "Interpretation of radiomics features: a pictorial review". In: *Computer Methods and Programs in Biomedicine* (2021), p. 106609.

[5] Reza Azad et al. "Loss Functions in the Era of Semantic Segmentation: A Survey and Outlook". In: *arXiv preprint arXiv:2312.05391* (2023).

[6] M Bach Cuadra, Valérie Duay, and J-Ph Thiran. "Atlas-based segmentation". In: *Handbook of Biomedical Imaging: Methodologies and Clinical Research* (2015), pp. 221–244.

[7] Nicholas P Baskerville et al. "The loss surfaces of neural networks with general activation functions". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.6 (2021), p. 064001.

[8] Giulio Benetti. "Radiomica per la predizione della risposta alla chemioterapia nel tumore della mammella." PhD thesis. 2019.

[9] E Bercovich and MC Javitt. *Medical Imaging: From Roentgen to the Digital Revolution, and Beyond. Rambam Maimonides Medical Journal, 9, e0034*. 2018.

[10] NJ Bundred. "Prognostic and predictive factors in breast cancer". In: *Cancer treatment reviews* 27.3 (2001), pp. 137–142.

[11] Renee Cattell, Shenglan Chen, and Chuan Huang. "Robustness of radiomic features in magnetic resonance imaging: review and a phantom study". In: *Visual computing for industry, biomedicine, and art* 2 (2019), pp. 1–16.

[12] Gökçen Çetinel, Fuldem Mutlu, and Sevda Gül. "Decision support system for breast lesions via dynamic contrast enhanced magnetic resonance imaging". In: *Physical and Engineering Sciences in Medicine* 43 (2020), pp. 1029–1048.

[13] Yeun-Chung Chang et al. "Computerized breast lesions detection using kinetic and morphologic analysis for dynamic contrast-enhanced MRI". In: *Magnetic resonance imaging* 32.5 (2014), pp. 514–522.

[14] Dar-Ren Chen et al. "Multiview contouring for breast tumor on magnetic resonance imaging". In: *Journal of digital imaging* 32 (2019), pp. 713–727.

[15] Jeon-Hor Chen et al. "Quantitative analysis of peri-tumor fat in different molecular subtypes of breast cancer". In: *Magnetic resonance imaging* 53 (2018), pp. 34–39.

[16] Anna Choromanska, Yann LeCun, and Gérard Ben Arous. "Open problem: The landscape of the loss surfaces of multilayer networks". In: *Conference on Learning Theory*. PMLR. 2015, pp. 1756–1760.

[17] Anna Choromanska et al. "The loss surfaces of multilayer networks". In: *Artificial intelligence and statistics*. PMLR. 2015, pp. 192–204.

[18] Özgün Çiçek et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 424–432.

[19] Alan S Coates et al. "Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015". In: *Annals of oncology* 26.8 (2015), pp. 1533–1546.

[20] Mehmet Ufuk Dalmış et al. "A computer-aided diagnosis system for breast DCE-MRI at high spatiotemporal resolution". In: *Medical physics* 43.1 (2016), pp. 84–94.

[21] Aydin Demircioğlu. "The effect of preprocessing filters on predictive performance in radiomics". In: *European Radiology Experimental* 6.1 (2022), p. 40.

[22] Lee R Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), pp. 297–302.

[23] Basak Erguvan-Dogan et al. "Bi-RADS-MRI: a primer". In: *American Journal of Roentgenology* 187.2 (2006), W152–W160.

[24] Gokhan Ertas, Simon J Doran, and Martin O Leach. "A computerized volumetric segmentation method applicable to multi-centre MRI data to support computer-aided breast tissue analysis, density assessment and lesion localization". In: *Medical & biological engineering & computing* 55 (2017), pp. 57–68.

[25] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[26] Roberta Fusco et al. "Breast DCE-MRI: lesion classification using dynamic and morphological features by means of a multiple classifier system". In: *European radiology experimental* 1.1 (2017), pp. 1–7.

[27] Antonio Galli et al. "A pipelined tracer-aware approach for lesion segmentation in breast DCE-MRI". In: *Journal of imaging* 7.12 (2021), p. 276.

[28] Raoof Gholami and Nikoo Fakhari. "Support vector machine: principles, parameters, and applications". In: *Handbook of neural computation*. Elsevier, 2017, pp. 515–535.

[29] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. "Radiomics: images are more than pictures, they are data". In: *Radiology* 278.2 (2016), pp. 563–577.

[30] Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[32] Albert Gubern-Mérida et al. "Automated localization of breast cancer in DCE-MRI". In: *Medical image analysis* 20.1 (2015), pp. 265–274.

[33] Luying Gui, Chunming Li, and Xiaoping Yang. "Medical image segmentation based on level set and isoperimetric constraint". In: *Physica Medica* 42 (2017), pp. 162–173.

[34] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[35] Moritz Hardt, Ben Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent". In: *International conference on machine learning*. PMLR. 2016, pp. 1225–1234.

[36] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[37] Jianghua He, Prabhakar Chalise, et al. "Nested and repeated cross validation for classification model with high-dimensional data". In: *Revista Colombiana de Estadística* 43.1 (2020), pp. 103–125.

[38] Byung-Woo Hong. "Joint estimation of shape and deformation for the detection of lesions in dynamic contrast-enhanced breast MRI". In: *Physics in Medicine & Biology* 58.21 (2013), p. 7757.

[39] Liang Hu et al. "Image manifold revealing for breast lesion segmentation in DCE-MRI". In: *Bio-medical materials and engineering* 26.s1 (2015), S1353–S1360.

[40] Ignacio Alvarez Illan et al. "Automated detection and segmentation of nonmass-enhancing breast tumors with dynamic contrast-enhanced magnetic resonance imaging". In: *Contrast Media & Molecular Imaging* 2018 (2018).

[41] Paul Jaccard. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". In: *Bull Soc Vaudoise Sci Nat* 37 (1901), pp. 547–579.

[42] Shruti Jadon. "A survey of loss functions for semantic segmentation". In: *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE. 2020, pp. 1–7.

[43] Jagadaeesan Jayender et al. "Automatic segmentation of invasive breast carcinomas from dynamic contrast-enhanced MRI using time series analysis". In: *Journal of Magnetic Resonance Imaging* 40.2 (2014), pp. 467–475.

[44] Gaoxia Jiang and Wenjian Wang. "Error estimation based on variance analysis of k-fold cross-validation". In: *Pattern Recognition* 69 (2017), pp. 94–106.

[45] Rasha Kamal et al. "Contrast-enhanced mammography in comparison with dynamic contrast-enhanced MRI: which modality is appropriate for whom?" In: *Egyptian Journal of Radiology and Nuclear Medicine* 52 (2021), pp. 1–14.

[46] Ji-Hyun Kim. "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap". In: *Computational statistics & data analysis* 53.11 (2009), pp. 3735–3745.

[47] TK Koo and MY Li. *A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016; 15 (2): 155–63*. 2000.

[48] Subbiahpillai Neelakantapillai Kumar, A Lenin Fred, and P Sebastin Varghese. "Suspicious lesion segmentation on brain, mammograms and breast MR images using new optimized spatial feature based super-pixel fuzzy c-means clustering". In: *Journal of digital imaging* 32 (2019), pp. 322–335.

[49] Ruben THM Larue et al. "Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study". In: *Acta oncologica* 56.11 (2017), pp. 1544–1553.

[50] Erin LeDell, Maya Petersen, and Mark van der Laan. "Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates". In: *Electronic journal of statistics* 9.1 (2015), p. 1583.

[51] Shaoyuan Lei et al. "Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020". In: *Cancer Communications* 41.11 (2021), pp. 1183–1194.

[52] Jacob Levman et al. "Semi-automatic region-of-interest segmentation based computer-aided diagnosis of mass lesions from dynamic contrast-enhanced magnetic resonance imaging based breast cancer screening". In: *Journal of digital imaging* 27 (2014), pp. 670–678.

[53] Jundong Li et al. "Feature selection: A data perspective". In: *ACM computing surveys (CSUR)* 50.6 (2017), pp. 1–45.

[54] Wei Li et al. "Molecular subtypes recognition of breast cancer in dynamic contrast-enhanced breast magnetic resonance imaging phenotypes from radiomics data". In: *Computational and Mathematical Methods in Medicine* 2019 (2019).

[55] David Liljequist, Britt Elfving, and Kirsti Skavberg Roaldsen. "Intraclass correlation–A discussion and demonstration of basic features". In: *PloS one* 14.7 (2019), e0219854.

[56] Hai-Qing Liu et al. "Machine learning on MRI radiomic features: Identification of molecular subtype alteration in breast cancer after neoadjuvant therapy". In: *European Radiology* 33.4 (2023), pp. 2965–2974.

[57] Hui Liu et al. "A new background distribution-based active contour model for three-dimensional lesion segmentation in breast DCE-MRI". In: *Medical physics* 41.8Part1 (2014), p. 082303.

[58] Jun Ma et al. "Loss odyssey in medical image segmentation". In: *Medical Image Analysis* 71 (2021), p. 102035.

[59] Katarzyna J Macura et al. "Patterns of enhancement on breast MR images: interpretation and imaging pitfalls". In: *Radiographics* 26.6 (2006), pp. 1719–1734.

[60] Maria Adele Marino et al. "Multiparametric MRI of the breast: A review". In: *Journal of Magnetic Resonance Imaging* 47.2 (2018), pp. 301–315.

[61] Darryl McClymont et al. "Fully automatic lesion segmentation in breast MRI using mean-shift and graph-cuts on a region adjacency graph". In: *Journal of Magnetic Resonance Imaging* 39.4 (2014), pp. 795–804.

[62] Kenneth O McGraw and Seok P Wong. "Forming inferences about some intraclass correlation coefficients." In: *Psychological methods* 1.1 (1996), p. 30.

[63] Pankaj Mehta et al. "A high-bias, low-variance introduction to machine learning for physicists". In: *Physics reports* 810 (2019), pp. 1–124.

[64] Carmelo Militello et al. "3D DCE-MRI radiomic analysis for malignant lesion prediction in breast cancer patients". In: *Academic Radiology* 29.6 (2022), pp. 830–840.

[65] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571.

[66] Hang Min et al. "Automatic lesion detection, segmentation and characterization via 3D multi-scale morphological sifting in breast MRI". In: *Biomedical Physics & Engineering Express* 6.6 (2020), p. 065027.

[67] Stefania Montemezzi et al. "3T DCE-MRI radiomics improves predictive models of complete response to neoadjuvant chemotherapy in breast cancer". In: *Frontiers in oncology* 11 (2021), p. 630780.

[68] Monica Morrow, Janet Waters, and Elizabeth Morris. "MRI for breast cancer screening, diagnosis, and treatment". In: *The Lancet* 378.9805 (2011), pp. 1804–1811.

[69] Quynh Nguyen and Matthias Hein. "The loss surface of deep and wide neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 2603–2612.

[70] Nobuyuki Otsu. "A threshold selection method from gray-level histograms". In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66.

[71] Dinesh Pandey et al. "Automatic breast lesion segmentation in phase preserved DCE-MRIs". In: *Health Information Science and Systems* 10.1 (2022), p. 9.

[72] Nikolaos Papanikolaou, Celso Matos, and Dow Mu Koh. "How to develop a meaningful radiomic signature for clinical use in oncologic patients". In: *Cancer Imaging* 20 (2020), pp. 1–10.

[73] Doohyun Park et al. "Importance of CT image normalization in radiomics analysis: prediction of 3-year recurrence-free survival in non-small cell lung cancer". In: *European Radiology* 32.12 (2022), pp. 8716–8725.

[74] Chengtao Peng et al. "LMA-Net: A lesion morphology aware network for medical image segmentation towards breast tumors". In: *Computers in Biology and Medicine* 147 (2022), p. 105685.

[75] Hanchuan Peng, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on pattern analysis and machine intelligence* 27.8 (2005), pp. 1226–1238.

[76] Gabriele Piantadosi et al. "DCE-MRI breast lesions segmentation with a 3TP U-Net deep convolutional neural network". In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2019, pp. 628–633.

[77] Gabriele Piantadosi et al. "Multi-planar 3D breast segmentation in MRI via deep convolutional neural networks". In: *Artificial Intelligence in Medicine* 103 (2020), p. 101781.

[78] SE Pinder et al. "Laboratory handling and histology reporting of breast specimens from patients who have received neoadjuvant chemotherapy". In: *Histopathology* 50.4 (2007), pp. 409–417.

[79] Robert H Press et al. "The use of quantitative imaging in radiation oncology: a quantitative imaging network (QIN) perspective". In: *International Journal of Radiation Oncology* Biology* Physics* 102.4 (2018), pp. 1219–1235.

[80] ChuanBo Qin et al. "Joint Dense Residual and Recurrent Attention Network for DCE-MRI Breast Tumor Segmentation". In: *Computational Intelligence and Neuroscience* 2022 (2022).

[81] Masoomeh Rahimpour et al. "Visual ensemble selection of deep convolutional neural networks for 3D segmentation of breast tumors on dynamic contrast enhanced MRI". In: *European Radiology* 33.2 (2023), pp. 959–969.

[82] Meredith H Redden and George M Fuhrman. "Neoadjuvant chemotherapy in the treatment of breast cancer". In: *Surgical Clinics* 93.2 (2013), pp. 493–499.

[83] Jorge S Reis-Filho and Lajos Pusztai. "Gene expression profiling in breast cancer: classification, prognostication, and prediction". In: *The Lancet* 378.9805 (2011), pp. 1812–1823.

[84] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[85] Jintao Ru et al. "Attention guided neural ODE network for breast tumor segmentation in medical images". In: *Computers in Biology and Medicine* 159 (2023), p. 106884.

[86] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (2016).

[87] David Rydning, John Reinsel, and John Gantz. "The digitization of the world from edge to core". In: *Framingham: International Data Corporation* 16 (2018), pp. 1–28.

[88] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. "Tversky loss function for image segmentation using 3D fully convolutional deep networks". In: *International workshop on machine learning in medical imaging*. Springer. 2017, pp. 379–387.

[89] Debbie Saslow et al. "American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography". In: *CA: a cancer journal for clinicians* 57.2 (2007), pp. 75–89.

[90] Tapas Si et al. "Identification of breast lesion through integrated study of gorilla troops optimization and rotation-based learning from MRI images". In: *Scientific Reports* 13.1 (2023), p. 11577.

[91] Kok Swee Sim et al. "Breast cancer detection from MR images through an auto-probing discrete Fourier transform system". In: *Computers in biology and medicine* 49 (2014), pp. 46–59.

[92] Pawel Smialowski, Dmitrij Frishman, and Stefan Kramer. "Pitfalls of supervised feature selection". In: *Bioinformatics* 26.3 (2010), pp. 440–443.

[93] Thorvald Sørensen. "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons". In: *Biologiske skrifter* 5 (1948), pp. 1–34.

[94] Jost Tobias Springenberg et al. "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806* (2014).

[95] Karl D Spuhler et al. "Task-based assessment of a convolutional neural network for segmenting breast lesions for radiomic analysis". In: *Magnetic resonance in medicine* 82.2 (2019), pp. 786–795.

[96] Carole H Sudre et al. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer. 2017, pp. 240–248.

[97] Liang Sun et al. "A collaborative multi-task learning method for BI-RADS category 4 breast lesion segmentation and classification of MRI images". In: *Computer Methods and Programs in Biomedicine* 240 (2023), p. 107705.

[98] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[99] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC medical imaging* 15.1 (2015), pp. 1–28.

[100] J Terven et al. "Loss Functions and Metrics in Deep Learning". In: *arXiv preprint arXiv:2307.02694* (2023).

[101] Jie Tian et al. *Radiomics and Its Clinical Application: Artificial Intelligence and Medical Big Data*. Academic Press, 2021.

[102] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[103] Michal R Tomaszewski and Robert J Gillies. "The biological meaning of radiomic features". In: *Radiology* 298.3 (2021), pp. 505–516.

[104] Janita E Van Timmeren et al. "Radiomics in medical imaging—"how-to" guide and critical reflection". In: *Insights into imaging* 11.1 (2020), pp. 1–16.

[105] Erik Verburg et al. "Knowledge-based and deep learning-based automated chest wall segmentation in magnetic resonance images of extremely dense breasts". In: *Medical physics* 46.10 (2019), pp. 4405–4416.

[106] Joel Vidal, Joan C Vilanova, Robert Martí, et al. "A U-Net Ensemble for breast lesion segmentation in DCE MRI". In: *Computers in Biology and Medicine* 140 (2022), p. 105093.

[107] Wolf-Dieter Vogl et al. "Automatic segmentation and classification of breast lesions through identification of informative multiparametric PET/MRI features". In: *European radiology experimental* 3 (2019), pp. 1–13.

[108] Chuin-Mu Wang, Chieh-Ling Huang, and Sheng-Chih Yang. "3D shape-weighted level set method for breast MRI 3D tumor segmentation". In: *Journal of Healthcare Engineering* 2018 (2018).

[109] Weimiao Wu et al. "Exploratory study to identify radiomics classifiers for lung cancer histology". In: *Frontiers in oncology* 6 (2016), p. 71.

[110] Chang Yan Xu, Zi Jiang Sang, and Ye Qin Shao. "MSA-VNet: Multi-scale Attention-based V-Net for DCE-MRI Lesion Segmentation". In: *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. IEEE. 2022, pp. 309–312.

[111] Cindy Xue et al. "Radiomics feature reliability assessed by intraclass correlation coefficient: a systematic review". In: *Quantitative Imaging in Medicine and Surgery* 11.10 (2021), p. 4431.

[112] Hao Hao Yang and Fang Fang Zhang. "Magnetic Resonance Imaging Features in Diagnosis of Breast Cancer and Evaluation of Effect of Epidermal Growth Factor Receptor-Targeted Therapy". In: *BioMed Research International* 2022 (2022).

[113] Yi Yang et al. " 2, 1-norm regularized discriminative feature selection for unsupervised learning". In: *IJCAI international joint conference on artificial intelligence*. 2011.

[114] Milita Zaheed et al. "Sequencing of anthracyclines and taxanes in neoadjuvant and adjuvant therapy for early breast cancer". In: *Cochrane Database of Systematic Reviews* 2 (2019).

[115] Jing Zhang et al. "Fully automatic classification of breast lesions on multi-parameter MRI using a radiomics model with minimal number of stable, interpretable features". In: *La radiologia medica* 128.2 (2023), pp. 160–170.

[116] Wenchao Zhang, Yu Guo, and Qiyu Jin. "Radiomics and Its Feature Selection: A Review". In: *Symmetry* 15.10 (2023), p. 1834.

[117] Yang Zhang et al. "Automatic detection and segmentation of breast cancer on MRI using mask R-CNN trained on non–fat-sat images and tested on fat-sat images". In: *Academic Radiology* 29 (2022), S135–S144.

[118] Xiaoming Zhao et al. "BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification". In: *Computers in Biology and Medicine* 164 (2023), p. 107255.

[119] Heng Zhou et al. "Multitask Deep Learning-Based Whole-Process System for Automatic Diagnosis of Breast Lesions and Axillary Lymph Node Metastasis Discrimination from Dynamic Contrast-Enhanced-MRI: A Multicenter Study". In: *Journal of Magnetic Resonance Imaging* (2023).

[120] Pan Zhou et al. "Towards theoretically understanding why sgd generalizes better than adam in deep learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21285–21296.

[121] Jingjin Zhu et al. "Development and validation of a deep learning model for breast lesion segmentation and characterization in multiparametric MRI". In: *Frontiers in oncology* 12 (2022), p. 946580.

[122] Alex Zwanenburg et al. "Assessing robustness of radiomic features by image perturbation". In: *Scientific reports* 9.1 (2019), p. 614.

[123] Alex Zwanenburg et al. "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping". In: *Radiology* 295.2 (2020), pp. 328–338.

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor and co-supervisor, for their invaluable guidance and support throughout the entire research process. I am also deeply grateful to the AOVR staff for their support and resources, which have made this project possible.

Finally, I extend heartfelt thanks to my dearest friend G. and my parents for their unwavering support and encouragement throughout my academic journey.