Definition of predictive models to assess the response to neadjuvant chemotherapy from breast magnetic resonance images

Marie-Judith Saint Martin

université
PARIS-SACLAY

Definition of predictive models to assess the response to neoadjuvant chemotherapy from breast magnetic resonance images

Modèles de prédiction de la réponse à la chimiothérapie néoadjuvante à partir d'examens d'IRM mammaire

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n°575 Electrical, Optical, Bio: PHYSICS AND ENGINEERING (EOBE)
Spécialité de doctorat: Sciences de l'information et de la communication
Graduate School : Sciences de l'ingénierie et des systèmes
Référent : Faculté des sciences d'Orsay

Thèse préparée dans le Laboratoire d'Imagerie Translationnelle en Oncologie (LITO) (U1288, Inserm/Institut Curie) sous la direction de Frédérique FROUIN, chargée de recherche, et le co-encadrement de Fanny ORLHAC, chargée de recherche

Thèse soutenue à Orsay, le 18 novembre 2022, par

Marie-Judith SAINT MARTIN

Composition du jury

Cyril Poupon Directeur de recherche, Neurospin - Institut des Sciences du Vivant Frédéric Joliot - CEA	Président du jury
Carole Lartizien Directrice de recherche, CREATIS UMR 5520 - CNRS - Université Lyon 1	Examinatrice
Benjamin Lemasson Chargé de recherche, GIN U836 - Inserm - Université Grenoble Alpes	Rapporteur & Examinateur
Nicolas Passat Professeur des universités, CReSTIC - Université de Reims Champagne-Ardenne	Rapporteur & Examinateur
Frédérique Frouin Chargée de recherche, LITO, U1288 - Inserm - Institut Curie	Directrice de thèse

THESE DE DOCTORAT
NNT : 2022UPAST129

Acknowledgements

A lot happened during the course of my thesis at Institut Curie. We went from working full time at the laboratory to working from home due to COVID-19, endured multiple lockdowns and discovered the pros and cons of online meetings and conferences. Nevertheless, it was an enriching and rewarding experience from both the professional and human points of views thanks to the many people that I was lucky to meet at LITO.

First, I am extremely grateful to Frédérique Frouin for her trust, constant support, patience and involvement throughout these three years. Frédérique allowed me to spread my wings and try many approaches and ideas while preventing me from being sidetracked.

I would also like to extend my deepest gratitude to Fanny Orlhac for her encouragement and unwavering guidance. Whenever I was feeling down or getting worried, she was always there to cheer me up (sometimes with a tea break) and reaffirm her profound belief in my abilities.

Frédérique and Fanny, that both oversaw me during this thesis, created an enjoyable and proactive environment to work and discuss. They were extremely involved in my work, sometimes even at the wee hours of the morning. I was incredibly lucky to have such supportive advisors.

I cannot begin to express my thanks to Dr. Carole Malhaire, who initiated this whole study. Her help was instrumental in the analysis of MR images and development of models. I would like to thank her for her kindness and patience in introducing me to the physiology and mechanisms of breast cancer. Thanks also to the members of the radiology department of Institut Curie for their help and support for the practical experiments.

I am also grateful to Irène Buvat for her critical review and advice and for the vibrancy and dedication that she brings to the LITO and the PhD students in particular.

Special thanks to Christophe Nioche for his help with LIFEx and computer set-ups in general but also for his inimitable pep talks.

I also had the great pleasure of working with all the members of the LITO that made the lab a fun and pleasant environment for research.

I would finally like to thank Nicolas Passat and Benjamin Lemasson, for examining my thesis and for their valuable suggestions and recommendations. I would also like to thank Carole Lartizien and Cyril Poupon for agreeing to be on my thesis committee and for their interest in my work.

At last, I want to thank my family and friends: my mother for introducing me from an early age into the world of research, my brother for showing me what a PhD in a scientific lab look like and finally Pauline, Nolwenn and Clément for their love and continuous support.

Synthèse en Français

Ce travail a été mené au Laboratoire d'Imagerie Translationnelle en Oncologie (LITO) de l'Institut Curie sous la supervision de Frédérique Frouin (directrice de thèse) et de Fanny Orlhac, en collaboration avec le département de radiologie de l'Institut Curie et plus particulièrement avec le Dr Caroline Malhaire, qui a recruté la cohorte de patientes et initié l'étude, le Dr Pia Akl et le Dr Fatine Selhane. L'approche de segmentation automatique présentée au chapitre 7 a été réalisée en collaboration avec Michel Koole et Masoomeh Rahimpour de l'équipe "Nuclear Medicine & Molecular Imaging" de l'Université KU Leuven (Belgique).

Le cancer du sein est le cancer le plus fréquent chez les femmes en France. Il est devenu le cancer le plus diagnostiqué à l'échelle mondiale en 2020 avec près de 2,3 millions de cas recensés [1]. Il s'agit d'une maladie très hétérogène, divisée en quatre sous-types moléculaires (Luminal A, Luminal B, HER2+, Triple Négatif) qui ont des caractéristiques, des pronostics et des réponses aux traitements différentes [2]. Administrée avant la chirurgie, la chimiothérapie néoadjuvante (CNA) vise à diminuer la taille des tumeurs pour faciliter l'intervention chirurgicale et réduire le recours aux mastectomies [3, 4]. La CNA est devenue le traitement de référence pour les cancers agressifs ou localement avancés. Elle est principalement prescrite pour les cancers de type Luminal B, HER2+ ou Triple Négatif. Cependant, le taux de réponse pathologique complète à la CNA dépend des sous-types moléculaires et est globalement de 20 à 30% [5, 6]. Avec le développement de la médecine personnalisée, l'intérêt pour la prédiction précoce de la réponse pathologique complète à la CNA s'est développé. Identifier les patientes qui ne répondraient pas à la chimiothérapie avant le début du traitement constituerait une avancée majeure dans leur prise en charge. Les patientes identifiées comme mauvaises répondeuses pourraient être orientées plus rapidement vers d'autres thérapies ce qui leur éviterait les effets nocifs de la CNA sans retarder leur prise en charge thérapeutique.

Notre hypothèse est que l'imagerie médicale donne accès à des informations complémentaires à celles fournies par la biopsie, qui vont contribuer à prédire la réponse pathologique complète d'une tumeur avant ou en cours de traitement. L'imagerie *in vivo* permet d'évaluer de manière non-invasive, à différents moments du traitement, la tumeur dans son ensemble ainsi que son micro-environnement. L'intelligence artificielle a augmenté le potentiel de prédiction de l'imagerie médicale, en considérant les images comme une masse importante de données. Ainsi, la radiomique, une discipline récente, repose sur l'hypothèse que la morphologie et l'hétérogénéité d'une tumeur, mesurées macroscopiquement, traduiraient ses caractéristiques biologiques. En extrayant des indices de forme, des indices d'intensité et des indices de texture, il serait donc possible de quantifier des informations difficilement appréciables visuellement et de développer de nouveaux modèles de prédiction [7, 8].

L'IRM est la modalité d'imagerie médicale la plus usitée pour le suivi de la réponse à la chimiothérapie néoadjuvante dans le cancer du sein [9, 10]. Mener des analyses quan-

titatives en IRM soulève cependant plusieurs problèmes. Les images IRM sont sujettes à un biais du champ magnétique, modifiant la distribution d'intensités de voxels similaires en fonction de leur position dans le champ de vue, et à l'arbitraire des unités dans lesquelles sont exprimées les intensités. Ces deux phénomènes rendent difficiles les comparaisons entre différentes acquisitions effectuées sur un même appareil avec un paramétrage identique. Par ailleurs, certains indices radiomiques sont dépendants des paramètres d'acquisition (type d'imageur, antenne, séquences...), effet que nous résumons sous le terme d'« effet scanner» [11]. En conséquence, un modèle prédictif défini à partir d'indices issus d'images acquises dans un premier centre d'imagerie risque d'avoir des performances dégradées en utilisant des images d'un second centre d'imagerie.

Après avoir brièvement défini le contexte clinique de ce travail dans le chapitre 1, le chapitre 2 de cette thèse fait un état de l'art de la radiomique utilisant des indices prédéfinis (indices de forme, issus de l'histogramme et/ou de texture) extraits des images, sous la forme d'un schéma récapitulant les 8 étapes nécessaires pour mener de telles études (Figure 1).

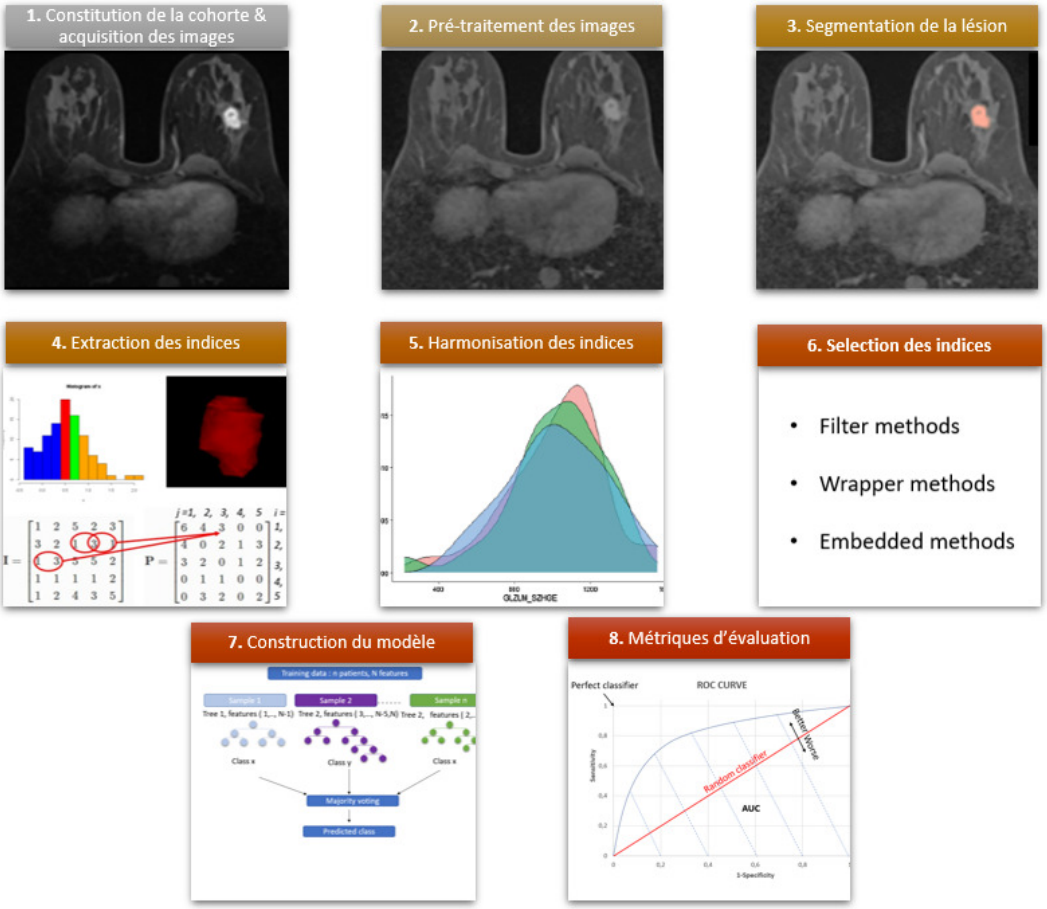


Figure 1: Schéma récapitulatif des étapes d'une analyse radiomique.

Le chapitre propose ensuite une analyse critique de l'état de l'art des études radiomiques utilisant l'IRM et s'intéressant à la prédiction précoce de la réponse à la CNA dans le cancer du sein. De cette étude sont retenus 36 articles utilisant des indices prédéfinis et 11 études utilisant

des méthodes d'apprentissage profond. Plusieurs enseignements se dégagent de l'analyse. Tout d'abord, la prédiction de la réponse à partir des examens IRM est une question difficile et d'une étude à l'autre, les conclusions varient sur l'intérêt des indices radiomiques extraits des images acquises avant le début de la chimiothérapie pour prédire la réponse [12, 13], sur l'utilité de combiner indices radiomiques et données cliniques et biologiques [12, 14], sur la nature des séquences d'IRM à privilégier [12, 15] et sur la nécessité de développer des modèles restreints à un sous-type moléculaire particulier [16]. Les questions de correction du biais dans le champ magnétique, de normalisation des images et d'harmonisation des indices radiomiques pour réduire l'effet scanner sont également peu discutées : moins de 15% des études mentionnent ces sujets. Enfin, l'évaluation des modèles développés sur des bases de données de test indépendantes et multicentriques reste rare (<14%). Pourtant ce point est crucial pour mesurer la robustesse et la généralisation des modèles proposés.

L'objectif du travail de thèse est de prédire la réponse à la CNA à partir d'une base multicentrique d'IRM mammaires acquises avant le début de la prise en charge thérapeutique, tout en apportant des éléments de réponse aux problèmes de standardisation des données d'imagerie et de généralisation des modèles radiomiques. Le travail a été réalisé sur des images acquises classiquement suivant les protocoles cliniques, à savoir des images pondérées en T1 après injection de produit de contraste et des images pondérées en T2. Une cohorte de 136 patientes, ayant toutes été prises en charge pour leur thérapie à l'Institut Curie entre 2016 et 2020, a été constituée. Les patientes présentaient une tumeur à un stade avancé ou localement agressive, nécessitant une chimiothérapie néoadjuvante. La base de données a été divisée en un ensemble d'apprentissage de 103 patientes dont les images ont été acquises sur l'une des trois machines de l'Institut Curie et d'un ensemble de test indépendant de 33 patientes dont les images ont été acquises en majorité à l'extérieur de l'Institut Curie, préalablement à la prise en charge, dans une quinzaine de centres d'imagerie différents. A partir des images IRM, les tumeurs ont été segmentées séparément par deux radiologues à l'exception de 30 lésions de l'ensemble d'apprentissage qui ont été segmentées par les deux radiologues afin d'étudier la reproductibilité inter-opérateur.

Le chapitre 3 décrit les conditions d'inclusion des patientes et d'acquisition des images IRM ainsi que les données cliniques et biologiques. Les radiologues ont également évalué visuellement cette base de données en utilisant le lexique standardisé BI-RADS (*Breast Imaging-Reporting And Data System*) [17], spécifique de l'imagerie mammaire. De premiers modèles prédictifs utilisant seulement les caractéristiques BI-RADS et les données cliniques et biologiques ont été construits en utilisant différentes méthodes de sélection des données (RFE pour « *recursive feature elimination* » ou « élimination récursive de caractéristiques », algorithme de Boruta, méthode mRMR pour « *minimum Redundancy Maximum Relevance* » ou « redondance minimale pertinence maximale») et types de modèles (SVM pour « *Support Vector Machine* » ou « Séparateur à Vaste Marge », *Random Forest* ou forêt aléatoire, et régression logistique). Les performances en terme « AUC » d'aire sous la courbe ROC (« *Receiver operating characteristic* » ou « fonction d'efficacité du récepteur ») obtenues sur l'ensemble de test se situent dans l'intervalle [0,65, 0,76] et sont équivalentes à celles publiées par d'autres équipes [18].

Le chapitre 4 présente un ensemble de méthodes (appelé pipeline) de correction des images

IRM développé et testé sur deux fantômes anatomiques de sein et dont les résultats ont été publiés dans *Magnetic Resonance Materials in Physics, Biology and Medicine* en 2021 [19]. Les fantômes sont composés de matériaux ayant des propriétés élastiques proches de celles du sein. Ils ont des incrustations de dureté et taille différentes, simulant des nodules et sont dédiés à l'entrainement de la réalisation des biopsies mammaires par les radiologues. Les images de fantôme, plus simples à analyser que les images cliniques, ont ainsi été acquises sur les deux machines de l'Institut Curie et les 3 antennes mammaires utilisées en clinique, suivant un protocole d'imagerie proche de celui utilisé pour les patientes, à l'exception de l'injection d'un produit de contraste.

De nombreux travaux portant sur la correction du biais lié au champ magnétique dans des acquisition d'images cérébrales ont été publiés et plusieurs algorithmes de correction, dont l'algorithme N4 de Tustison et al. [20], ont été proposés. La correction de cet effet dans l'IRM du sein demeure cependant sous-étudiée. Appliquer l'algorithme N4 utilisant les hyper paramètres définis pour le cerveau sur les images de sein n'a pas permis une correction satisfaisante des effets. Nous avons montré que l'algorithme devait utiliser un masque du sein pour estimer le champ de biais et 5 niveaux de décomposition du volume d'images (au lieu de 4) pour le corriger efficacement. Deux types de normalisation, un z-score classique et une méthode de concordance d'histogrammes [21, 22], ont ensuite été testées sur les images de fantômes. Corriger le champ de biais et normaliser les images quelle que soit la méthode a permis de réduire les variations intra et inter-acquisitions mais n'a pas conduit à gommer totalement l'effet scanner sur les valeurs des indices radiomiques. Une harmonisation spécifique de ces indices par l'approche ComBat [23] est en effet nécessaire après les étapes de correction de biais et de normalisation. La méthode ComBat consiste à aligner les distributions des indices issus d'images acquises avec des protocoles différents selon le principe suivant : pour un indice y mesuré dans la région j du centre d'imagerie ou du protocole i, l'indice peut être défini par :

$$y_{ij} = \alpha + \gamma_i + \delta_i \epsilon_{ij} \tag{1}$$

où α est la valeur moyenne de y, γ_i est un effet centre additif et $\delta_i \epsilon_{ij}$ un effet centre multiplicatif associé à un terme d'erreur. La méthode ComBat corrige les distributions des paramètres en calculant $\hat{\alpha}$, $\hat{\gamma_i}$ et $\hat{\delta_i}$ estimateurs de α, γ_i et δ_i suivant l'estimation du maximum de vraisemblance, tel que :

$$y_{ijcorrected} = \frac{y_{ij} - \hat{\alpha} - \hat{\gamma_i}}{\hat{\delta_i}} + \hat{\alpha} \tag{2}$$

La forme non-paramétrique de la méthode a été employée sans l'hypothèse empirique de Bayes. Une transformation spécifique pour chaque indice séparément a été définie et a permis de réduire considérablement l'effet scanner.

Le chapitre 5 a pour objectif d'adapter le pipeline défini au chapitre précédent sur les images de fantômes aux images patientes et propose une déclinaison du schéma global d'analyse en 8 étapes présenté au chapitre 2 :

- Constitution de la cohorte et acquisition des images
- Prétraitement des images :
  - Correction du champ de biais magnétique grâce à l'algorithme de N4 paramétré spécifiquement pour l'IRM mammaire

  - Rééchantillonnage spatial des images pondérées T1 après injection de produit de contraste et des images pondérées T2.
  - Normalisation des images en utilisant un z-score dont les paramètres (moyenne, écart-type) ont été calculés dans le parenchyme mammaire, à l'exclusion de la tumeur

- Segmentation de la lésion tumorale

- Extraction des indices des images natives et des images filtrées par des filtres en ondelettes à l'aide du logiciel Pyradiomics en utilisant une discrétisation absolue des images et un intervalle de taille fixe

- Harmonisation des indices radiomiques en utilisant la méthode ComBat

- Sélection des indices :

  - Standardisation des indices (variables centrées réduites)
  - Sélection des indices robustes à la méthode de segmentation en se basant sur le coefficient de corrélation intra-classe de chaque indice (ICC $>$ 0,8) calculé sur les 30 lésions segmentées par les deux radiologues
  - Sélection des indices dont la borne inférieure de l'AUC pour prédire la réponse à la CNA est strictement supérieure à 0,5
  - Rejet des indices présentant un coefficient de corrélation de Spearman supérieur au seuil de 0,8
  - Sélection des 5 indices les plus fréquemment sélectionnés au cours de 100 répétitions de l'algorithme de Boruta, pour construire les modèles prédictifs

- Construction des modèles prédictifs

- Métriques d'évaluation

Pour valider la méthode de correction du biais dans les volumes d'images, le coefficient de variation dans un tissu de référence (couche adipeuse dans les images pondérées T1 et quelques coupes du sternum dans les images pondérées T2) a été calculé. La diminution de ce coefficient après la correction sur les images des ensembles d'apprentissage et de test dans les deux modalités a mis en évidence la réduction des inhomogénéités d'intensité dans le champ de vue. Le choix entre les deux méthodes de normalisation (z-score et concordance d'histogrammes) s'est fait selon plusieurs critères : la qualité de l'alignement des distributions d'intensités dans les images entre les différentes acquisitions, le nombre d'indices après normalisation concordants avec les indices avant normalisation en se basant sur le coefficient de corrélation de concordance et le nombre d'indices associés à la réponse à la CNA [24]. Il n'y avait pas de supériorité d'une des deux méthodes dans l'alignement des distributions mais la normalisation par z-score menait à un plus grand nombre d'indices concordants et associés à la CNA. Compte tenu également de la plus grande facilité de mise en œuvre, l'approche z-score a été finalement retenue.

Le chapitre 6 intègre la succession d'analyses définie dans le chapitre précédent pour développer des modèles radiomiques afin de prédire la réponse à la CNA dont les premiers résultats ont été présentés à la conférence IEEE-EMBS et publiés [25]. Dans ce chapitre, nous nous sommes intéressés à comprendre si d'autres informations que celles classiquement estimées dans la région tumorale pouvaient contribuer à la prédiction de la CNA. Différents volumes d'intérêt desquels sont extraits les indices radiomiques ont été définis sur les images: la région tumorale classique, une boîte parallélépipédique englobant la région tumorale, une boîte de taille fixe située dans la plus grande majorité des cas exclusivement à l'intérieur de la tumeur et la boîte englobante précédemment définie mais en considérant uniquement deux niveaux d'intensité distinguant la tumeur de l'extérieur. Pour déterminer l'intérêt de combiner les indices extraits de plusieurs volumes, 15 expériences ont été menées correspondant à toutes les combinations possibles utilisant les indices issus d'un, de deux, de trois ou de quatre volumes d'intérêt. Ces expériences ont été répétées en utilisant dans un premier temps les indices issus des images pondérées T1 après injection de contraste, puis des images pondérées T2 et enfin les deux types de séquences. Les résultats ont montré l'intérêt d'associer les paramètres radiomiques issus de la région tumorale classique, de la tumeur « binarisée » placée dans une boîte englobante et d'une boîte de taille fixe située à l'intérieur de la tumeur avec des performances significativement supérieures à celles obtenues dans les expériences utilisant seulement la région tumorale classique. Sur la base de test indépendante et multicentrique, l'index de Youden médian était égal à 0,44 et l'écart interquartile était de [0,43, 0,50] en prenant la meilleure expérience, alors qu'il était égal à 0,16 et l'écart interquartile à [0,10, 0,18] pour l'expérience utilisant seulement la région tumorale. Considérer uniquement la forme de la tumeur en la plaçant dans une boîte englobante pour calculer des indices de texture a permis de caractériser sa forme plus précisément que ce que ne peuvent faire les indices de forme classiquement utilisés dans les logiciels d'analyse radiomique. Des modèles spécifiques sur le sous-groupe de patientes présentant des tumeurs HER2 positives et des tumeurs triple négatives (73 patientes en apprentissage et 24 patientes en test) ont ensuite été développés. Les résultats préliminaires obtenus soutiennent l'hypothèse que des modèles dédiés à certains sous-types moléculaires peuvent améliorer les performances de prédiction de réponse complète à la CNA. Des analyses complémentaires devront cependant être menées sur de plus grandes cohortes.

Pour obtenir ces résulats dans le chapitre 6, l'ensemble de test, comportant 33 études, a été analysé de la même manière que l'ensemble d'apprentissage à la différence que l'harmonisation des indices n'a pas pu être réalisée par la méthode ComBat. En effet, cette approche requiert qu'il y ait entre 20 et 30 images par configuration d'acquisition alors que l'ensemble de test contient entre un et cinq examens par centre pour les images acquises en dehors de l'Institut Curie. Une solution originale a alors été développée pour harmoniser les caractéristiques radiomiques extraites de l'ensemble de test. En s'appuyant sur les indices radiomiques calculés dans une région de taille constante placée dans le sein controlatéral de chaque patiente de la base d'apprentissage, et après analyse par composantes principales, trois groupes correspondant aux trois configurations matérielles utilisées pour l'acquisition des images peuvent être mis en évidence. Les centroïdes de ces « clusters » ont été calculés, puis chaque examen de l'ensemble de test a été associé au cluster dont il était le plus proche, au sens de la distance euclidienne. Chaque examen de l'ensemble de test est ainsi assigné à une des machines de l'ensemble d'apprentissage. Les transformations spécifiques à chaque indice pour les

machines de l'ensemble d'apprentissage précédemment déterminées ont pu alors être utilisées pour harmoniser l'ensemble de test. Dans la tâche spécifique de la prédiction de la réponse à la CNA, de meilleures performances ont été observées dans 73% des expériences réalisées après harmonisation et des performances meilleures ou équivalentes dans 82% des cas après harmonisation suivant la méthode proposée.

Le dernier chapitre propose une approche originale, basée sur l'apprentissage profond, de segmentation des tumeurs mammaires à partir d'images pondérées T1 après injection de produit de contraste. Cette approche est décrite sous la forme d'un article publié dans *European Radiology* [26]. Segmenter les tumeurs est une tâche fastidieuse pour les radiologues, qui n'est pas nécessaire dans le soin courant, ce qui ralentit ainsi la constitution de cohortes dédiées aux études radiomiques. Les indices radiomiques sont également affectés par la variabilité inter-opérateur de segmentation. Idéalement une segmentation largement automatisée des tumeurs permettrait de réduire la variabilité inter-opérateur tout en facilitant l'augmentation de la taille des cohortes pour les études radiomiques. Ce chapitre portant sur la segmentation utilise une cohorte de patientes légèrement différente de celle utilisée précédemment car elle inclut quelques images acquises au cours du traitement. Pour cette étude, la distinction entre l'ensemble d'apprentissage et l'ensemble de test s'est faite de façon à ce que l'ensemble de test soit composé uniquement des 30 lésions segmentées par les deux radiologues et que les performances obtenues puissent être analysées au regard de la variabilité inter-opérateur. Dans cette approche, trois modèles ont été définis, utilisant soit des images après injection, soit une fusion précoce ou tardive des images après injection du produit de contraste et des images de soustraction (images acquises après injection du contraste et images acquises juste avant l'injection). Comme la fusion des trois modèles ne donnait pas de résultats satisfaisants, il a été demandé à l'expert de choisir le meilleur modèle parmi les trois proposés pour chaque patient. Les performances des différents modèles et de la méthode d'ensemble ont été évaluées quantitativement en utilisant le score de Dice et la distance de Hausdorff et visuellement par un radiologue. La méthode d'ensemble a obtenu sur la base de test des performances équivalentes à l'accord entre les deux radiologues. Dans 77% des cas, les segmentations choisies par la méthode d'ensemble ont été qualifiées d'« excellentes » ou d' « utiles » par le radiologue alors que ce pourcentage n'était que de 60% pour le meilleur des trois modèles automatiques. Ainsi, la contribution de l'expert a permis d'améliorer sensiblement les performances, sans augmenter de façon notoire sa charge de travail (relecture de trois contours pour une même étude au lieu d'un seul contour).

En conclusion, ce travail de thèse a apporté des contributions méthodologiques dans l'élaboration d'une chaîne complète d'analyse pour développer des modèles radiomiques, fondés sur des images par résonance magnétique, robustes à la segmentation et exportables sur des acquisitions faites dans des centres extérieurs. Des tendances dans la recherche d'informations pertinentes pour prédire la réponse à la chimiothérapie néoadjuvante ont été dégagées. Les perspectives visent à conforter ces résultats à partir de cohortes plus conséquentes, tout en explorant d'autres axes de recherche comme l'exploitation des images acquises au cours du traitement mais également des images de diffusion pour prédire la réponse à la chimiothérapie néoadjuvante dans le cancer du sein le plus précocement possible.

Contents

Introduction

Breast cancer is the leading cause of cancer death in women worldwide. A highly heterogeneous disease, breast cancer has been classified into four main molecular subtypes with different characteristics, treatments, and prognoses. For aggressive and locally advanced tumors, neoadjuvant chemotherapy (NAC), that administers treatments before surgery, has become the standard of care. NAC aims to facilitate surgeries and decrease mastectomy rates. Chances of responding to NAC are however extremely variable depending on molecular subtypes and globally only 20 to 30% of patients achieve pathological complete response (pCR). In the context of precision medicine, early prediction of pCR is becoming more important. Being able to identify non-responders beforehand would indeed greatly improve patient care as they could be offered specific alternative treatments more quickly and not suffer the side-effects of intense chemotherapy.

Trying to predict the response to treatment requires to be able to characterize the lesion accurately. Biopsies are used to determine the molecular and genetic profile of the tumor, but they only characterize a small sample of a large and frequently heterogeneous lesion. Medical imaging appears as an essential complement to biopsies as they offer a non-invasive, easily repeatable way of assessing lesions in their entirety and potentially their micro-environment. With the rise of artificial intelligence, the potential of medical imaging grew even more as images are considered as a mineable source of a considerable amount of data that could be used, for instance, in predictive modelling. The emerging field of radiomics is premised upon the idea that morphological aspects and heterogeneity of tumors convey information about their biological properties. By extracting shape, intensity-based or texture features, radiomics can contribute to access this hidden or not easily quantifiable biological information in the images and use it to build decision-making tools.

Magnetic Resonance Imaging (MRI), thanks to its high contrast resolution in soft tissues, is one of the most precise modalities to monitor patient response to neoadjuvant chemotherapy in breast cancer and assess residual disease. However, performing quantitative analysis on MRI raises several challenges. Among them, the bias field gain, creating local inhomogeneities within tissues, and the comparison of MR intensities between different images since the MR information is not conventionally expressed in standard units, can be mentioned. Furthermore, radiomic features extracted from medical imaging are heavily influenced by acquisition parameters (scanners, sequences...), this effect being called the "scanner effect". This makes the use of radiomic models built using images from one imaging center on datasets from other centers a very sensitive step.

In recent years, many MRI-based radiomic studies investigated the prediction of pCR to NAC in breast cancer, sometimes limiting the models to a specific molecular subtype. These models have various performances, but these numerous studies prove that this prediction is a

difficult task to solve. Little interest was shown in addressing normalization issues in breast MRI and very few studies reported an export of the radiomic models on multicentric test data.

The focus of this thesis is thus to improve the prediction of pCR to NAC compared with the models found in the literature, with a particular focus on the standardization of images and radiomic features and the exportability of radiomic models.

This work was conducted at the "Laboratoire d'Imagerie Translationnelle en Oncologie" (LITO) under the supervision of Frédérique Frouin and Fanny Orlhac, in collaboration with the Radiology Department of Institut Curie and especially with Dr. Caroline Malhaire, who defined the cohort, Dr. Pia Akl and Dr. Fatine Selhane. The segmentation approach detailed in this work (Chapter 7) was developed in collaboration with Michel Koole and Masoomeh Rahimpour from the "Nuclear Medicine & Molecular Imaging" team of KU Leuven University (Belgium).

This thesis consists of seven chapters.

Chapter 1 introduces breast cancer and protocols used in breast MR imaging. It provides some epidemiological data related to breast cancer, its risk factors and highlights the heterogeneity of the disease by going over the different classifications used in clinical routine. Advantages and drawbacks of neoadjuvant chemotherapy are introduced as well as the role of MRI in monitoring patient response to NAC.

Chapter 2 focuses on radiomics for medical imaging, describing the general pipeline of a handcrafted radiomic study from the image acquisition to the evaluation of radiomic models. A critical review of the literature on the prediction of pCR to NAC in breast cancer using MRI-based radiomics is also conducted.

Chapter 3 introduces the patient cohort. A cohort of 136 patients with locally advanced or aggressive breast cancers, all treated at Institut Curie using standard of care therapy was collected. At Institut Curie, three MR settings were used to acquire T1-weighted dynamic contrast-enhanced (DCE) images and fat-saturated T2-weighted images. In addition, MR images coming from several other imaging centers were used if their quality was deemed satisfactory. Clinical and biological data that were collected on top of MR images are described. Finally, predictive models, using only clinical and biological data, to predict pCR to NAC were built and their performances compared with those obtained by equivalent models from the literature.

Chapter 4 describes experiments conducted on imaging breast phantoms, scanned at Institut Curie with the routine clinical protocol that was used for the patient cohort. These experiments were used to develop a correction pipeline to address inhomogeneities issues on MR images and the impact of the "scanner effect" on radiomic features, with the hope to apply it in a later stage on patient data. This chapter is mainly based on a paper published in *Magnetic Resonance Materials in Physics, Biology and Medicine* [19].

Chapter 5 describes the choices and experiments that were carried out to export the pipeline proposed in Chapter 4 to patient images. It goes over the radiomic pipeline presented in

Chapter 2 and explains, step by step, the methodological choices taken in our study to build the radiomic models, including the harmonization of features with the statistical ComBat method to reduce the "scanner effect". A dedicated approach to correct bias field inhomogeneities and to normalize images is detailed.

Chapter 6 aims to further define the relevant information that can be found in MR images to predict pCR to NAC, with the pipeline presented in Chapter 5. Using different types of volume of interest based on tumor delineation and simple boxes inside the tumor or englobing the tumor, the influence of shape and margins of the lesions on the prediction are studied. Different models were built on the training database (103 patients imaged at Institut Curie) and tested on the test set of 33 patients (imaged in 15 centers). As it was not possible to apply the ComBat procedure on the test set due to the reduced number of acquisitions in the different centers (between 1 and 5), an original harmonization strategy to correct the "scanner effect" where conventional harmonization methods cannot be used due to small sample size, is also detailed. This chapter is mainly based on the draft of a paper to be submitted. Some preliminary results are shown using the same methodology but applied to a subgroup of patients with specific molecular subtypes.

Chapter 7 presents a deep learning-based segmentation approach to segment breast tumors on T1-weighted DCE images. Requiring a final user validation, it allows to choose from three methods the best segmentation for a given patient. This work was done in collaboration with KU Leuven university (Belgium). The chapter is mainly based on a paper published in *European Radiology* [26].

Conclusions and plans for future work, including new methodological developments, are finally exposed.

Chapter 1

Breast cancer

Preface

This chapter focuses on the clinical context starting from the epidemiology of breast cancer and its risk factors, then going over the classifications used in clinical routine, highlighting the heterogeneity of this disease. It presents next the advantages and drawbacks of neoadjuvant chemotherapy, the varying rates of pathological complete response among molecular subtypes and the predictive biomarkers associated with this response. The final section describes the role of MRI in breast cancer and especially in the assessment of response to neoadjuvant chemotherapy. It outlines MRI clinical acquisition protocols and introduces the Breast Imaging Reporting & Data System (BI-RADS).

1.1 Epidemiology

In 2020, female breast cancer became the most frequently diagnosed cancer worldwide, with almost 2.3 millions cases detected, representing 11.7% of the total new cancer cases, as stated by GLOBOCAN 2020 data [1]. Breast cancer age-standardized incidence rate has been proven to be positively associated with the human development index (HDI) of countries by multiple studies [27–29]. Transitioned countries with high HDI, displaying a higher prevalence of breast cancer risk factors such as obesity and carrying intensive screening, indeed report incidence rates 88% higher than transitioning countries with low/medium HDI. Highest incidence rates are observed in Northern and Western Europe, Northern America and Australia and New Zealand (>80 per 100 000) [1, 30].

Breast cancer accounted for 6.9% of worldwide cancer deaths, that being an estimated 685 000 deaths, behind lung, colorectal, liver and stomach cancers affecting men and women alike. It is nevertheless the leading cancer cause of deaths in women in 110 countries, representing almost 1 in 6 cancer deaths globally (Figure 1.1). Despite lower incidence rate than in transitioned countries, women in transitioning countries have a 17% higher mortality rate (15.0 versus 12.8 per 100,000) with highest rates found in countries in Oceania, Polynesia, Western Africa, and the Caribbean [1].

Incidence rates of breast cancer have globally increased these last two decades. This trend can be explained by the impressive uptake in mammographic screening, the aging of the population in western countries, the worldwide increase of obesity but also by sociocultural

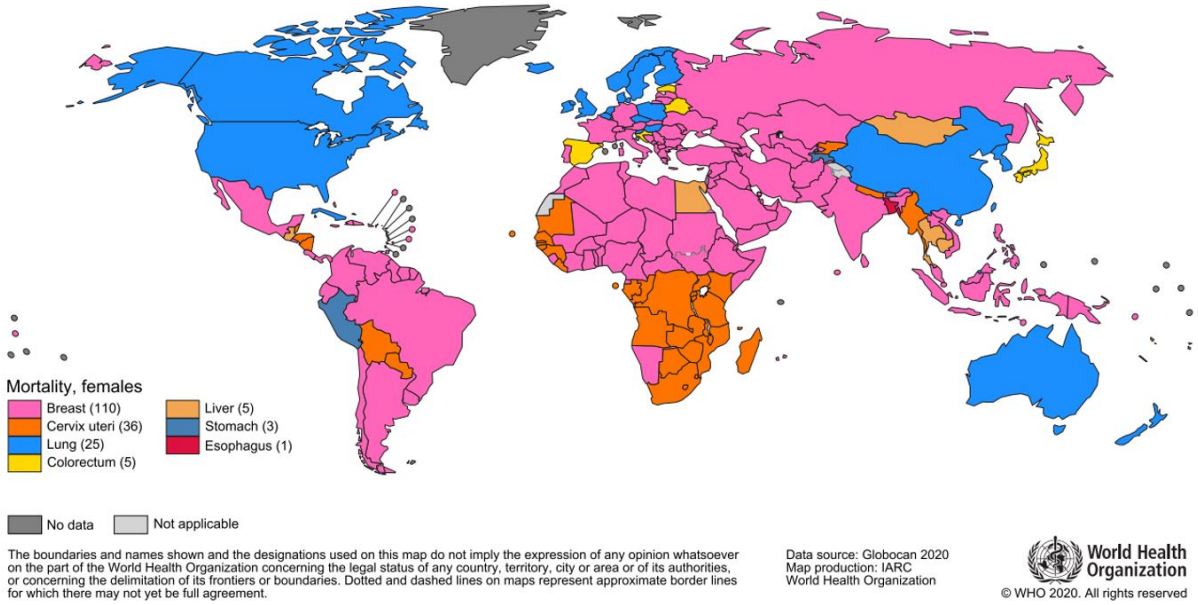


Figure 1.1: Most common type of cancer mortality by country in 2020 among women. The numbers of countries represented in each ranking group are included in the legend. Source: GLOBOCAN 2020.

changes like the delayed age of first pregnancy. In some countries like China or Korea, it is associated with the evolution to a more "westernized" lifestyle presenting higher risk factors. Conversely, most developed countries reported a decrease in mortality rates between 2000 and 2015 due to early detection of breast cancer and the development of advanced treatments while mortality rate increased or remained stable in low income countries with limited access to repeated preventive screening or timely treatments [1, 28, 30, 31].

1.2 Risk factors

Breast cancer risk factors are wide-ranging and extremely diverse. While some of these risk factors are inherent, others can be linked to patient lifestyle and environment.

Non-modifiable risk factors include for example, age, as more than 80% of breast cancer patients are over 50 years old, ethnicity, with a higher incidence observed in white non Hispanic women, and high breast tissue density. Family history of breast or ovarian cancers and to an even greater extent, personal history of previous breast cancers or radiation therapies are also associated with a higher risk of breast cancer. Genetic mutations on genes such as BRCA1 and BRCA2, which are correlated with a rise of carcinogenesis, have a considerable impact on the probabilities of developing breast cancer [30, 32]. Besides, studies have highlighted the link between exposure to endogenous hormones, such as oestrogen, progesterone or other sex hormones, and incidence of breast cancer. A high concentration of oestrogen in postmenopausal women is associated with an increased incidence of breast cancer. Events that affect hormonal balance and especially reproductive habits have therefore been investigated. Early age of first menstruation, delayed age of first full-term pregnancy and late menopause are linked to an increase risk of breast cancer while breastfeeding, especially over a long period, reduces it [30,

33].

Lifestyle and physical behaviours can also influence the occurrence and development of breast cancer. Dietary habits are amongst the first extrinsic factors associated with higher risk of breast cancer. The consumption of highly processed food rich in saturated fat and sodium increases this risk [34]. The lack of physical activity, excessive weight or obesity, elevated alcohol consumption and active smoking are also associated with it [30, 35].

1.3 Breast cancer classification

1.3.1 Histological classification

Breast cancer is a highly heterogeneous disease, presenting a multitude of different clinical, morphological and molecular profiles affecting its behaviour, response to treatments and eventual outcome [36].

Breast cancer is considered invasive as opposed to *in situ* when cancer cells have broken from the milk ducts or lobules (glands producing the milk) where they usually originate and have started proliferating into the surrounding stroma (Figure 1.2). Histological classification of invasive breast cancer nevertheless mainly rests on the number, cell types characteristics and profiles rather than the location where these cells first appeared to predict breast cancer types [37]. The two main histological types of invasive breast cancer are invasive ductal carcinoma and invasive lobular carcinoma. Invasive ductal carcinoma is the default type, also called the “no specific type" that gathers tumors which do not present sufficient characteristics to be classified as a special type of breast cancer. Fifty to 80% of new breast cancer cases are classified as invasive ductal carcinoma. Invasive lobular carcinoma on the other hand represents 5 to 15% of new cancer cases and is more common in women of advanced age.

1.3.2 Classification based on grade and stage

The American Joint Comittee on Cancer (AJCC) defined in its eighth edition an updated staging system of breast cancer to estimate patients' prognosis [30, 38]. This system, which was globally recognized, combines anatomical assessment and biological factors, such as the evaluation of biomarkers expression or histological grading.

To assess tumor growth and spread, clinicians rely on the anatomical TNM staging, which evaluates tumor size and extension (T), nodal status (N) and distant metastases (M) using multiple categories. A global stage varying between 0 to IV is defined, inversely associated to an estimate prognosis [38–40] (Table 1.1).

Histological grading can be established using various methods but the Scarff-Bloom Richardson grading system modified by Elston-Ellis, often called the Nottingham grading system, remains a reference (Table 1.2) [41]. This grading aimed to transcribe tumor biology by analyzing the extent of differentiation in tumor tissue [42]. Tumor grade is assessed by evaluating three types of morphological features: the mitotic count, the formation of glands or tubules, and the size and shape of cell nuclei. A total score is then calculated by summing up scores of each morphological feature. A score between 3 and 5 is associated to grade 1, 6-7 to grade 2 and 8-9 to grade 3. Low-grade tumors which show clear differentiation of structures in tissue, have been found to have a better prognosis than high-grade tumors. Besides, works have

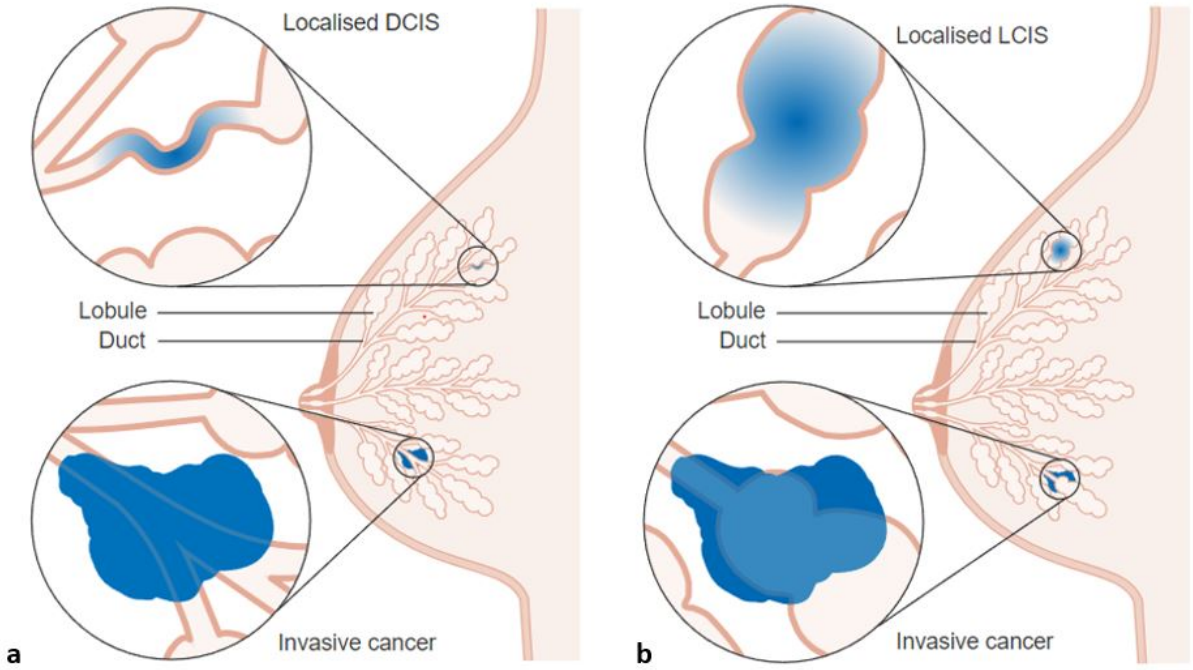


*Figure 1.2: Diagrams showing **(a)** progression from a ductal carcinoma in situ (DCIS) to an invasive ductal carcinoma. **(b)** progression from a lobular carcinoma in situ (LCIS) to an invasive lobular carcinoma. Source: Cancer Research UK/ Wikimedia Commons.*

showed that histological grade remained a prognostic factor independently from tumor size and lymph node status [43].

The AJCC updated staging system also recommended that all invasive carcinoma be tested for estrogen receptor (ER), progesterone receptor (PR) and Human epidermal growth factor receptor 2 (HER2). A tumor is said to be ER or PR positive when at least 1% of the cells collected for immunohistochemistry testing display respectively ER or PR receptors. Around three out of four diagnosed breast tumors are hormone positive (ER or PR) [45]. It is however worth mentioning that the 1% threshold is not always adopted as in France, a 10%-threshold is preferred.

Finally, the AJCC paved the way to use more genetic profiling during the staging process to further specify tumors and offer a personalized estimate of patients' outcome.

1.3.3 Molecular classification

Going beyond morphological and histological classification, attempts have been made to gather tumors according to molecular patterns. This stratification of breast cancer in molecular subtypes aimed to better apprehend its diversity and develop specific therapies for each major trend of breast cancer. Analysing the gene expressions of tumors using the cDNA microarray technique, fundamental differences between five different molecular subtypes were established [46, 47]. However, due to the high cost and technology difficulty of resorting to these techniques in clinical practice, a surrogate classification based on immunohistochemistry (IHC) was developed and recommended at the Saint Gallen conference in 2011 [2]. This IHC classification

		Stage	Primary tumour (T)*	Regional lymph node status (L)	Distant metastasis (M)
T- Tumour		**0**	Tis	N0	M0
T1	Tumour ≤ 2 cm	**I**	T1	N0	M0
T2	Tumour ≥ 2 cm but ≥ 5 cm		T0	N1	M0
T3	Tumour ≥ 5 cm	**IIA**	T1	N1	M0
T4	Tumour of any size with direct extension to chest wall or skin		T2	N0	M0
N- Lymph node		**IIB**	T2	N1	M0
N0	No cancer in regional node		T3	N0	M0
N1	Regional movable metastasis	**III A**	T0	N2	M0
N2	Non-movable regional metastases		T1	N2	M0
N3	Cancer in the internal mammary lymph nodes		T2	N2	M0
M- Metastasis			T3	N1/N2	M0
M0	No distant metastases	**III B**	T4	Any N	M0
M1	Distant metastases	**III C**	Any T	N3	M0
		IV	Any T	Any N	M1

Criteria for staging breast tumours according to the UICC ICD-10 TNM classification.

*Size measurements are for the tumour's greatest dimension.

Table 1.1: TNM classification for breast cancer. Source: Ljuslinder [40]

was founded on the expression of four prognostic biomarkers: the estrogen receptor, progesterone receptor, Human epidermal growth factor receptor 2 and cell proliferation regulator (Ki67). Four molecular subtypes were thus defined [2, 37, 45, 48, 49]:

- Luminal A tumors represent almost 50% of invasive breast cancer cases. They are characterised by being ER positive, HER2-negative and presenting high expression of PR ($\geq$ 20%) and low Ki67 levels ($<$ 14%). Luminal A tumors usually have a good prognosis and are historically treated using hormone therapies.
- Luminal B tumors gather 20% to 30% of new invasive breast cancers. They can be divided into two groups: one characterised by being ER positive, HER2-negative and displaying low expression of PR ($\leq$ 20%), and high level of Ki67 ($\geq$ 14%) and the other characterized by being ER positive and HER2 positive, and having various levels of expression of PR and Ki67. Luminal B tumors tend to have a poorer prognosis than Luminal A tumors and are treated using hormone therapies as an alternative or alongside chemotherapies.

Feature graded	Criterion	Score
Tubule (gland) formation	>75%	1
	10%-75%	2
	<10%	3
Nuclear pleomorphism	Small, regular uniform cells	1
	Moderate increase and variability	2
	Marked variation	3
Mitotic count (/10 hpf)	0-5	1
	6-10	2
	>11	3

Table 1.2: Nottingham grading system for invasive breast cancers. Source: Atanda et al. [44].

- HER2-enriched tumors (HER2 positive, ER negative, PR $\leq$ 20%, high Ki67 levels) account for 15 to 20% of new breast cancer cases. Assessment of HER2 status may need complementary testing using Fluorescence in situ hybridization (FISH) technique to confirm the gene amplification when membrane staining in immunohistochemical tests is incomplete or moderate. Though HER2 positive tumors were originally associated with a poor outcome, the development of HER2-target therapies used as adjuvant therapies to chemotherapies considerably improved patients' prognosis [50]. These therapies are predominantly based on a monoclonal antibody called trastuzumab (commercially sold under the name of Herceptin).

- Triple-negative (TN) tumors represent 10% to 20% of breast cancer cases. They display an ER, PR and HER2 negative status but high level of Ki67 expression ($\geq$ 14%), reflecting their important ability to proliferate. They are very aggressive tumors, have a poorer prognosis and higher risks of recurrence than other subtypes especially within the first five years of treatment. TN tumors usually affect more young women below the age of forty and patients with BRCA1 mutations. TN subtype presents a high heterogeneity and can be itself divided into multiple subtypes [51]. Due to the lack of hormone or HER2 receptors, TN tumors do not respond to endocrine therapies or anti-HER2 therapies and their heterogeneity makes it difficult to develop other target therapies. The main systemic treatment of TN tumors thus consists in chemotherapies (adjuvant or neoadjuvant) combined with surgeries [52].

Despite the stratification of breast cancers defined at the Saint Gallen conference, no clear cut-off of the expression of Ki67 to separate Luminal A from Luminal B tumors could be found. Though a cutoff at 14% was first suggested [2], several studies have recently advocated to use a 20% threshold [53, 54].

A noteworthy difference between breast cancer subtypes is their association to histological grade. Luminal A tumors were found to be well differentiated (Grade 1) while other subtypes were associated with higher grades (Grade 2 or 3) [37, 55, 56].

Table 1.3 summarizes differences observed between molecular subtypes.

Molecular Subtypes	Luminal A	Luminal B		HER2+	TN
		(HER2-)	(HER2+)		
Biomarkers	ER+ PR+ HER2- Ki67low	ER+ PR- HER2- Ki67high	ER+ PR-/+ HER2+ Ki67low/high	ER- PR- HER2+ Ki67high	ER- PR- HER2- Ki67high
Frequency of Cases (%)	40–50	20–30		15–20	10–20
Histological Grade	Well Differentiated (Grade I)	Moderately Differentiated (Grade II)		Little Differentiated (Grade III)	Little Differentiated (Grade III)
Prognosis	Good	Intermediate		Poor	Poor
Response to Therapies	Endocrine	Endocrine Chemotherapy	Endocrine Chemotherapy Target Therapy	Target Therapy Chemotherapy	Chemotherapy PARP Inhibitors

ER: estrogen receptor; PR: progesterone receptor; HER2: human epidermal growth factor receptor 2.

Table 1.3: Molecular classification of breast cancers associated to their prognosis and therapies. Source: Gomes do Nascimento et al. [37]

1.4 Neoadjuvant chemotherapy in breast cancer

1.4.1 Purpose and benefits

Neoadjuvant chemotherapy (NAC) is a systemic treatment based on cytotoxic drugs delivered to patients before the administration of local treatments [3].

NAC for breast cancer was first proposed in the beginning of the 80's to patients with inoperable tumors with the aim to reduce lesions to facilitate ensuing surgeries or radiation therapies. It was later extended first to operable breast cancers requiring mastectomies in the hope to downstage tumors so that breast-conserving surgeries could be performed instead and then to other early-stage or operable tumors to prevent post-surgical complications [4]. In some rare cases ($< 5\%$), progression of the disease during NAC could nevertheless have an adverse effect on breast conservation [57].

Clinical trials showed that there was no difference in overall survival between patients administered NAC or adjuvant chemotherapy (chemotherapy given after surgery) though the breast conservation rates in NAC patients was higher [58, 59]. Questions have arisen about a possible link between NAC and higher locoregional recurrence rate but firm evidence remains lacking. Observed increase could be also explained by the lower rate of mastectomies [3, 57].

On top of favouring breast conservation, NAC is also associated with fewer adverse effects (complications due to chemotherapy in particular). Besides, administering chemotherapy before surgery allows to monitor closely and *in vivo* the effects of the treatment on the tumor. Patients resistant to drugs can be identified and their treatment altered accordingly [57]. Pathological complete response (pCR) to NAC is also associated with better overall survival (OS) and event-free survival (EFS) [60]. NAC treatments thus offered the additional opportunity for the research field to assess the prognosis power of new biomarkers using pCR as a surrogate marker for OS and EFS.

1.4.2 Pathological complete response

There is a global lack of standardization in the definition of pathological complete response to NAC and methods to assess it. Several classification systems have been developed to report on post-neoadjuvant specimens with the main ones including the Chevalier method, the Sataloff method, the Miller-Payne system, the residual cancer burden (RCB) score or the ypTNM staging [61]. Some of these classifications evaluate changes in cellularity and sizes between pre and post-NAC tumor specimens (Miller-Payne and Sataloff methods) while other systems focus only on the residual tumor present in the breast and axillary nodes (RCB, ypTNM) [62].

Pathological complete response can be at first described as the lack of residual invasive disease in the breast. This definition is however often extended to include axillary lymph nodes as a better prognosis value is conferred to pathological complete response that takes into account both breast parenchyma and nodes. Patients with no invasive residuals in breast and nodes have better overall survival [63–65]. Nevertheless, the prognosis power of achieving pCR is intrinsically linked to the molecular subtype of tumors. Indeed, pCR has been proven to be a good prognosis factor in triple-negative and HER2-enriched tumors but its entailment in other subtypes must be further explored [6]. Based on the analysis of the Collaborative Trials in Neoadjuvant Breast Cancer (CTNeoBC), the Food and Drug Administration (FDA), the federal agency that regulates the market access of drugs in the United States, has thus approved the use of pCR to NAC in high risk agressive breast cancer (triple-negative and HER2-enriched particularly) as an endpoint to accelerate approval of new drugs though confirmatory trials should be later conducted [66].

1.4.3 pCR rates among molecular subtypes

The probability of achieving pathological complete response strongly depends on molecular subtypes. Results found in the literature must be analysed carefully as to take into account the definition of pCR selected in the study, the different chemotherapy regimens administered to patients, the number of chemotherapy cycles followed, the potential use of complementary targeted or endocrine therapies and the general tumor state at the beginning of NAC. Using the immunohistochemistry classification, studies suggested that:

- Luminal A tumors have very low pCR rates and the lowest rate among all molecular subtypes. Haque et al. [5] analysed pCR to NAC in 13 929 women among which 322 patients were diagnosed with Luminal A tumors. Of these, only 0.3% achieved pCR. Other works in the literature reported rates between 0% to 7.5% as summarized by Wang-Lopez et al. [6]. Thus, treating Luminal A tumors with NAC is a controversial topic as neoadjuvant endocrine therapies or direct surgery is often preferred [67, 68].

- Luminal B tumors display higher chemosensitivity than Luminal A tumors with reported pCR rates between 1% and 16% [5]. Most chemotherapy treatments prescribed for Luminal A and B tumors consisted of a sequence or combination of several types of cytotoxic drugs, anthracyclines and taxanes being the most common, in a various number of cycles between 4 and 8. However, Luminal B tumors with HER2+ status can be treated with NAC combined with anti-HER2 targeted therapies (trastuzumab mainly) as HER2-enriched tumors, which can lead to higher response rate [22%-48%] [5].

- HER2-enriched tumors seem to have the highest rate of pCR among other subtypes [33% -70%]. Most studies evaluated pCR rates using chemotherapy combined with trastuzumab and reported improved pCR rates than when using chemotherapy alone. Research into using dual inhibition with two antibodies (trastuzumab + pertuzumab or trastuzumab + lapatinib) combined with chemotherapy yielded promising results, leading to higher pCR rates than chemotherapy with single inhibition, or dual inhibition without chemotherapy [6]. In all these combinations of treatments, HER2-enriched tumors achieved higher pCR rates than Luminal B/HER2+, attesting a higher sensitivity to chemotherapy and anti-HER2 therapies.

- Triple-negative tumors have relatively high pCR rates [20% -34%] assessed using treatments based in majority on an anthracyclines/taxanes regimen [5, 6]. However, recent studies have highlighted the potential benefit, in triple-negative tumors, of combining NAC with platinum salts (cisplatin and carboplatin), that are DNA-damaging drugs. TN tumors sensitivity to platinum agents could be linked to the high proportion of BRCA-mutated tumors among them. BRCA1 & BRCA2 genes are indeed involved in the DNA double strand repair process, and BRCA-mutated tumors would thus be more sensitive to cross-linking agents [69]. Simultaneously, there has been a huge increase in the development of targeted therapies for TN tumors [52]. Multiple agents like bevacizumab, an antibody inhibiting endothelial growth of blood vessels, have been investigated. Bevacizumab was found to improve pCR rates but its influence on OS has not been established [6].

1.4.4 Predictive biomarkers of pCR

Overall, neoadjuvant chemotherapy is prescribed to patients diagnosed with Luminal B, HER2-enriched or triple-negative tumors. Though Luminal B/HER2- tumors seldom achieved pCR, some benefits can still be gained from following NAC without pCR as studies still report in these conditions, tumor downstaging and higher breast conservation rates [68]. However, even with favorable molecular subtypes, a large proportion of patients does not respond to NAC. Being able to identify beforehand patients resistant to chemotherapy would considerably improve patient care. It would reduce patients' exposure to chemotherapy toxicity and enable them to be offered alternative treatments more quickly. A swift change of treatment is important as ineffective NAC gives time to the tumor to progress and metastasize and could favor ensuing chemo-resistance of tumors.

Predictive biomarkers not yet integrated into clinical practice were thus researched. High Ki67 proliferation index, ER negative status, high histological grade, small tumor sizes were significantly linked with pCR [70–72]. The potential interest of tumor-infiltrating lymphocytes (TILs) was also pointed out in some subtypes. TILs can invade either the stroma surrounding the lesion or the tumor-epithelial cells. High level of TILs in both occasions has been associated with pCR in HER2-positive and triple-negative tumors [4, 73, 74]. High level of stroma TILs alone, whose count was found to be more reproducible than intratumoral TILs, was validated as a strong predictor of both pCR and EFS in HER2-positive tumors [75] with several TILs level cut-off proposed (30%, 40%, 50%) [76]. Numerous studies attempted to integrate the previously mentioned biomarkers, sometimes with covariates like age or menopause status, to build nomograms predictive of pCR [70–72]. Further large-scale and independent validation

process must however take place before such predictive tools could be used in practice to help clinicians make a decision in the treatment pipeline.

Triple-negative tumors are highly heterogeneous. Novel genetic and molecular profiling have led to distinguish six subgroups within the TN subtype. These subgroups respond very differently to NAC with pCR rates going from 52% to 10% depending on the subgroup [4]. In the future, the constant development of genetic profiling could allow to select more precisely tumors that would likely achieve pCR.

1.5 MRI in breast cancer

1.5.1 MRI in breast cancer treatment pipeline

Magnetic resonance imaging (MRI) is a non-invasive imaging technique using high magnetic field, magnetic field gradients and radio waves to create three-dimensional anatomical and functional images. Because of its excellent soft tissue contrast resolution, MRI can be used to screen, diagnose and monitor patients during and after treatments for a great variety of conditions and regions (brain, thoracic and abdominal organs, pelvic organs, blood vessels, lymph nodes...).

In breast cancer screening routine, X-ray mammography, often associated with ultrasound, is the standard imaging modality. The sensitivity of mammography is however much lower in women with dense breast, as tumors can be easily masked within the fibrous and glandular tissue as they have the same appearance in mammography. X-ray performance is also reduced in young women and carriers of BRCA mutations. 3D MRI is not influenced by breast density and has a higher sensitivity than mammography [77]. Therefore, in 2007, the American Cancer Society recommended in its guidelines that women with high risk to develop breast cancer ($\geq$ 20%) due to family history and carriers of BRCA mutations and their untested relatives be offered MRI screening alongside mammography. MRI screening should also be proposed to women with an history of chest radiations between the age of 10 to 30 or affected with certain syndromes. No firm evidence could be found to advise for or against the additional use of MRI screening in women with dense breast [78].

There has been a significant uptake in the use of pre-operative MRI in newly diagnosed breast cancer in the last two decades. Arnaout et al. [79] reported a sharp increase from 3% to 24% of new breast cancer patients imaged with MRI between 2003 and 2012 in a retrospective study based on the Ontario cancer registry in Canada. However, the use of pre-operative MRI is still the subject of an intense debate as it is a costly modality, requiring patients to be screened in often uncomfortable positions for a relatively long amount of time and most importantly, has not been linked with improved outcomes [9]. Because of its higher sensitivity, MRI is able to detect more tumor foci. In a systematic review and subsequent meta-analysis, Houssami et al. [80] indeed reported a median increase of 16% in detecting lesions over 19 studies. However, due to the difficulty of separating benign from malignant lesions in MRI, only two-thirds of newly detected lesions were confirmed malignant after biopsies. Detecting more lesions led quite often to delay surgeries in order to carry on histological testing but also affected surgical procedures. Women with confirmed multifocal or multicentric tumors were recommended more extensive surgeries (especially mastectomies) in more than 11% of cases. False positive detection unfortunately also happened in about 5.5% of cases and

despite widespread use of histological confirmation, some women were still offered unnecessary extensive surgeries [81]. To date, the benefits of additional detection and altered surgeries are not clearly established. Several randomized trials reported no significant differences in the rate of re-operations (margin re-excision or conversion to mastectomies) in patients assessed with conventional imaging (mammography or/and ultrasound) or MRI. On the long-term, tumor staging with pre-operative MRI was not associated with a reduction of local or distant recurrence of breast cancer [82]. Nonetheless, in patients specifically diagnosed with infiltrating lobular carcinoma, MRI was found to better assess the extent of the tumor. MRI can also help in the detection of contralateral breast cancer though false positive findings are high [9, 83]. Based upon this mixed findings, guidelines of different national and international institutions may recommend pre-operative MRI for some subgroups of patients but most consider the use of MRI as optional.

1.5.2 MRI in neoadjuvant chemotherapy

The only context in which benefits of pre-operative MRI in breast cancer are clearly demonstrated is in the monitoring of the response to neoadjuvant chemotherapy. Though pathological complete response is definitively established by analyzing the surgical specimens, estimations with imaging can be used as surrogate markers. MRI has proven to be the most accurate method before ultrasound, mammography and clinical examination to estimate residual tumor size and determine pCR after NAC. Studies report high rates of correct identification of residual tumor (83%-92%) but intermediate rate of pCR estimations (47%-63%) [9]. Correct assessment of residual disease is important for surgical planning to select the type of breast surgery to perform and increase probabilities of achieving negative margins, implying that all tumor cells have been removed, and thus lower re-excision rate.

MRI has the combined benefit of allowing morphological and physiological monitoring of the response. Diameters or volume changes between the different cycles of chemotherapy can be measured accurately as recommended by Response Evaluation Criteria in Solid Tumors (RECIST) guidelines commonly used in oncology. In specific MRI sequences using contrast media injection, pharmacokinetic parameters allow to track the perfusion of contrast agent highlighting physiological changes occurring within the tumors. MR imaging also assesses well the shrinkage pattern of tumors whether it is concentric or scattered possibly leaving the tumor fragmented which is an important information in the selection of subsequent breast surgery [10].

MRI accuracy can also be influenced by tumor morphology, as the residual extent of non-mass lesions can be particularly underestimated, and by tumor molecular subtypes. For instance, accuracy to predict pCR based upon MR imaging captured at the end of NAC was higher in HER2-enriched and triple-negative tumors [10]. Aggressive tumors with a high Ki67 proliferation index or a high grade are also measured more accurately than other types of tumors [84, 85].

1.5.3 MRI clinical protocols for NAC

Women are usually imaged in MR scanners in the prone position with their breast positioned in dedicated coils equipped with multiple channels. Modern coils tend to have a higher number

of channels (16 or more) to increase signal-to-noise ratio and reinforce parallel imaging to reduce acquisition time. As biopsies are routinely performed to inspect newly found masses on MRI, patients can be imaged with coils dedicated to biopsy that allow an easy access to the breast area. Images are most often acquired in the axial plane that gives a complete overview of both breasts [9]. MRI routine clinical protocols involve almost always nowadays multiple MRI techniques based on several sequence types.

MRI clinical protocols rely mainly on the use of T1-weighted dynamic contrast-enhanced (T1-DCE) MRI. In this type of sequence, a gadolinium-based contrast agent is injected via intravenous access into patients at a dose of 0.1-0.2 mmol/kg [86]. The subsequent temporal enhancement of the breast is analysed as the increased concentration of the contrast agent at a given point will shorten the local T1 time, leading to a higher signal intensity. Dynamic series include a native pre-contrast image acquired before the administration of the gadolinium chelate and at least one post-contrast image taken 60 to 90s after injection corresponding to the common peak enhancement time in breast cancer. Subtraction of the pre and post-contrast image is commonly used in clinical routine for lesion detection, using maximum intensity projection (MIP) images [9]. Several other post-contrast images can be acquired usually till 5 to 7 minutes after injection.

This succession of images is used to create a time-signal intensity curve that assesses microvascular properties of the tissue such as blood vessel permeability and tissue perfusion [87]. In order to grow, tumors need to create an additional blood supply system by forming new blood vessels. This process is called angiogenesis. New vessels formed due to tumor growth can be more permeable leading to a quicker accumulation of contrast agent in tissue than in the normal vasculature. The reported differences in wash-in and wash-out of the gadolinium-based agent between benign and malignant lesions are reflected in their time-signal intensity curves and are further used in the diagnostic process: in 83% of benign cases, a slow onset with progressive enhancement can be observed (type I curve in Figure 1.3) while a strong and quick enhancement followed either by a plateau (type II in Figure 1.3) or a fast wash-out (type III in Figure 1.3) are reported in 93% of cases in malignant lesions [86]. A quick wash-out is the most common pattern observed in malignant lesions with the exception of DCIS and more diffuse lobular cancers that can show a persistent curve [9].

Several parameters that characterize the pharmacokinetics of the contrast agent in tissues can be defined, including K_{Trans} the volume transfer constant of the contrast agent from blood plasma to extravascular extracellular space (EES), V_e the fractional volume of the EES (volume of EES per unit volume of tissue) and K_{ep}, the rate constant of tranfer from the EES to blood plasma.

Routine protocols also include T2-weighting imaging with and without fat suppression. Fat suppression allows a better visualization of cysts [9]. Compared to T1-weighted imaging, T2-weighted imaging has the additional benefit of showing edemas in the breast that can have, like peritumoral and prepectoral edemas, diagnostic and prognostic values [88, 89].

Diffusion-weighted MRI (DW-MRI) is also part of the standard MRI protocol. This imaging technique measures the Brownian motion (random motion) of water molecules within tissue. The underlying principle of DW-MRI relies on T2* signal attenuation, depending on the ease of diffusion of water molecules. DW imaging (DWI) usually uses T2-weighted spin-echo (SE) or spin-echo echo planar pulse sequences. By applying two symmetrical gradients, one before and one after the 180° pulse, the sequence is made sensitive to diffusion. For stationary water

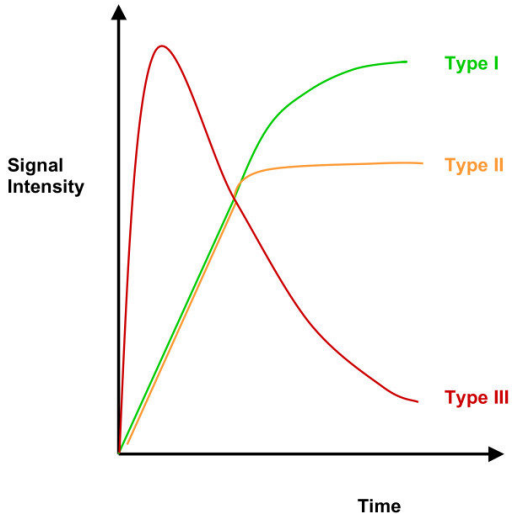


Figure 1.3: Diagrams depicting the three types (type I, type II and type III) of time-signal intensity curves usually observed in breast lesions imaged with dynamic contrast-enhanced MRI. Source: O'Flynn et al. [86]

molecules, the effect of the first gradient is reversed by the second one and there is no signal loss. Similarly, moving molecules gained phase information with the first gradient. However, since they move, they are not subjected to the exact same gradient the second time. They are thus not rephased properly and there is therefore an attenuation of the signal along the axis to which the gradient is applied. The greater the diffusion, the bigger the phase difference leading to a greater signal loss. The degree of the diffusion weighting applied in a sequence is measured by the "b-value", which depends on the paired gradient pulse amplitude, duration and interval [90].

Using the previously defined sequence, DW images are generated according to the following process. First, an image without diffusion gradients (b=0) is acquired, usually referred as the "b0 image". Then, at least three images are acquired assessing the diffusion in orthogonal directions with potentially different b-values. An isotropic diffusion image is obtained by combining the diffusion-weighted images using the geometric mean. The isotropic diffusion image is however T2-weighted and tissues with very long T2 decay time may appear bright though there is an actual diffusion, this effect is called the "T2-shine through". An apparent diffusion coefficient map (ADC) where the T2 effects are removed, can also be calculated from the isotropic diffusion image and the b0 image. Regions where diffusion is restricted have low ADC values and appear black on ADC maps contrary to DW images.

Thus, DWI can transcribe the cellularity of tissues. Tumors, characterised by a high cellularity and where diffusion is hindered, are therefore represented by low ADC values [9, 86]. DW-imaging has been used to differentiate lesions as benign lesions have higher ADC values than malignant ones [91].

As common MRI clinical protocols involve a long acquisition time, with 20 minutes needed on average, there has been in recent years an increased interest in developing faster protocols [92]. So-called "FAST" protocols have emerged which advocate the use of only one post-contrast image in addition to T1 and T2-weighted morphological images to reduce acquisition

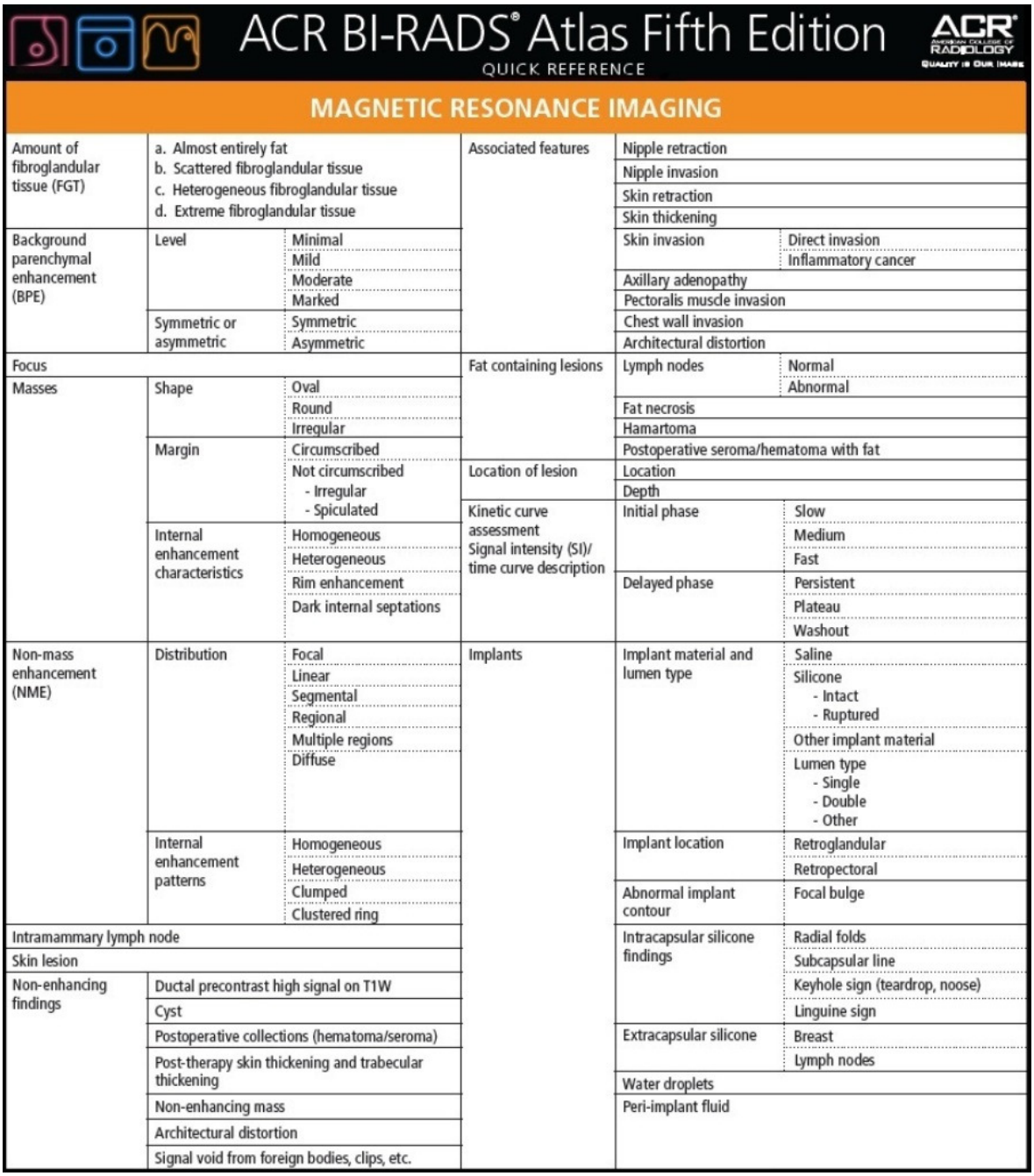
time. These abbreviated protocols have been proven to successfully save time and resources while maintaining equivalent diagnostic and lesion characterisation potentials though a slight decrease in specificity was sometimes reported [93, 94]. They nevertheless prevent the use of kinetic analyses, that require several post-contrast images. In this context, high-temporal resolution dynamic contrast-enhanced MRI (HTR-DCE MRI) was introduced. HTR-DCE MRI focuses more on the analysis of the contrast wash-in than of the wash-out in the lesions by oversampling the first minute after injection. Mann et al. [95] found that the dynamic information captured during this lapse of time could be used to detect and classify breast lesions with as high accuracy as conventional time-signal intensity curves. Milon et al. [92] reported equivalent performances using a FAST protocol combined with HTR-DCE MR imaging to a full standard protocol (T2-weighted images, T1-weighted DCE series, DWI series), questioning the future of the standard protocol. Ramtohul et al. [96] found that the wash-in slope from ultrafast breast MRI brought relevant information for the prediction of pCR to NAC.

1.5.4 MR imaging analysis

In order to reduce inter and intra-radiologist variabilities in assessing breast images and facilitating communication between medical doctors, the Breast Imaging Reporting & Data System (BI-RADS) atlas was introduced by the American College of Radiology (ACR) in 1993. Its fifth edition was released in 2013 [17]. It offers standardized terminology to characterize lesions, report on their structures and classify them within seven official "BI-RADS assessment categories". Management recommendations are associated with every category. The BI-RADS atlas was originally developed for mammography but nowadays encompasses three different lexicons of descriptors for mammography, ultrasound, and MRI. The MRI lexicon (Figure 1.4) includes an estimation of breast density, defined as the relative amount of fibroglandular tissue compared with fat in the breast. It describes different types of enhancements, characterises masses and non-mass enhancement, qualitatively estimates kinetic properties and reports on potential specific features such as implant characteristics.

BI-RADS descriptors assessed on MR images have been integrated over the years into multiple machine learning models to answer a wide variety of clinical questions. They were notably used to distinguish malignant from benign lesions [97], to predict molecular subtypes of breast lesions [98, 99], predict lymph node metastasis [100] or help predict invasiveness [101]. Many studies highlighted the association of BI-RADS features from pretreatment MR images with response to NAC. Harada et al. [102] found in a retrospective study that characterisation of breast edema on T2-weighted images could help predict the response to neoadjuvant chemotherapy. Malhaire et al. [103] showed that oval or round shape, no multifocality, non-spiculated margins and the absence of non mass enhancement indicated response to NAC. Uematsu et al. [104] reported in chemoresistant cancers large tumour sizes, the absence of mass effect, and very high intratumoral signal intensity on T2-weighted images, that could be a sign of intratumoral necrosis. As breast cancer is a highly heterogeneous disease, some BI-RADS descriptors were significantly associated with response to NAC only in specific molecular subtypes. Bae et al. [105] found that in triple-negative cancers, round or oval masses and the absence of peritumoral edema and intratumoral T2 high signal intensity seemed to be indicators of pCR.

The performances of univariate or multivariate models using BI-RADS descriptors only



ACR BI-RADS® Atlas Fifth Edition
QUICK REFERENCE
ACR AMERICAN COLLEGE OF RADIOLOGY
QUALITY IS OUR IMAGE

MAGNETIC RESONANCE IMAGING

Amount of fibroglandular tissue (FGT)	a. Almost entirely fat b. Scattered fibroglandular tissue c. Heterogeneous fibroglandular tissue d. Extreme fibroglandular tissue		Associated features	Nipple retraction	
				Nipple invasion	
				Skin retraction	
				Skin thickening	
Background parenchymal enhancement (BPE)	Level	Minimal		Skin invasion	Direct invasion
		Mild			Inflammatory cancer
		Moderate		Axillary adenopathy	
		Marked		Pectoralis muscle invasion	
	Symmetric or asymmetric	Symmetric		Chest wall invasion	
		Asymmetric		Architectural distortion	
Focus			Fat containing lesions	Lymph nodes	Normal
Masses	Shape	Oval			Abnormal
		Round		Fat necrosis	
		Irregular		Hamartoma	
	Margin	Circumscribed		Postoperative seroma/hematoma with fat	
		Not circumscribed - Irregular - Spiculated	Location of lesion	Location	
				Depth	
	Internal enhancement characteristics	Homogeneous	Kinetic curve assessment Signal intensity (SI)/ time curve description	Initial phase	Slow
		Heterogeneous			Medium
		Rim enhancement			Fast
		Dark internal septations		Delayed phase	Persistent
					Plateau
					Washout
Non-mass enhancement (NME)	Distribution	Focal	Implants	Implant material and lumen type	Saline
		Linear			Silicone - Intact - Ruptured
		Segmental			Other implant material
		Regional			Lumen type - Single - Double - Other
		Multiple regions			
		Diffuse			
	Internal enhancement patterns	Homogeneous		Implant location	Retroglandular
		Heterogeneous			Retropectoral
		Clumped		Abnormal implant contour	Focal bulge
		Clustered ring			
Intramammary lymph node				Intracapsular silicone findings	Radial folds
Skin lesion					Subcapsular line
Non-enhancing findings	Ductal precontrast high signal on T1W				Keyhole sign (teardrop, noose)
	Cyst				Linguine sign
	Postoperative collections (hematoma/seroma)			Extracapsular silicone	Breast
	Post-therapy skin thickening and trabecular thickening				Lymph nodes
	Non-enhancing mass			Water droplets	
	Architectural distortion			Peri-implant fluid	
	Signal void from foreign bodies, clips, etc.				

Figure 1.4: BI-RADS MRI lexicon. Source: ACR BI-RADS® Atlas Fifth Edition

were often not very high. Models combined BI-RADS descriptors with biological factors and kinetic parameters [106, 107] to increase the predictive power of models.

Conclusion

Breast cancer is a highly heterogeneous disease that is the leading cancer cause of deaths in women and whose burden has been continually increasing this last decade. Four main molec-

ular subtypes (Luminal A, Luminal B, HER2-enriched and triple-negative) of breast cancer have been identified thanks to immunohistochemistry testing. These subtypes have different characteristics, prognosis and reactions to treatments. Neoadjuvant chemotherapy has become the standard of care in locally advanced or aggressive Luminal B, HER2-enriched and triple-negative breast cancers. However, the variable response rate to NAC and the significant adverse effects associated to it, led to ponder how patients that would not achieve pCR could be identified before treatment or after a few cycles of chemotherapy. MRI was found to be one of the most precise imaging modalities to monitor tumor response and assess residual disease. Predictive models based on biological factors, kinetic parameters and BI-RADS reading from pretreatment MR images were designed. However, an advanced analysis of MRI emerged as the way forward to improve predictions of pCR to NAC.

Chapter 2

MRI-based radiomic analyses in breast cancer

Preface

This chapter presents an overview of advanced quantitative image analysis applied to MRI of breast cancer and its underlying principles. It describes the main steps of the radiomic study pipeline from image pre-processing to evaluation of predictive models. Finally, an analysis of the state-of-the-art of MRI-based radiomics in breast cancer, and especially in the prediction of the response to neoadjuvant chemotherapy, is developed.

2.1 Radiomics in cancer imaging

2.1.1 Introduction

Medical imaging offers a non-invasive opportunity to globally assess tumors in addition to biopsies that are invasive and extract and analyse a small piece of tissue from an often heterogeneous lesion. Analyses of radiological images rely on the idea that morphological aspects and heterogeneity of tumors or surrounding tissues convey information about their biological properties. Indeed, the BI-RADS initiative elaborates malignancy assessments and management recommendations based on visual and qualitative image analysis. However, more advanced and quantitative analysis of medical imaging is presumed to be able to go even further and transcribe the molecular characteristics, phenotype and microenvironment of tumors, thus bringing novel and complementary information that could potentially be combined with other clinical data. This is the idea on which radiomics, an emerging field of study in advanced medical image analysis, is built. It has known a considerable development in the last decade with the rise of artificial intelligence in the medical field [8, 108]. Radiomics can be defined as the high-throughput extraction of quantitative imaging features from radiological images [7, 8]. Radiomic studies focus mainly on three imaging modalities, computed tomography (CT), MRI and positron emission tomography (PET), though ultrasound-based radiomic studies are on the rise [108, 109]. The field of radiomics can distinguish three main types of studies depending on the machine learning techniques they used: handcrafted radiomic studies, deep learning studies and deep radiomic studies.

2.1.2 Handcrafted radiomics

Handcrafted radiomics extracts from designated regions of interest (ROI) (lesions, peritumoral regions...) features that quantify the shape, volume and intensity of lesions but also its texture, reflecting the heterogeneity of tumors. These features can be separated into three main categories: shape and volume descriptors, first-order statistics from the image intensity histogram and higher-order statistics that analyse the relationship between neighbor image voxels defining the texture patterns. Though its application fields are wide-ranging like in cardiac or brain imaging [110, 111], handcrafted radiomics is particularly relevant in oncology where tumor heterogeneity has been studied extensively. By extracting a large amount of features bringing objective information about tumors and using them in predictive modelling based on classical machine learning, handcrafted radiomics could provide a personalized approach to patient care and help build tools to assist medical doctors in decision-making. Handcrafted radiomics has been used to predict malignancy of suspicious lesions [112], molecular subtype [113] or histology [114] of tumors, overall survival [115, 116] but also disease-free survival [117] of patients in studies based on a large variety of organs (brain [118], lung [119], breast [120], rectum [116], pancreas [121]...). One of the leading trends in the radiomic field is the early prediction of response to therapies including radiotherapies [122], chemotherapies [13, 121, 123] or immunotherapies [124, 125].

2.1.3 Deep learning approaches

Deep learning is a sub-area of the global machine learning field that uses complex architectures made of numerous stacked layers of neurons, called neural networks [126]. These often quite large networks are optimized during the training process by minimizing a loss function with algorithms based on the gradient descent method. Amongst deep learning methods, convolutional neural networks (CNNs) are particularly adapted to image analysis. Unlike handcrafted radiomics which is based on the extraction of features from a precisely defined ROI, deep learning studies usually take the image as a whole and aim, thanks to filters and convolutions, to analyse the spatial relationship between voxels. CNNs have been widely used in medical imaging to segment lesions and organs, detect abnormalities but also to tackle the same kind of classification and prediction tasks that handcrafted radiomic models handle [126, 127]. They have the advantage of not requiring a delineation of the tumor which often constitutes a bottleneck in the development of radiomic studies. However, deep learning networks need a substantial amount of data to build relevant models and reduce the risk of overfitting the training set. As CNNs usually have millions of parameters, the optimization process also requires great computational power.

2.1.4 Deep radiomics

Often pitted against each other, handcrafted radiomic and deep learning approaches can nevertheless be combined to improve analyses, by delivering a new approach to extract features from images. Rather than extracting handcrafted features from ROIs, features could be extracted from images using deep learning models. Many strategies have been developed to extract so-called deep features but most of them are based on autoencoders [128, 129].

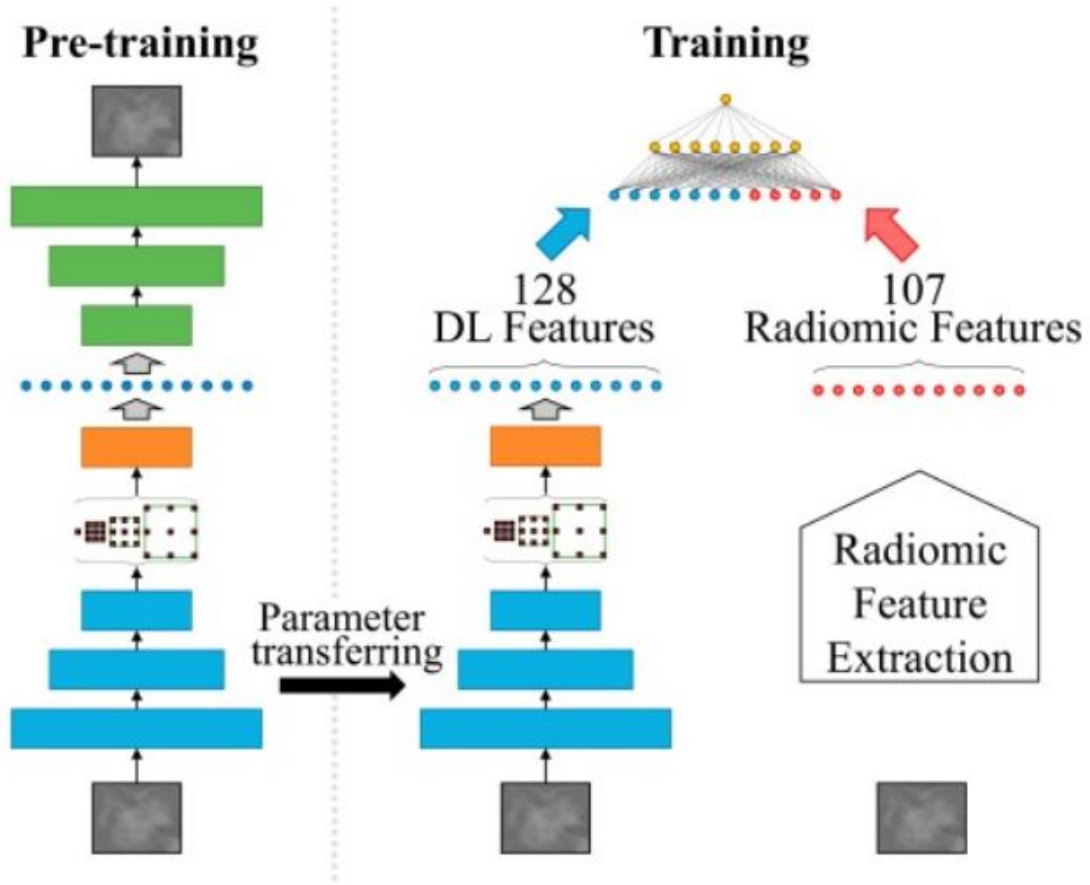


Figure 2.1: Extraction of deep radiomic features with an autoencoder. During the pre-training phase, an autoencoder is optimized. Then, deep learned (DL) features extracted from the latent representation (orange module) can be combined with other radiomic features and fed into predictive models. Source: Wang et al. [128].

Autoencoders are a type of neural network that thanks to two blocks, an encoder and a decoder, and without any labelling, learn a compressed representation of high dimensional data. The encoder module learns hidden features in the input data referred as latent features, from which the decoder part attempts to reconstruct the original image (Figure 2.1). Autoencoders are optimized by minimizing the loss estimating the divergence between the original and reconstructed data. In this context, the encoder acts as a feature extractor. Other strategies suggest to train CNNs for a specific clinical question and then used the second to last layer to extract features (Figure 2.2) [127].

The reasoning behind deep features is that they could convey a more abstract representation of the information contained in an image or find "hidden" information, that would be difficult to quantified visually. As they are extracted from the neural networks optimized on the training set, they could also be more data specific than the conventional radiomic features defined for all images no matter the imaging modality or the organ involved. But on the other hand, because of their abstract nature, deep features suffer from a lack of interpretability which makes it difficult to relate with other clinical data.

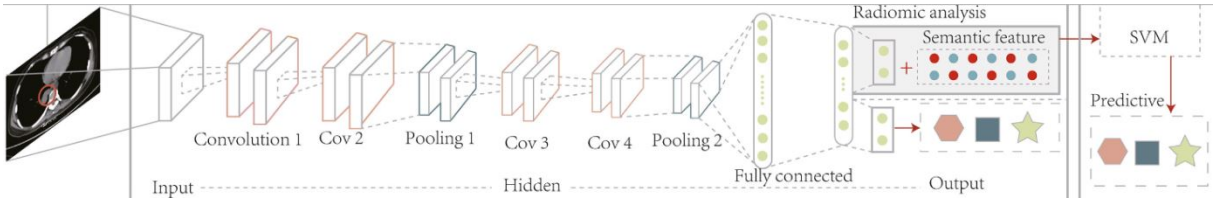


Figure 2.2: Extraction of deep radiomic features from the second to last layer of a CNN trained for a specific clinical question. Deep features can then be combined with other types of features to improve performances when fed into a classifier (SVM, Bayes classifier,...). Source: Zhang et al. [127].

2.1.5 Conclusion

This work will focus on the handcrafted radiomics approach presented previously, which constitutes the bulk of the literature on MRI-based radiomic analyses to predict pCR to NAC in breast cancer (Section 2.3.2). This approach has the advantage of building low computational and more interpretable models and needing fewer data.

2.2 Handcrafted radiomic analysis pipeline

This section will endeavour to describe a generic handcrafted radiomic analysis pipeline and analyze the challenges raised by each step (Figure 2.3).

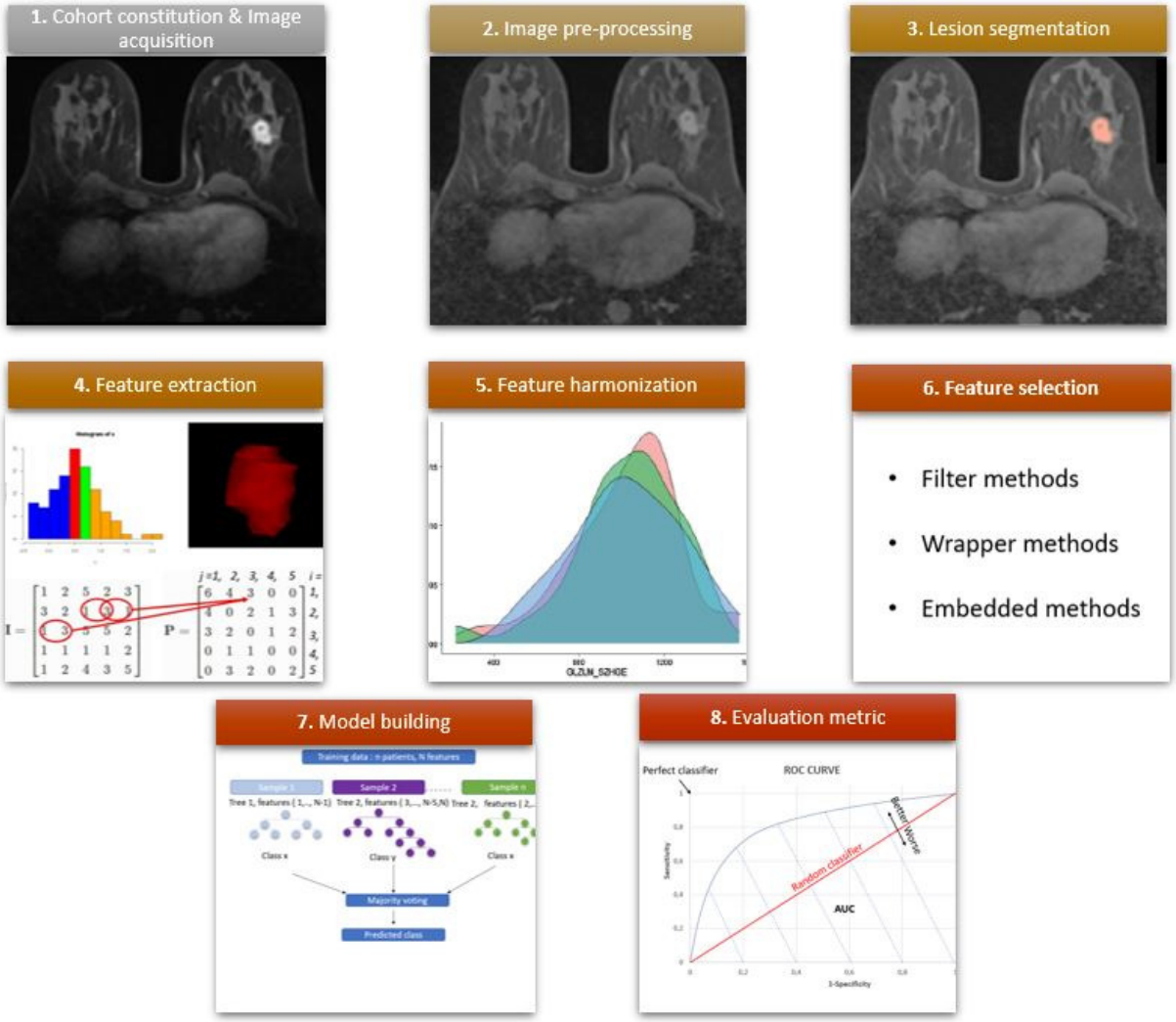


Figure 2.3: Main steps of the handcrafted radiomic analysis pipeline.

2.2.1 Cohort constitution & Image acquisition

There has been a steady increase in recent years in the number of multimodality studies published, with the goal to achieve a better characterization of tumors by combining the respective abilities of each modality [130]. Combinations of features from PET/CT [131, 132], PET/multiparametric MRI [133] or multiparametric MRI [134] have notably been used to predict early response to NAC in breast cancer.

The majority of published radiomic studies, no matter the organ, are retrospective studies including as few as around thirty patients up to hundreds of patients. They gather images either from a single institution or from multiple institutions (multicentric cohorts) [127]. Taking

into account the conditions of image acquisition is paramount when conducting a radiomic study because of their influence on radiomic feature values. This effect, called the "scanner effect" or the "center effect", has been reported in PET and CT [135], and in MRI [136–138]. Acquisition and reconstruction parameters including pulse sequence parameters, voxel sizes and field strengths, types and generations of scanners but also types of receiver coils, impact the statistical distribution of radiomic feature values in MRI [136–140]. Though single-institution studies are not homogeneous as several scanners, coils or sequences can be used, additional variations are to be observed in multicentric studies where radiologists' preferences could lead to differences in imaging protocols and patient positioning for example. The "scanner effect" has two main consequences on radiomic studies. First, in studies based on images acquired in diverse conditions, variations due to biological effects could be overlooked or minimized by the "scanner effect". Besides, it damages the exportability of radiomic models as threshold values determined for a specific dataset could be unsuitable for another set of images acquired in different conditions. It thus calls for corrective measures in order to improve the power and exportability of models.

Chapter 3 will introduce the cohort of our study and the imaging protocol while Chapter 4 will present a pipeline dedicated to breast MRI to reduce the "scanner effect" affecting radiomic features.

2.2.2 Image pre-processing

After acquisition, a first pre-processing step is required to correct potential artefacts, improve image quality (noise reduction, spatial smoothing) and adapt images to a standardized scale. In MR imaging, pre-processing is particularly important as images suffer from the bias field non-uniformity, which creates regional intensity variations [141, 142]. It is also necessary to correct bias field gain before the segmentation step as it can affect the ROI delineation. Besides, MR images are affected by the arbitrary units in which intensities are expressed, that vary between patients, scanners and acquisitions and that make comparisons difficult to interpret. Bias field correction, rarely applied in radiomic studies [18], and intensity normalization are thus required to correct intra and inter-acquisition inhomogeneities. Spatial resampling is also commonly applied as isotropic voxels are preferred to calculate some texture features [130, 143]. Resampling is usually calculated using B-Spline or higher-order interpolation. Image pre-processing is the first step in reducing the "scanner effect" in multi-scanner studies.

Chapter 4 and Chapter 5 will further explore the issue of pre-processing in breast MRI first using phantom experiments and then patient images.

2.2.3 Lesion segmentation

Lesion segmentation to define a ROI constitutes a critical step in handcrafted radiomic analyses. Though peritumoral tissue or parenchyma are sometimes investigated in breast MRI [144–146], the great majority of radiomic studies rely on a precise delineation of the tumors, demanding an expert radiologist involvement. Tumors are segmented in 3D or in 2D using the most representative slice. Tumor segmentation is a time-consuming and tedious task for radiologists. It thus constitutes a major bottleneck for large cohorts. Moreover, an inter and intra-radiologist variability in segmenting lesions is unavoidable due to preferences, radiolo-

gists' experience or different protocols. This variability affects radiomic features extracted from these segmented lesions [147, 148].

To make radiologists' work easier and attempt to reduce inter-operator variability, semi-automated and automated techniques have been proposed. Semi-automated methods usually need a first manual input to define a global localization of the tumor from which a more precised delineation is obtained using for instance, thresholding operations, clustering algorithms like the fuzzy c-means algorithm or active contours [149–151]. Figure 2.4 shows the semi-automated pipeline to segment breast lesions in MRI proposed by Teruel et al. [149]. Methods to automatically segment lesions are nowadays dominated by deep learning approaches. However, they are usually modality and organ specific and require a substantial amount of data and computational power to train models.

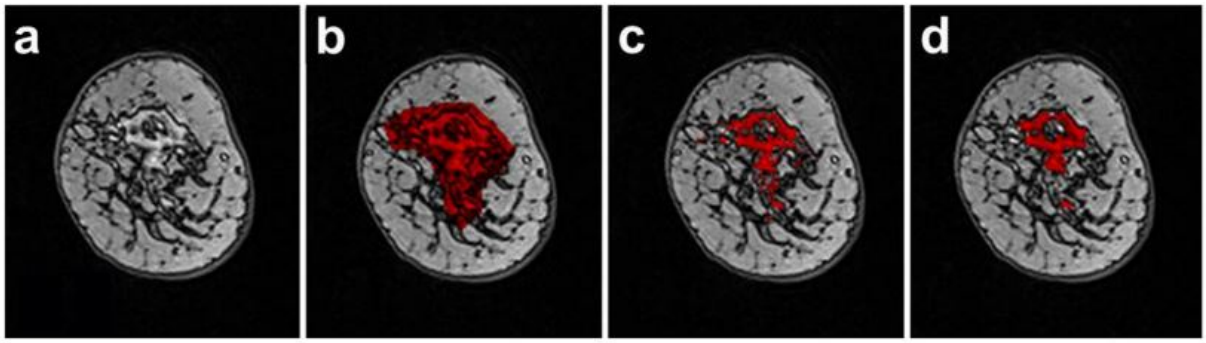


Figure 2.4: Steps of a semi-automated pipeline to segment breast tumors: ***(a)*** *first post-contrast image showing a breast tumor;* ***(b)*** *large manual ROI covering the whole tumor area;* ***(c)*** *same ROI after thresholding using a relative enhancement ratio (RER) criterion;* ***(d)*** *Segmentation obtained after performing morphological operations. Source: Teruel et al. [149].*

Chapter 7 will propose a 3D automated deep learning approach to segment tumors on breast T1-DCE images.

2.2.4 Feature extraction

As radiomics became more frequently used, difficulties arose with the lack of standardization in calculating features and the lack of accurate reporting on the extraction pipeline in studies. Therefore, in 2020, the Image Biomarker Standardization Initiative (IBSI) [143] proposed standardized mathematical definitions of 174 common radiomic features, recommendation guidelines and reference values. Pyradiomics [152], a python package dedicated to the extraction of radiomic features from medical images, is involved in the standardization process inspired by IBSI though it has some specific features. For convenience, as radiomic features were predominantly calculated with Pyradiomics (v3.0.1) in this thesis, the following description of radiomic features used the Pyradiomics glossary.

Before extracting features, image intensities are discretized either using a fixed number of bins or using a fixed bin width depending on the modality and normalization chosen [130]. Hundreds of features can be extracted but they are commonly gathered in 3 subgroups:

- Shape descriptors that quantify in 2D or 3D the geometric properties of the ROI including the description of its surface and volume. These features are globally not or very little affected by the "scanner effect" but they are very sensitive to segmentation variabilities. Table 2.1 indicates the shape features (3D) extracted by Pyradiomics (v3.0.1).

- First-order statistics or image intensity histogram features. These features analyse the intensity distributions inside the ROI represented by the image intensity histogram without considering neighborhood relationships between voxel intensities. Common statistics calculated from the histogram include mean, median, range, skewness, kurtosis... (Table 2.1).

- Higher-order statistics or texture features. These features analyse the statistical relationship between the intensities of 2, 3 or more neighboring voxels defining the texture patterns. Texture features are extracted from matrices that report the spatial relationship between voxels in the image. Five matrices, defined in IBSI [143], are commonly used:

  - the Gray Level Co-occurrence Matrix (GLCM) that measures the number of times two co-occurring values are represented in an image. Let I be the original image, and P the GLCM matrix. *P(i,j)*, with *i*, *j* respectively the row and column of the GLCM matrix, is equal to the number of times the combination of two voxels of levels *i* and *j* separated by *N* pixels along angle θ is present in the image I. In Pyradiomics, by default a GLCM matrix is calculated for each angle on which features are calculated. Features are then averaged across all matrices. In 2D, there are 8 possible angles and in 3D, 26 angles. However, matrices are often represented according to "directions", grouping angle θ with angle $\theta + 180$, to get symmetrical matrices.

    Figure 2.5 depicts an example of the construction of a GLCM matrix using a two-dimensional original image I with parameters *N*=1, $\theta = 0°, 180°$ (horizontal direction).

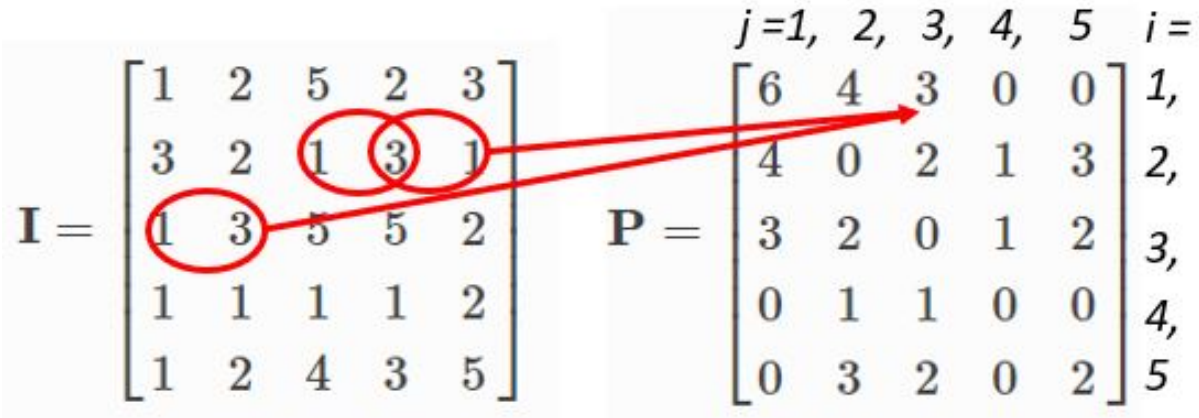


Figure 2.5: GLCM matrix (P) obtained using a 5x5 image (I) discretized to have 5 gray levels with N=1, $\theta = 0°, 180°$. Red arrows give an example of how a value is obtained in the GLCM matrix.

- the Gray Level Size Zone matrix (GLSZM) measures in an image the number of zones of size *N*, defined as a line of *N* connected voxels of the same gray level value. Two voxels are considered connected if the distance between them is equal to 1 using the infinity norm (8-connectivity in 2D). The GLSZM matrix is rotation invariant with only one matrix calculated from an image as opposed to the GLCM for instance. Let I be the original image, and P the GLSZM matrix. *P(i,j)* is equal to the number of zones of size *j* of voxels of level *i* in the image I. Figure 2.6 depicts an example of the construction of a GLSZM matrix using a two-dimensional original image I.

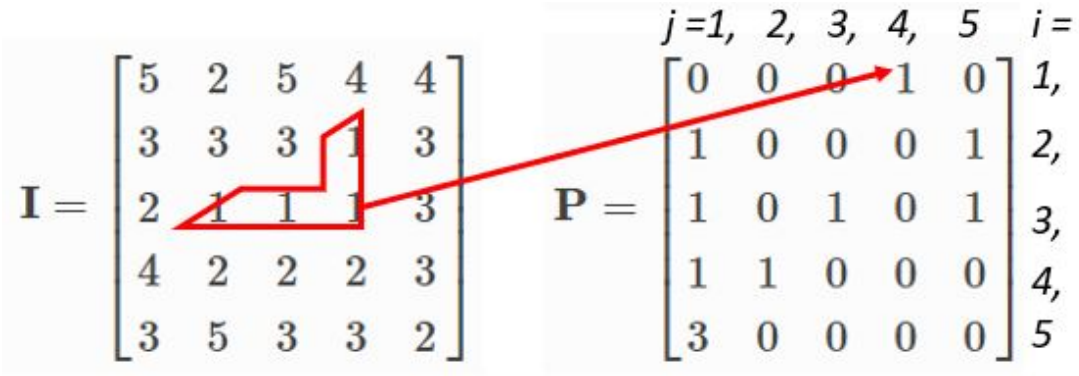


Figure 2.6: GLSZM matrix (P) obtained using a 5x5 image (I) discretized to have 5 gray levels. Red arrows give an example of how a value is obtained in the GLSZM matrix.

- the Gray Level Run Length Matrix (GLRLM) measures the number of runs of size *N* in the image, a run being a line of *N* consecutive voxels of the same gray level value. Let I be the original image, and P the GLRLM matrix. *P(i,j)* is equal to the number of runs of size *j* of voxels of level *i* along angle θ in the image I. There are 4 possible angles in 2D and 13 in 3D. Figure 2.7 depicts an example of the construction of a GLRLM matrix using a two-dimensional original image I with parameter $\theta = 0°$ (horizontal direction).

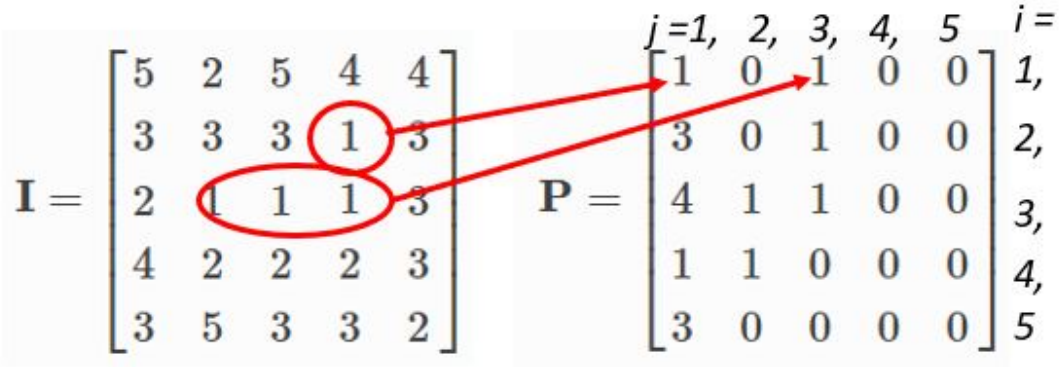


Figure 2.7: GLRLM matrix (P) obtained using a 5x5 image (I) discretized to have 5 gray levels with angle $\theta = 0°$. Red arrows give an example of how two values are obtained in the GLRLM matrix.

- the Gray Level Dependence Matrix (GLDM) counts the number of gray level dependencies in the image. A gray level dependency is defined as the number of voxels connected by a distance δ to a dependent center voxel. Two voxels of respective intensities *m* and *n* are dependent with a level α if $|m - n| \leq \alpha$. Let I be the original image, and P the GLDM matrix. *P(i,j)* is equal to the number of center voxels of gray level *i* with *j+1* dependent voxels for determined α and δ in the image I.

  Figure 2.8 presents the construction of a GLDM matrix using a two-dimensional original image with parameters $\alpha = 0$ and $\delta = 1$ (8-connectivity).

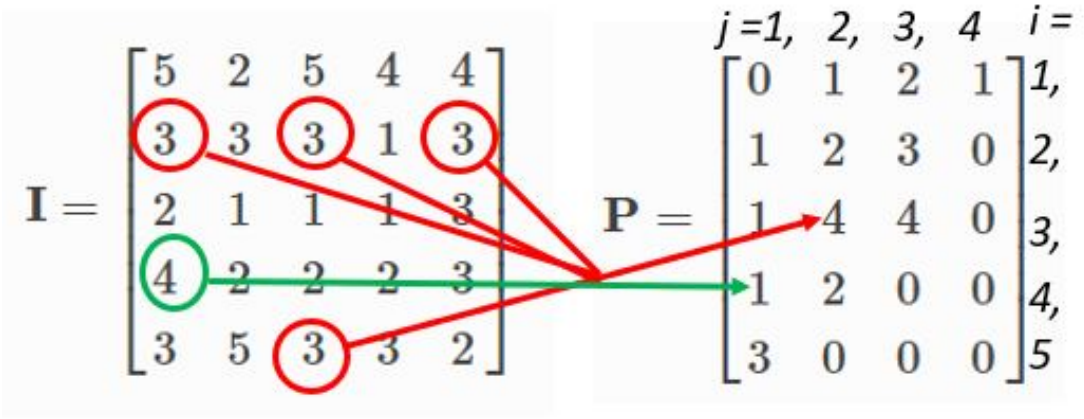
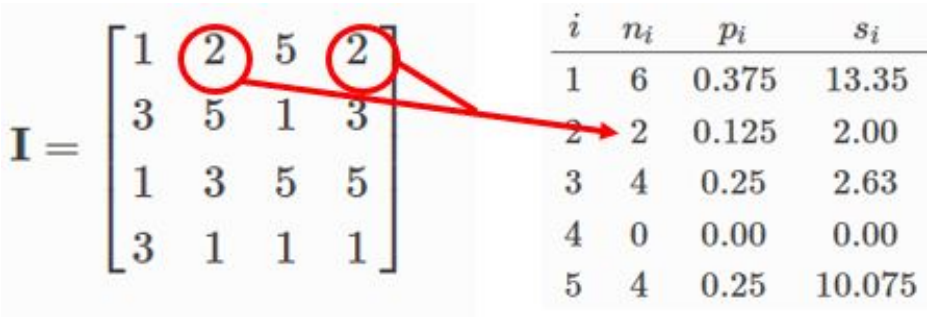


Figure 2.8: GLDM matrix (P) obtained using a 5x5 image (I) discretized to have 5 gray levels with $\delta = 1$ and $\alpha = 0$. Red and green arrows give an example of how a value is obtained in the GLDM matrix.

- the Neighboring Gray Tone Difference Matrix (NGTDM) measures the difference between the gray level of voxel and the average gray level of its neighbors within a distance δ. The sum of the absolute differences for a level *i* is then stored in the NGTDM matrix. The NGTDM matrix has n rows where n is the maximum pixel value of the matrix. Figure 2.9 shows the NGTDM matrix obtained using a two-dimensional original image with $\delta = 1$ using infinity norm (8-connectivity).



i	n_i	p_i	s_i
1	6	0.375	13.35
2	2	0.125	2.00
3	4	0.25	2.63
4	0	0.00	0.00
5	4	0.25	10.075

Figure 2.9: NGTDM matrix obtained using a 4x4 image (I) discretized to have 5 gray levels. Let i be the gray level value selected, n_i the number of pixels of level i, p_i the gray level probability with $p_i = n_i/N$ where N is the total number of pixels in the image, s_i is equal to the sum of absolute differences for the level i.

From these five matrices are extracted multiple features summarized in the Table 2.2.

Table 2.1: List of shape descriptors and first-order statistics calculated by Pyradiomics.

first-order features	Shape Descriptors
10^{th} Percentile	Elongation
90^{th} Percentile	Flatness
Energy	Least Axis Length
Entropy	Major Axis Length
Interquartile Range	Maximum 2D Diameter Column
Kurtosis	Maximum 2D Diameter Row
Maximum	Maximum 2D Diameter Slice
Mean Absolute Deviation	Maximum 3D Diameter
Mean	Mesh Volume
Median	Minor Axis Length
Minimum	Sphericity
Variance	Voxel Volume
Range	Surface Area
Robust Mean Absolute Deviation	Surface Volume Ratio
Root Mean Squared	
Skewness	
Total Energy	
Uniformity	

Definition of the features is available online in Pyradiomics documentation [152].

Features can be extracted from the original images but also from images that have been filtered to enhance for instance edges or sharp or coarse variations of intensities. Common filters include logarithm, gradient, square, squareroot, exponential, Laplacian of Gaussian and wavelet filters. Wavelet filtering in 3D encompasses the application of low (L) or high (H) pass filters in the three dimensions (x, y, z) resulting in eight different filters: HHH, LHH, HLH, HHL, LLH, LHL, HLL and LLL. Wavelet filtering can also be computed in 2D. In Pyradiomics, by default, order1 Coiflet wavelets are used for filtering.

2.2.5 Feature harmonization

In order to reduce the "scanner effect", further harmonization of the intensity-based features and texture features may be necessary (shape features are not further processed). Chapters 4 and 5 will present conventional methods like the ComBat approach to harmonize features while Chapter 6 will introduce an original strategy to harmonize features in a small data sample.

2.2.6 Feature selection

Between the different types of features calculated and the potential use of filters, hundreds of features are extracted from a single ROI. All of these features are nevertheless not of interest to answer the clinical question. Besides, amongst the features of interest, there are redundant, highly correlated or multicollinear features, which can create problems when they are used in regression models. Removing these features using methods like the variance inflation factor to deal with multicollinearity [153], can thus be a first option. The risk of overfitting the data, defined as the development of models that are too specific to the training data and would

Table 2.2: List of texture features calculated by Pyradiomics according to the five different matrices.

GLCM	GLDM	GLRLM	GLSZM	NGTDM
Autocorrelation	Dependence Entropy	Gray Level Non-Uniformity	Gray Level Non-Uniformity	Busyness
Cluster Prominence	Dependence Non-Uniformity	Gray Level Non-Uniformity Normalized	Gray Level Non-Uniformity Normalized	Coarseness
Cluster Shade	Dependence Non-Uniformity Normalized	Gray Level Variance	Gray Level Variance	Complexity
Cluster Tendency	Dependence Variance	High Gray Level Run Emphasis	High Gray Level Zone Emphasis	Contrast
Contrast	Gray Level Non-Uniformity	Long Run Emphasis	Large Area Emphasis	Strength
Correlation	Gray Level Variance	Long Run High Gray Level Emphasis	Large Area High Gray Level Emphasis	
Difference Average	High Gray Level Emphasis	Long Run Low Gray Level Emphasis	Large Area Low Gray Level Emphasis	
Difference Entropy	Large Dependence Emphasis	Low Gray Level Run Emphasis	Low Gray Level Zone Emphasis	
Difference Variance	Large Dependence High Gray Level Emphasis	Run Entropy	Size Zone Non-Uniformity	
Id	Large Dependence Low Gray Level Emphasis	Run Length Non-Uniformity	Size Zone Non-Uniformity Normalized	
Idm	Low Gray Level Emphasis	Run Length Non-Uniformity Normalized	Small Area Emphasis	
Idmn	Small Dependence Emphasis	Run Percentage	Small Area High Gray Level Emphasis	
Idn	Small Dependence High Gray Level Emphasis	Run Variance	Small Area Low Gray Level Emphasis	
Imc1	Small Dependence Low Gray Level Emphasis	Short Run Emphasis	Zone Entropy	
Imc2		Short Run High Gray Level Emphasis	Zone Percentage	
Inverse Variance		Short Run Low Gray Level Emphasis	Zone Variance	
Joint Average				
Joint Energy				
Joint Entropy				
MCC				
Maximum Probability				
Sum Average				
Sum Entropy				
Sum Squares				

Definition of the features is available online in Pyradiomics documentation [152].

achieve poor performances when tested on other datasets despite high results obtained during the training process, must also be taken into account. To reduce the risk of overfitting, feature selection is thus highly recommended.

Several methods to select features are conventionally used in radiomics. They can be separated in two broad groups: supervised and unsupervised methods. Unsupervised methods do not use any data labelling. They commonly include clustering techniques or dimensionality reduction approaches like the principal component analysis (PCA) and mostly aim to remove redundancy among features [154].

Supervised techniques select features based on their importance in solving a specific task and thus require data labelling. Three main approaches can be identified [155]:

- Filter methods: these methods can operate on a univariate or multivariate basis and do not require to train models. Using univariate statistics, filter methods select features based on their association with the target variable without considering the relationships and potential redundancy between features. Statistics used in this step depend on the nature of the data (continuous, discrete or qualitative data) and its distribution. Wilcoxon rank sum test, Student's *t*-test, Fisher score or the Chi-squared score are commonly calculated. On the other hand, the minimum Redundancy Maximum Relevance method (mRMR) is an example of a filter method that takes into account other features [156]. mRMR finds the minimal-optimal subset by iteratively calculating the F-statistic of features and using Pearson correlation coefficients (for continuous variables).

- Wrapper methods: wrapper methods investigate within a set of extracted features all possible subsets of features. Their relevance is assessed by the performance of a predictive model trained on this subset. Wrapper approaches usually work iteratively by

adding or removing a feature at each step to find the optimal subset that maximizes performances. Wrapper methods include the well-known forward and backward selection approaches [157].
The Boruta algorithm [158] is also an example of a wrapper method, based on random forest models [159]. It works in an iterative way and its main characteristic is to duplicate all features and shuffle randomly values to create “shadow features". At each iteration, random forest models are trained on real and associated shadow features and accept real features if their importance is higher than the importance of all the shadow features created from them or reject them when deemed unimportant. The algorithm stops when all features have been selected or rejected or when the maximum number of random forest runs has been carried out. The number of runs (iterations) can be increased if, after all the iterations, there are still doubts about some features which are then classified as “Tentative" by the algorithm. The Boruta method selects all features relevant to the problem instead of finding a minimal-optimal subset like some other approaches like mRMR.

- Embedded methods: contrary to filter and wrapper approaches, embedded methods perform feature selection while building the predictive model. Ridge and LASSO [160, 161] regression or tree-based models like Random Forest [159] can be cited as emblematic embedded methods [157, 162].

Chapter 5 will define the feature selection process performed in our analyses.

2.2.7 Model building

The model building step depends on the purpose of the study and whether it can be assimilated to a regression or classification task. It is also impacted by the way in which feature selection is performed as described in the previous section. Classification problems have been investigated in many radiomic studies such as in the prediction of molecular subtype [113], of the malignancy of a lesion [112] or the response to a treatment [13]. Predicting pCR to NAC is an example of a particular case of classification, called binary classification, where the output is restricted to two classes. Among regression models, logistic regression is well adapted to binary classification as it offers only two possible outcomes. Other well-known models suitable to classify patients include support vector machines (SVM), k-nearest neighbors algorithm (k-NN), naive Bayes classifier, neural networks or tree-based models like Random Forest [118, 151]. In our studies, we resort most of the times to random forest models as they handle binary, categorical and numerical features well, are quick to train and robust to outliers and non-linear data.

Random Forest (RF) is an ensemble learning method based on multiple decision trees, introduced by Breiman [159], that can be used either for classification or regression. The RF algorithm builds N decision trees made of nodes. Each tree is trained on a bootstrap sample (random sample with replacement) from the training data using a random subset of features. The final result is obtained by aggregating the results from each tree based on majority voting for classification and averaging for regression. Figure 2.10 shows a schematized representation of a random forest model trained for a classification task.

To achieve the best possible performances, hyperparameters that configure the Random Forest model architecture, including N the number of trees built, the metric and criteria on

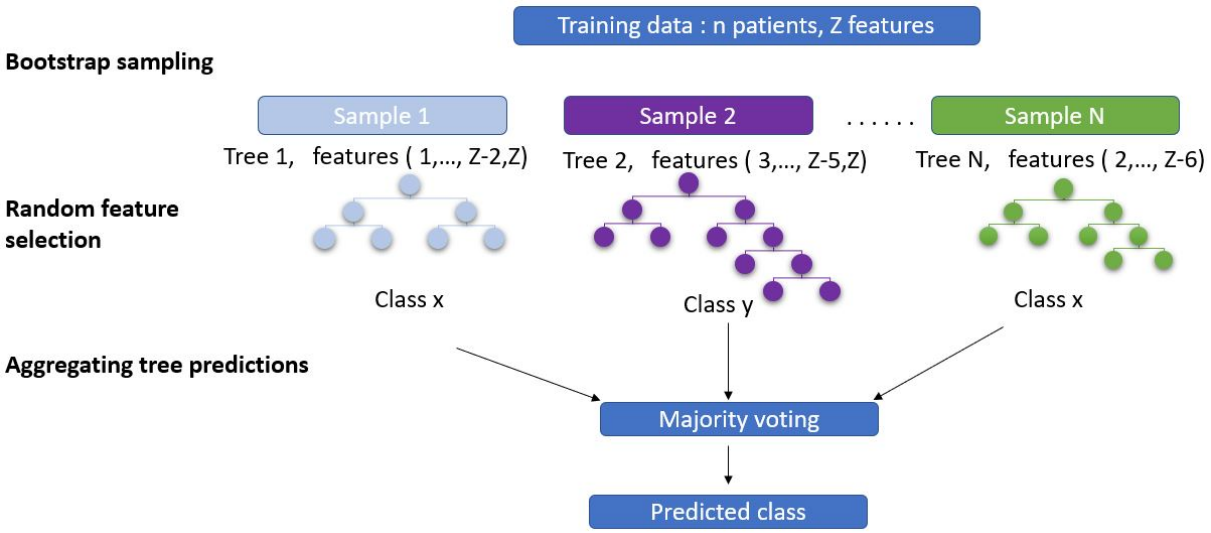


Figure 2.10: Random Forest model for classification built using N trees from a training set of n patients characterized by Z features.

which to split on at each node, the number of features randomly selected, usually equivalent to $\sqrt{N}$ and the maximum depth of the trees, must be optimized. This search for the best hyperparameters, called "model tuning" can be performed using different well-known techniques like random search or grid search, where all the potential combinations of hyperparameter values are points on a grid representing the search space that is exhaustively investigated for the optimal set of hyperparameters [163]. These methods are usually performed on a separate validation set or conjointly with K-fold cross-validation. K-fold cross validation is a resampling technique used to estimate the performance of a model by partitioning the data into K subsets and using iteratively K-1 subsets to train the models and the last subset to test it. Global performance is obtained by averaging the performances obtained on the test set in the K iterations. Figure 2.11 depicts a 5-fold cross-validation process.

This optimization process can take a considerable amount of time and depends on the number and types (continuous, discrete, categorical, conditional...) of the specific hyperparameters of the model trained. Depending on the circumstances, different approaches (random or grid search, Bayesian optimization, gradient descent...) may be considered to select hyperparameters [164] .

2.2.8 Model evaluation

Evaluation of radiomic models occurs first during the training phase to tune the classifiers and select the optimal set of hyperparameters and then during the testing phase to estimate the generalization ability of the models.

In the case of a binary classification task, predictions and ground truth of the model are commonly schematized in a confusion matrix (Figure 2.12) from which several evaluation metrics like the sensitivity (true positive rate or recall), specificity (true negative rate), precision,

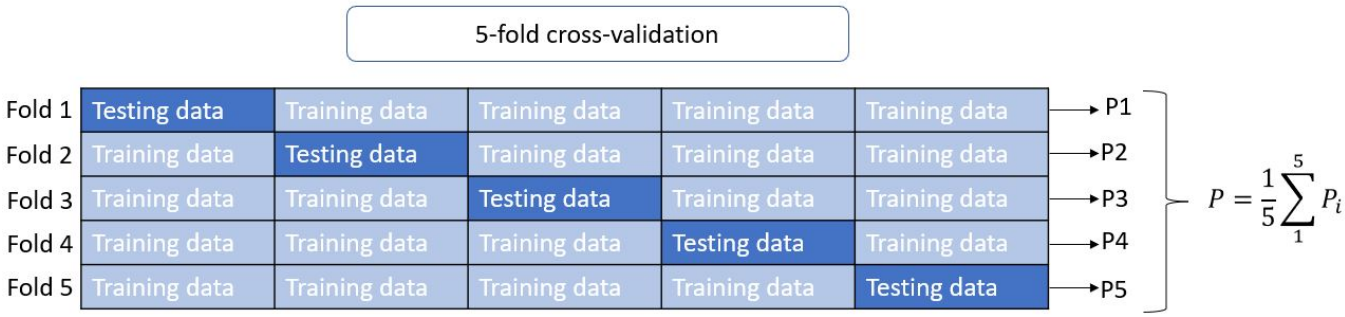


Figure 2.11: Example of a 5-fold cross-validation. Global performance P is obtained by averaging the performance P_i on the testing data of each fold i.

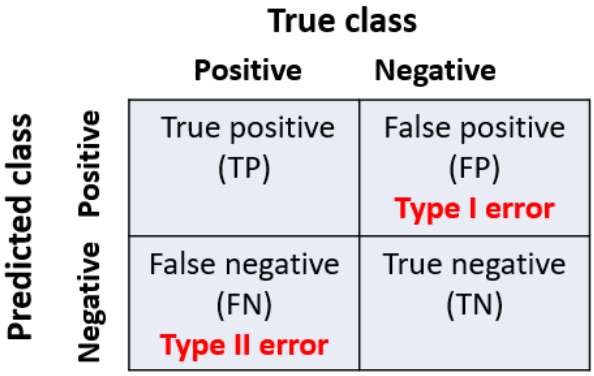


Figure 2.12: Confusion matrix.

Youden Index, F1-score or accuracy, amongst others, can be calculated:

$$Recall/Sensitivity/True\ positive\ rate = \frac{TP}{TP + FN} \tag{2.1}$$

$$Specificity/True\ negative\ rate = \frac{TN}{TN + FP} \tag{2.2}$$

$$Precision/Positive\ predictive\ value = \frac{TP}{TP + FP} \tag{2.3}$$

$$F_1 = \frac{2 \times (Recall \times Precision)}{Recall + Precision} \tag{2.4}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.5}$$

$$Youden\ Index = Sensitivity + Specificity - 1 \tag{2.6}$$

Accuracy is one of the most used metrics for binary and multi-class classification [165]. It is however not recommended in case of an imbalanced dataset as indeed, in these circumstances,

predicting the majority class every time will always lead to good performances. That is why the balanced accuracy, defined as

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{2.7}$$

was introduced. Precision can be defined as the probability that an observation classified as positive is actually positive whereas recall is the probability that a positive observation is classified as positive. The use of precision and recall as evaluation metrics depends on the predictive question the model intends to tackle. Precision will be favored when the cost of classifying a patient as positive is high whereas the cost of classifying him as negative is low. On the other hand, in cases where one wants to maximize the number of actual positive observations classified as positive, recall will be preferred. In problems where a good balance must be found between precision and recall without any preferences, the F1-score defined as the harmonic mean between them can be selected. The Youden index maximizes the sum of sensitivity and specificity and can equally be used when the cost of wrongly predicting an observation as positive or negative is comparable. Most of these metrics and the confusion matrix representation can be extended to multi-class problems. Sometimes, weighting of the different classes in metrics like the balanced accuracy can be introduced to ensure good performances in all classes [166].

Unlike the previously defined metrics, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve measures the quality of the predictions of the models while varying the classification threshold. The ROC curve graphically depicts the trade-off between sensitivity and specificity of a binary classifier (Figure 2.13). AUC evaluates the global ranking performance of a model rather than the final classification. It is one of the most popular metrics in radiomic studies [149, 167, 168].

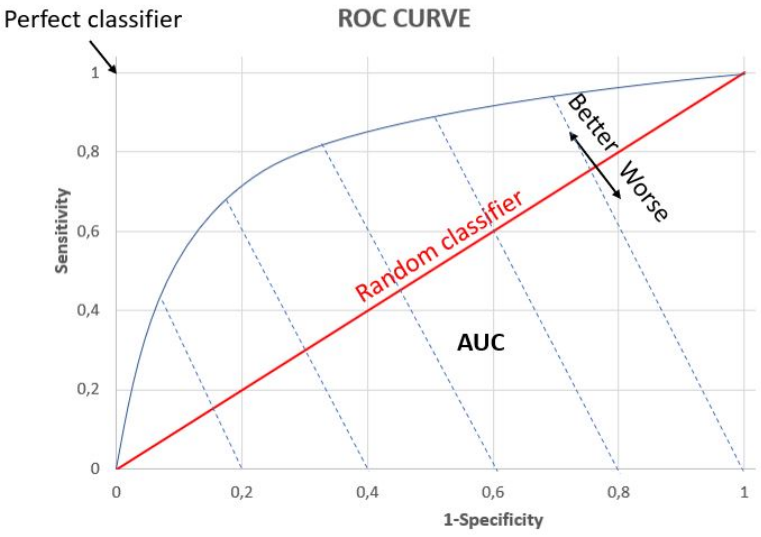


Figure 2.13: ROC curve showing an example of a curve obtained with a binary classifier. The left top corner indicates the point that should be reached by the curve of a perfect classifier with sensitivity = specificity = 1. The better the performance of the classifier, the greater is the area under the curve (AUC).

The evaluation of a radiomic model performance must also consider the set of patients on which to test the models to estimate its exportability. Having an independent test set remains the gold standard when testing predictive models. However when using small datasets, resorting to K-fold cross-validation is a widespread practice [146, 167–169]. Leave-one-out cross-validation (LOOCV) is an extension of K-fold cross-validation where K is equal to the number of patients, leading to train at each iteration the model on all the observations but one that is used to test the model [170]. This method can be computationally expensive but reduces the bias of the results as models are trained on almost all the dataset. Another benefit of K-fold cross-validation is to get a mean and standard deviation of the performance.

2.3 Breast radiomics: state of the art

2.3.1 Applications of radiomics in breast cancer

Radiomics has been applied in breast cancer to tackle many clinical questions, including the prediction of the response to NAC, using a wide variety of modalities [120, 171]. The paragraph below describes a few examples of radiomic applications in breast cancer other than the prediction of pCR to NAC.

First, radiomic studies have attempted to distinguish malignant from benign lesions. Nie et al. [172] combined volume, shape and GLCM features extracted from post-contrast MR images into a neural network to differentiate lesions with an AUC of 0.82 on a separate validation set. To textural features, Wang et al. [173] decided to add kinetic parameters (K_{Trans}, K_{ep} and V_e, that can be represented in anatomical maps) to improve identification of malignant lesions. Radiomic-based model can also predict lymph node metastases. In a prospective study, Liu et al. [134] used features from the wavelet-transformed images combined with shape, size, first-order statistics and texture features to predict metastases or lymph node metastases with an AUC of 0.76. Radiomic features have simultaneously been used to identify molecular subtypes of breast cancer [113] or histological type of tumors [174]. Besides, an important application of radiomics in breast cancer concerns the prediction of survival outcomes. Park et al. [117] built a radiomic signature with clinical factors and the accompanying nomogram to predict disease-free-survival in patients diagnosed with invasive breast cancer. Cancer recurrence prediction is a topic in which radiomics can bring value as well. Chan et al. [175], for example, identified patients with high risk of recurrence using features extracted from the patterns of wash-in and wash-out of DCE pre-treatment images.

2.3.2 Prediction of pCR to NAC in breast cancer using MRI

Literature review methodology

There has been a growing interest in the prediction of pCR to NAC in recent years. From a few studies between 2010 and 2015, about half a dozen studies have been published every year since. Researchers have attempted to address this question using both deep learning and handcrafted radiomic approaches. Radiomics studies designed to predict pCR have been based on CT [176, 177], PET [178, 179], MR [13] or ultrasound [180] images but also using combinations of these modalities with PET/CT [131, 132, 181] and PET/MRI [133] studies. The

following section will specifically focus on MRI, the modality used in this thesis. We searched PubMed, Web of Science and Google Scholar databases until June 1, 2022 to gather all published articles in English describing studies designed with MR images only and that calculated handcrafted texture features or used deep learning methods to build predictive models to predict pCR to NAC in breast cancer. The following keywords were used to search databases: "MRI" AND "Breast" AND ("Neoadjuvant chemotherapy" OR "Neoadjuvant therapy") AND ("Radiomic" OR "Texture" OR "Deep learning"). Articles that only measured the association of features with pCR without building models were not reported. Articles that used PET/MRI or used other modalities were rejected as well.

Table 2.3 lists studies that calculated handcrafted texture features to predict pCR. The first columns of Table 2.3 provide the reference, year of publication, molecular subtypes used to build the model and the number of patients of the study. When several lines describe molecular subtypes and patients for a specific study, it means that several predictive models were built with usually one model designed for all the subtypes and other ones for specific subtypes. The table then reports the MR sequences of the images. The "Multicentric" column indicates if images were acquired in one or multiple institutions. The mention "MS" standing for "Multiscanner" means that patients were scanned in one center but on several scanners. Table 2.3 indicates if image pre-processing was applied before feature extraction. The "ROI" column describes the region from which features were extracted in the images. Though in the vast majority of cases, the ROI only considers the tumor region, other sub-regions, such as the peritumoral regions, were investigated in a few studies. Finally, the selected features, evaluation methods and AUC values on the test set or using cross-validation on the training set are mentioned.

Similarly, Table 2.4 lists all published studies using deep learning approaches to predict pCR and notably indicates if clinical and biological variables were combined with images to build the models.

Literature review analysis

This review process found 36 handcrafted radiomic and 11 deep learning-based studies designed to predict pCR to NAC in breast cancer. From the main characteristics of the studies (described in Tables 2.3 and 2.4), a number of observations can be made while some questions need to be further explored.

First, we can see an important rise in the last few years in the number of studies published about the prediction of pCR in breast cancer using either handcrafted or deep learning approaches (no study combining them was found). There are comparatively fewer deep learning studies than handcrafted ones but the deep learning trend in radiomics seems to have pick up more recently which is in par with the global deep learning surge. Deep learning studies have on average 194 patients (median=141, interquartile range (IQR) [121, 234]) while handcrafted studies gather 146 patients (median=95, IQR [60, 160]). This difference is due to the necessity to constitute larger cohorts to use deep learning approaches. As it may prove tricky, this can also explain why deep learning studies are less frequent.

Building models for all the molecular subtypes of breast cancer or for a subset of them is one of the first main questions to tackle in the design of a radiomic study. Indeed, as

Table 2.3: Listing of handcrafted radiomic studies to predict pCR to NAC in breast cancer.

Reference	Year	Subtype	Number of Patients	Treatment	Modality	Multicentric	2D/3D	Image Preprocessing	ROI	Features	Evaluation	AUC
Golden [167]	2013	TN	60	Pre/Post	DCE	Yes	2D	Kinetic map	Tumor	BI-RADS Clinical/Bio FO/Texture	MCCV	0.68±0.05
Teruel [149]	2014	All	58	Pre	DCE	No	2D	-	Tumor	FO/Texture	UA	0.69
Wu [170]	2016	All	35	Pre/Mid	DCE	No	3D	Temporal PCA	Intratumor partitioning	FO/Texture	LOOCV	0.79
Giannini [182]	2017	All	44	Pre	DCE	No	3D	-	Tumor	FO/Texture	CV	0.80
Henderson [183]	2017	All	88	Pre/Mid	T2	No	3D	-	Tumor	Texture	UA	0.85
Banerjee [168]	2017	TN	41	Pre/Post	DCE	Yes	3D	Kinetic map	Tumor	FO/Texture	MCCV	0.83±0.01
Braman [144]	2017	All TN/HER2+ HR+/HER2-	117 47 70	Pre	DCE	Yes	3D	-	Tumor Peritumoral	FO/Texture Kinetics	Test set (39) CV CV	0.74 0.83 ± 0.03 0.93 ± 0.02
Chamming [184]	2017	All HR+/HER2-	85 69	Pre	DCE, T2	No	2D	-	Tumor	FO/Texture	UA	NA 0.67
Thibault [169]	2017	All	38	Pre/Mid	DCE, T2	No	3D	-	Tumor	FO/Texture Kinetics	CV	1
Fan [145]	2017	All	103	Pre	DCE, T2	MS	3D	-	Tumor BPE	FO/Texture Kinetics	Test set (46)	0.70
Machirredy [185]	2019	All	55	Pre/Mid	DCE	No	2D	Parametric fractal map	Tumor	Multi-resolution fractal analysis Kinetics	Test set (15)	0.80
Cain [150]	2019	All TN/HER2+	269 151	Pre	DCE	MS	3D	-	Tumor	FO/Texture	Test set (134) Test set (75)	0.60 ± 0.05 0.70 ± 0.06
Liu [134]	2019	All HR+/HER2- TN	586	Pre	DCE, T2, DWI	Yes	3D	Z-score	Tumor	Clinical/Bio FO/Texture	Test sets	0.79 0.87 0.84
Drukker [186]	2019	All	158	Pre	DCE	MS	3D	-	Tumor	FO/Texture	Bootstrap test	0.82± 0.03
Braman [187]	2019	HER2+	70	Pre	DCE	Yes	3D	-	Tumor Peritumoral	FO/Texture	Test set(28)	0.80± 0.09
Tahmassebi [107]	2019	All	38	Pre/Mid	DCE, T2, ADC	No	3D	-	Tumor	BI-RADS Kinetics	CV	0.86± 0.06
Bitencourt[188]	2020	HER2+	311	Pre	DCE	Yes	3D	-	Tumor	FO/Texture	Test set (134) Test set (75)	0.60 ± 0.05 0.70 ± 0.06
Chen [189]	2020	All	158	Pre	DCE, T2	No	2D	-	Tumor	FO/Texture	Test set (48)	0.84
Zhou[146]	2020	All	55	Pre	DCE	No	3D	-	Tumor Peritumoral	FO/Texture	CV	0.89± 0.03
Chen [190]	2020	All	158	Pre	DCE, ADC	No	3D	-	Tumor	FO/Texture	Test set (28)	0.84

Reference	Year	Subtype	Number of Patients	Treatment	Modality	Multicentric	2D/3D	Image Preprocessing	ROI	Features	Evaluation	AUC
Xiong [191]	2020	All	125	Pre	DCE	No	3D	-	Tumor	Clinical/Bio FO/Texture	Test set(62)	0.94
Fusco [192]	2020	All	45	Pre	DCE	Yes	3D	-	Tumor	FO/Texture Kinetics	UA	0.93
Bian [193]	2020	All	152	Pre	DCE, T2, ADC	No	2D	-	Tumor	FO/Texture	Test set (45)	0.93
Eun [13]	2020	All	136	Pre/Mid	DCE, T2 DWI,ADC	No	2D	-	Tumor	FO/Texture	CV	0.82±0.03
Sutton [194]	2020	All	273	Pre/Post	DCE	MS	3D	Histogram standardization	Tumor	Clinical/Bio FO/Texture	CV	0.83±0.05
Hussain [12]	2021	All	166	Pre/Mid	DCE, T2	Yes	3D	-	Tumor Peritumoral	Clinical/Bio FO/Texture	Test set	0.98
Granzier [18]	2021	All	320	Pre	DCE	Yes	3D	Bias field correction Histogram matching	Tumor	Clinical/Bio FO/Texture	Test sets	NA*
Pesapane [14]	2021	All	83	Pre	DCE	No	3D	Intensity Normalization	Tumor	Clinical/Bio FO/Texture	CV	0.83±0.05
Montemzzi[195]	2021	All	60	Pre	DCE	No	3D	Intensity Normalization (mean=0, sd=1000)	Tumor	Clinical/Bio FO/Texture Kinetics	LOOCV	0.91
Kolios[196]	2021	All	102	Pre	T2	No	2D	-	Tumor Margins	FO/Texture	LOOCV	0.78
Nemeth [15]	2021	TN	75	Pre	DCE, T2, DWI	No	3D	-	Tumor Parenchyma	FO/Texture	Test set (18)	0.83
Caballo [16]	2022	All Luminal A Luminal B HER2+ TN	251 107 47 25 72	Pre	DCE	MS	3D	-	Tumor Peritumoral	FO/Texture Kinetics	LOOCV	0.71 0.82 0.82 0.84 0.80
Yoshida [197]	2022	All	78	Pre	DCE, T2, DWI	No	3D	-	Tumor	BI-RADS Clinical/Bio FO/Texture Kinetics	Test set (20)	0.76
Jimenez [198]	2022	TN	80	Pre	DCE	No	3D	-	Tumor	TII FO/Texture	CV	0.75±0.03
Li [199]	2022	All	448	Pre	DCE,T2 ADC	No	3D	-	Tumor Peritumoral	FO/Texture	Test set(86)	0.92
Peng [200]	2022	All	356	Pre	DCE	No	3D	-	Tumor	Clinical/Bio FO/Texture Kinetic	CV	0.78±0.02

Kinetic map**: texture is calculated on maps of kinetic parameters instead of native or filtered images; **Pre**: Pre-treatment; **DCE**: T1-weighted dynamic-contrast enhanced sequence; **Mid**: Mid-treatment; **Post**: Post-treatment; **MS**: Multiscanner but not multicentric study; **FO/Texture**: first-order, shape and texture features; **Clinical/Bio**: Clinical and biological variables; **Kinetics**: Kinetic parameters; **TIL**: Tumor-infiltrating lymphocytes; **UA**: univariate analysis, **CV**: Cross-validation, **LOOCV**: Leave-one-out cross validation; **MCCV**: Monte-Carlo cross-validation; **Test set(N)**: test set of N patients with N included in the global number of patients; **AUC**: area under the curve on the test set **NA: No performance value is indicated for this study as the highlighted conclusion was that handcrafted radiomics cannot predict successfully pCR.*

Table 2.4: Listing of deep learning radiomic studies to predict pCR to NAC in breast cancer.

Reference	Year	Subtype	Number of Patients	Treatment	Sequence	Multicentric	2D/3D	Clinical	Model	Evaluation	AUC
Huynh [201]	2017	All	64	Pre	DCE	No	3D	No	VGG CNN	LOOCV	0.85±0.03
Ha [202]	2018	All	141	Pre	DCE	MS	2D	No	VGG16	Test set(28)	0.98
Ravichandran [203]	2018	All	166	Pre	DCE	No	3D	Yes	AlexNet	Test set(33)	0.85
Adoui [204]	2019	All	42	Pre	DCE	No	3D	No	CNN	CV	0.92
Liu [205]	2020	All	131	Pre	DCE	Yes	3D	No	CNN	CV	0.72 ± 0.08
Braman [206]	2020	HER2+	157	Pre	DCE	Yes	3D	No	CNN	Test set(28)	0.85
Qu [207]	2020	All	302	Pre/Post	DCE	No	3D	No	CNN	Test set(48)	0.97
Duanmu [208]	2020	All	112	Pre	DCE	Yes	3D	Yes	CNN	Test set(22)	0.80
Joo [209]	2021	All	536	Pre	DCE, T2	No	3D	Yes	CNN	Test set(107)	0.89
Massafra [210]	2022	All	151	Pre	DCE	Yes	3D	Yes	CNN	Test set(45)	0.80
Peng [200]	2022	All	356	Pre	DCE	No	3D	Yes	ResNeXt50	CV	0.83 ±0.02

***Pre**: Pre-treatment; **Mid**: Mid-treatment; **Post**: Post-treatment; **MS**: Multiscanner but not multicentric study; **CV**: Cross-validation, **LOOCV**: Leave-one-out cross validation; **VGG, AlexNet, ResNetXt50**: specific CNNs common in the deep learning literature; **CNN**: ad-hoc CNNs that were not built using a specific model; **Test set(N)**: test set of N patients with N included in the global number of patients.*

explained in Chapter 1, molecular subtypes have different characteristics, prognoses, evolution patterns, treatments and chemosensitivities which may advocate for the design of separate models. Though the vast majority of the studies (26/36, 10/11) are not subtype-specific, TN [15, 167, 198] and HER2+ [187, 188, 206] based studies have been developed. Starting from a global model, some articles [16, 134, 144, 150] have also designed complementary models for a particular set of subtypes with common pools being TN/HER2+ and HR+/HER2-. Refining the molecular subtypes of the models frequently proved to increase performances like in the study by Liu et al. [134] where the global model has an AUC of 0.79 and the HR+/HER2- model an AUC of 0.87. However, restricting the subtypes tends to reduce the statistical power of the study and often leads to evaluate models with cross-validation methods instead of using an independent test set.

The time in the therapy at which to set the study must also be discussed. Most of the studies extract features from pre-treatment images but some works are based on the combination of pre, mid and sometimes even post-treatment images [12, 13, 107, 167, 168, 170, 183, 185, 207] or evaluate changes in radiomic features between the different time points. Being able to predict pCR before the beginning of NAC would indeed be ideal but as it is a complex question, stopping treatment after a few cycles of an ineffective chemotherapy would still be useful. Hussain et al. [12] noted a significant boost in AUC from 0.88 to 0.92 on an independent test set when adding mid-treatment images to pre-treatment images. Furthermore, Eun et al. [13] showed that mid-treatment T1-weighted DCE based models achieved higher performances than models trained on pre-treatment images only or on a combination of pre and mid-treatment images.

Selecting the MR sequences is another important step in the study design. The basic and most frequent sequence consists in T1-weighted DCE (34/36 in handcrafted studies and 11/11 in deep learning studies) to which is sometimes added T2-weighted images (14/36, 1/11), ADC maps (5/36, 0/11) and DW-weighted images (4/36, 0/11). Liu et al. [134] found that multiparametric signatures (T1-DCE, T2, ADC) outperformed any single sequence-based models as observed similarly by Nemeth et al. [15]. However, Eun et al. [13] came to a different conclusion with better performances when using T1-DCE mid-treatment features alone. The choice of the best T1-weighted DCE image has also been debated. While using the first post-contrast image is common because peak enhancement time of breast tumors usually happens quickly, Montemezzi et al. [195] chose the third dynamic and its associated subtraction image. Yoshida et al. [197] also used subtraction images while Huyn et al. [201] investigated different pre and post-contrast images and achieved best performances with pre-contrast features.

Once the cohort constituted and image acquisition parameters set, a ROI must be defined in handcrafted studies. This step raises two questions: which regions to include in the ROI and how to delineate it. Some works have indeed highlighted the potential benefit of combining tumoral with peritumoral and parenchymal features for improved performances [131, 145, 187]. Peritumoral tissue, in particular, is thought to bear the mark of the angiogenetic activity of the tumor and the invasion of lymphatic vessels which could prove useful in the prediction of the response.

As previously mentioned, delineation of the tumor can be done manually, semi-automatically and entirely automatically. Tumors can be segmented in 3D like for almost all reported stud-

ies, or in 2D selecting the most representative slice. However, inter-radiologist variability in manual and semi-automatic segmentation of tumors heavily impacts radiomic feature values. As the reproducibility of radiomic features is paramount to export models, efforts have been made to improve this point by asking several radiologists to segment lesions and using a final segmentation based on consensus [15, 144]. Granzier et al. [211] and Saha et al. [148] studied the robustness of radiomic features by calculating the intraclass correlation coefficients (ICC) of features extracted from different segmentations. Differences in radiologists' segmentations were assessed with the Dice similarity coefficient (DSC). As an alternative to altering the segmentations, they proposed to only work on features with an ICC superior to a selected cut-off value (0.8 for example), which would be deemed robust to segmentation variabilities.

After the segmentation step, features must be extracted and selected to build models. The choice of features to use for predictive modelling is extremely wide but features usually fall into one of the five categories: BI-RADS features, kinetic parameters, shape or first-order/texture features and as a complement, clinical and biological data. The evidence concerning the use of different sources of features is mixed. Pesapane et al. [14] found their clinical and biological model and clinical and biological combined with radiomics based model to have equivalent performances. By contrast, Peng et al. [200] and Hussain et al. [12] indicated that adding clinical or biological data improved their model performances. Finally, Granzier et al. [18] underlined that the clinical and biological model achieved the best performances as they fail to build a model better than the random classifier using MR-based radiomic features. Among texture parameters, GLCM features and features from the wavelet-transformed images have been highlighted by some studies as particularly useful to predict pCR [146, 170]. Mallat et al. [212] notably underscored the scale separation and linearization ability of wavelets, that is used in deep learning architectures.

The choice of classifiers to predict pCR is an open question and quite often studies compare the performances of several classical algorithms like naive Bayes, SVM, random forest, logistic regression [15].

Comparing performances of models is a tricky task as all studies might not have the same definition of pCR and the distribution of molecular subtypes can greatly influence performances. Including Luminal A tumors that have a well-know very low rate of achieving pCR into cohorts can help achieving good performances. On a global level, AUCs on the test set cover a range from 0.70 to 0.98. Comparisons between deep learning and handcrafted approaches are again difficult to make except in the study by Peng et al. [200] where a radiomic and a deep learning models were tested on the same cohort. The deep learning model performed significantly better than the radiomic approach (AUC 0.83 vs 0.78, p <0.001). In this review, performances were always reported with the AUC metric, which gives a certain leeway when defining a classification threshold to study misclassified patients.

Finally, some aspects of the "quality" of the studies as defined by the radiomic quality score (RQS) [213] could be mentioned. A few studies [12, 205, 208, 210] (1/36, 3/11), used images from the public database "Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis" (I-SPY1 TRIAL) [214]. The I-SPY1 database is a multicentric cohort acquired between 2002 and 2006 to test in a prospective study the

ability of MRI to predict pCR to NAC in breast cancer. Nevertheless, the majority of the databases were private. Though the imaging protocol was usually well described and most of the segmentations obtained by the consensus of several radiologists, very few studies talked about any of the pre-processing steps that are essential in MR studies. Among radiomic handcrafted studies, only one study performed bias field correction [18], and five papers (5/36) reported performing intensity normalization or spatial resampling before extracting features. Though numerous studies were multicentric, a fact that is welcome, only Granzier et al. [18] and Caballo et al. [16] conducted further harmonization of the features to reduce the "scanner effect". At last, a substantial number of studies (20/36 of handcrafted and 4/11 of deep learning studies) did not have an independent test set to evaluate the model and needed to resort to cross-validation techniques.

Conclusion

This chapter introduced the field of radiomics, its underlying biological principles and the main techniques (handcrafted features, deep learning approaches, deep features) associated with it. The radiomic pipeline for the handcrafted features was described precisely while at the same time analyzing the challenges raised by each step like the "scanner effect", the impact of inter-radiologist variability in segmenting lesions and the lack of standardization in defining features. A final section went through the literature on MRI-based approaches to predict pCR to NAC in breast cancer, notably highlighting the common lack of information on image pre-processing or feature harmonization in the majority of the studies.

Chapter 3

MR study design & first analyses

Preface

Being able to predict the response to NAC before the beginning of treatment or after a few cycles of chemotherapy could considerably improve patient care. Radiomic studies combining texture features, clinical and biological data, BI-RADS or kinetic parameters have attempted to address this complex issue (Chapter 2). This chapter will introduce the cohort used in this thesis and propose first models based on clinical and biological data and BI-RADS features. A reduced version of the cohort, its characteristics and association with pCR of its features were published by Malhaire et al. [103].

3.1 Study design

3.1.1 Cohort constitution

This retrospective study was initiated by Dr. Caroline Malhaire, radiologist at Institut Curie with 15 years of experience in breast MR imaging and was approved by the institutional review board of Institut Curie (IRB number OBS180204). The study included adult women with locally advanced or invasive breast cancer treated with NAC (anthracycline and taxane regimen with herceptin for HER2-positive tumors) at Institut Curie between 2016 and 2020 and who were scanned with MRI before the beginning of NAC. Pregnant women, women previously treated for ipsilateral breast cancer, who were breastfeeding or who had breast implants were not included. Among the 156 patients retrospectively identified, 139 women had been enrolled into a prospective trial called "Neoelasto" (NCT02834494) investigating the use of shear wave elastography, with MRI as a reference modality, to predict and analyse the results of NAC in breast cancer and for which they gave written informed consent. The requirement to obtain informed consent was waived for the other patients. Discarding patients with missing modalities or clinical data or whose images could not be used due to technical failures, 136 patients were finally considered (Figure 3.1).

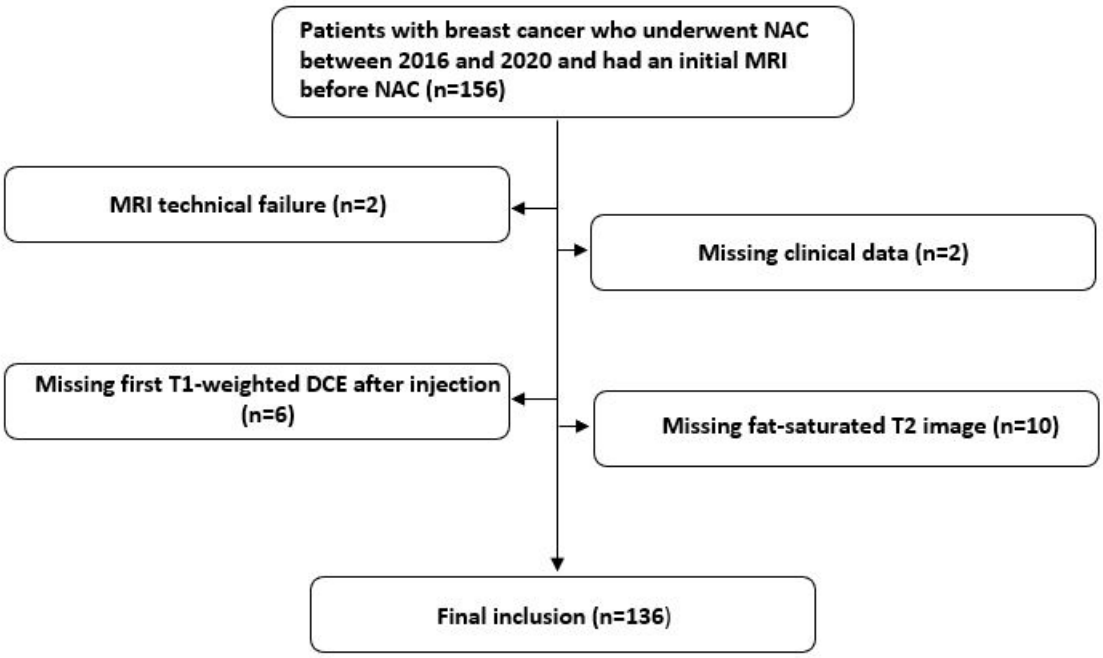


Figure 3.1: Inclusion flowchart.

3.1.2 Imaging protocol

All patients were imaged in the prone position in the axial plane with dedicated breast coils. A majority of patients (n=110) were scanned in one of the three imaging settings of Institut Curie: 28 patients were scanned on a GE Optima MR450w device with an 8-channel coil (coil 1), 19 on a MAGNETON Aera scanner using a 18-channel breast coil (coil 2) and 63 on the previously mentioned MAGNETON Aera scanner but with a Sentinelle 16-channel breast coil dedicated to biopsies (coil 3). Patients were imaged with the standard routine protocol of Institut Curie, including 3D T1-weighted DCE and fat-saturated T2 sequences. To perform dynamic contrast-enhanced imaging, an intravenous injection of a gadolinium-based contrast agent (gadoterate meglumine, commercialized under the name of Dotarem by Guerbet Healthcare) at a concentration of 0.2 mL per kilogram of body weight, was carried on using a power injector, followed by a 20 mL saline solution flush. Four to five images were acquired every 90s after injection. The sequence parameters of the three imaging settings are reported in Table 3.1.

The remaining patients (n=26) were scanned in a multitude of other centers with different scanners and coils. Slice thickness of images acquired in these centers ranged from 0.7 to 2.2 mm in T1-DCE images with a mean of 1.6 mm and 1.5 to 5 mm in T2 images with a mean of 3.4 mm. Images were reviewed to control quality by Dr. Malhaire.

Patients were divided into a training set, gathering the first 103 women imaged at Institut Curie, and a test set with all the patients imaged in other centers and 7 patients imaged at Institut Curie and included at a later stage in the study. Table 3.2 summarizes the scanning devices used in both sets.

Table 3.1: Scanning parameters of routine sequences of imaging devices of Institut Curie.

	T1			fat-saturated T2			T1-weighted DCE		
Coil	Coil 1	Coil 2	Coil 3	Coil 1	Coil 2	Coil 3	Coil 1	Coil 2	Coil 3
TR (ms)	6.9	592	545	5544	3310	6400	6.81	5.2	5.2
TE (ms)	4.2	13	13	90	88	88	3.3	2.4	2.4
Slice thickness (mm)	1.6	3.5	3.0	3.0	3.5	3.0	1.0	0.9	0.9
Spacing between slices (mm)	0.8	4.2	3.6	3.3	4.2	3.6	1.0	0.9	0.9
Pixel spacing (mm)	0.68x0.68	0.71x0.71	0.68x0.68	0.70x0.70	0.70x0.70	0.70x0.70	0.82x0.82	0.91x0.91	0.91x0.91
Pixel bandwidth (Hz/pixel)	244	130	130	558	315	375	434	355	355
Flip angle	20	148	148	160	150	180	15	10	10

***Coil 1**: Optima MR450w with 8-channel coil; **Coil 2**: Magneton Aera with 18-channel breast coil; **Coil 3**: Magneton Aera with Sentinelle breast coil.*

3.1.3 Patient & tumor characteristics

Clinical & biological data were collected from patient clinical records including the age, BMI, menopausal status and stage according to TNM staging. From biopsies performed before the beginning of NAC, histological type, tumor grade, Ki67 and TILs levels were assessed. TILs levels were binarized into "High" and "Low" classes using 30% as a threshold level [103]. Tumors were considered to be ER or PR positive when at least 10% of the cells collected for immunohistochemistry testing were stained, indicating respectively the presence of ER or PR receptors. HER2-overexpression was assessed with immunohistochemistry combined when necessary with FISH. Molecular subtypes can be described in different manners. Luminal B tumors with HER2 positive status are sometimes gathered with HER2-enriched tumors (Chapter 1) while on the contrary the Luminal B class can be sometimes broken down in two subgroups depending on the HER2 status, which was the case in this thesis. Tumor response was established at Institut Curie on the post-surgical specimens using the Residual Cancer Burden (RCB) score (Chapter 1). Grading systems to assess tumor response can vary between centers. Patients with RCB class I (minimal residual disease), class II (moderate residual disease) and class III (extensive residual disease) were considered non pCR (npCR) as opposed to patients with RCB class 0 who achieved pCR.

All MR images and associated BI-RADS descriptors were provided by Dr. Caroline Malhaire, with the assistance of Dr. Fatine Selhane, resident radiologist with one year of experience in breast MRI. Kinetic parameters were described using a time-intensity curve calculated in the most enhanced area within the tumor. In addition to BI-RADS, a breast edema score (BES), as defined by Harada et al. [102] on T2-weighted images, was recorded. Harada et al. found that BES was associated with the prognosis of breast cancer patients after NAC. BES was evaluated using five levels: BES1 indicates the lack of edema, BES2 the presence of a peritumoral edema, BES3 the presence of a prepectoral edema, BES4 the presence of a subcutaneous edema and

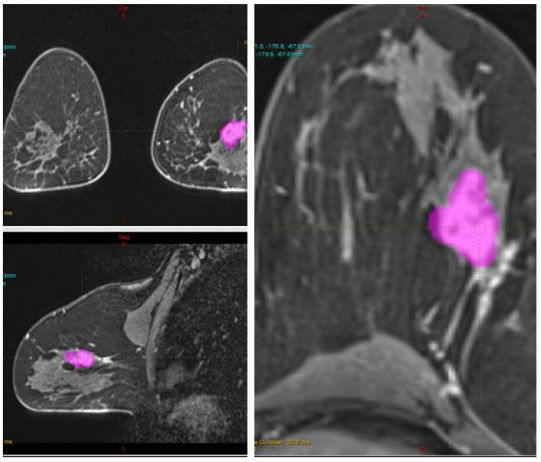
Table 3.2: Imaging devices of training and test sets.

Imaging centers	Manufacturers	Devices	Magnetic field strength (T)	Coils	Training	Testing
Institut Curie	GE	Optima MR450w	1.5	8-channel coil	25	3
Institut Curie	Siemens	MAGNETOM Aera	1.5	18-channel coil	19	0
Institut Curie	Siemens	MAGNETOM Aera	1.5	Sentinelle (16-channel) coil	59	4
Other center	Siemens	MAGNETOM Aera	1.5	16-channel coil	0	4
Other center	Siemens	MAGNETOM Aera	1.5	18-channel coil	0	3
Other center	Siemens	MAGNETOM Aera	1.5	Spine 32-channel coil	0	1
Other center	Siemens	MAGNETOM Amira	1.5	18-channel coil	0	1
Other center	Siemens	MAGNETOM Avanto eco	1.5	Breast matrix coil	0	1
Other center	Siemens	MAGNETOM Avanto eco	1.5	16-Channel AI Breast coil	0	1
Other center	Siemens	MAGNETOM ESSENZA	1.5	Breast matrix coil	0	1
Other center	GE	Discovery MR 750	3	HD Breast coil	0	1
Other center	GE	Optima MR360	1.5	HD Breast coil	0	4
Other center	GE	Optima MR450w	1.5	HD Breast coil	0	2
Other center	GE	Signa Artist	1.5	HD Breast coil	0	3
Other center	GE	Signa HDxt	1.5	HD Breast coil	0	2
Other center	GE	Signa Voyager	1.5	HD Breast coil	0	2

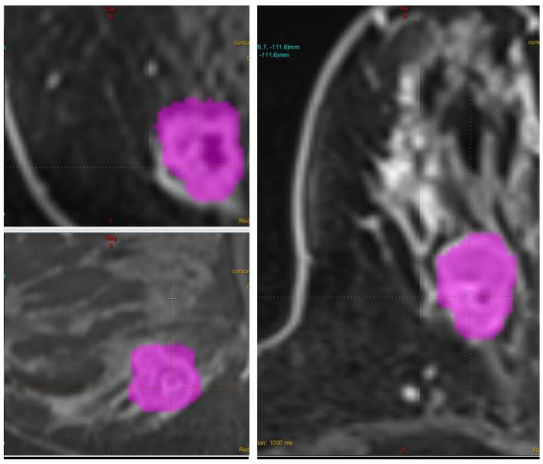
***Training**: Number of training patients; **Testing**: Number of test patients; **Other centers**: imaging centers not associated with Institut Curie.*

BES5 the presence of an edema noticeable during clinical exams without MR imaging. This score was included in the radiologists' report. When there were several malignant masses within the ipsilateral breast, tumors were described as multifocal. The presence of non-mass enhancement associated with the index lesion (the most extensive lesion) was also reported. The index lesion was measured and so was the combination of masses and potential non-mass enhancement under the name of "Maximal size of lesion". Measurements were carried out along the longest axis in one of the three planes (axial, coronal and sagittal) on the first subtracted image of the DCE sequence, obtained by subtracting the pre-contrast image to the first post-contrast image.

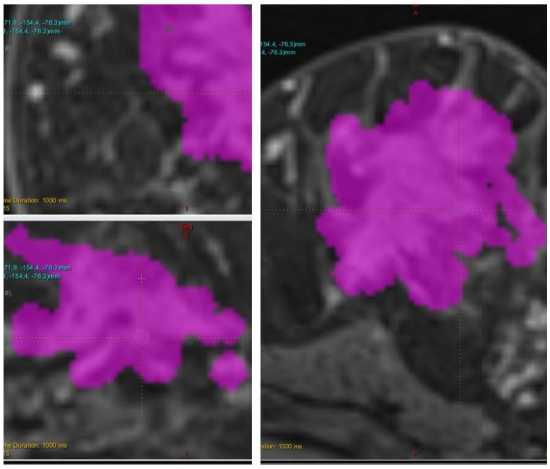
Tables 3.3 and 3.4 respectively summarize clinical & biological data and BI-RADS and other imaging descriptors in both training and test sets. Figure 3.2 illustrates the differences between the shape and margins descriptors as they both have an "irregular" class that can be confusing.



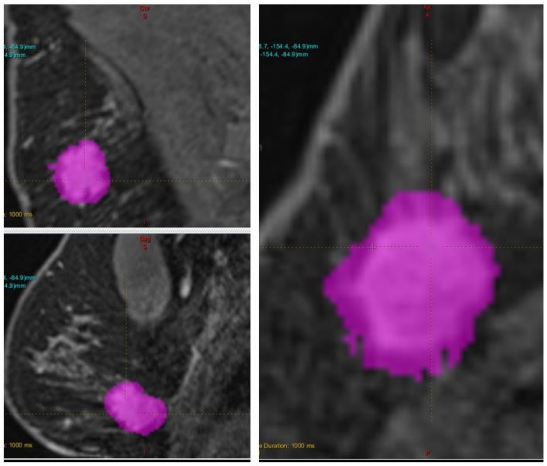
(a) Circumscribed/irregular margins & irregular shape



(b) Circumscribed/irregular margins & oval/round shape



(c) Spiculated margins & irregular shape



(d) Spiculated margins & oval/round shape

Figure 3.2: Illustration of shape and margins features in T1-weighted DCE images.

According to statistical tests, there were globally few differences between the training and test sets with the exception of the delayed phase enhancement ($p = 0.001$) and the background parenchymal enhancement ($p = 0.017$). T stage parameter was at the statistical limit ($p = 0.059$). The test set size is however limited and should temper these differences.

Table 3.3: Clinical & biological characteristics of training and test patients.

Label	Levels	Training	Testing	Total	p
Age (y)	Median (IQR)	48.0 (39.5 to 56.5)	46.0 (39.0 to 52.0)	47.5 (39.0 to 56.2)	0.594
BMI (kg.m^{-2})	Median (IQR)	23.4 (21.4 to 25.7)	23.4 (21.5 to 27.5)	23.4 (21.5 to 26.1)	0.359
Menopause	Postmenopausal	42 (40.8)	11 (33.3)	53 (39.0)	0.577
	Premenopausal	61 (59.2)	22 (66.7)	83 (61.0)	
T stage	0/I/II	91 (88.3)	24 (72.7)	115 (84.6)	0.059
	III/IV	12 (11.7)	9 (27.3)	21 (15.4)	
N stage	0	58 (56.3)	16 (48.5)	74 (54.4)	0.559
	I/II	45 (43.7)	17 (51.5)	62 (45.6)	
M stage	0	102 (99.0)	33 (100.0)	135 (99.3)	1.000
	I	1 (1.0)	0 (0.0)	1 (0.7)	
Histological Type	Ductal NOS	99 (96.1)	32 (97.0)	131 (96.3)	0.757
	Lobular	1 (1.0)	0 (0.0)	1 (0.7)	
	Mixt	2 (1.9)	0 (0.0)	2 (1.5)	
	Other	1 (1.0)	1 (3.0)	2 (1.5)	
Molecular subtype	HER2+	12 (11.7)	7 (21.2)	19 (14.0)	0.567
	Luminal B/HER2-	30 (29.1)	9 (27.3)	39 (28.7)	
	Luminal B/HER2+	13 (12.6)	3 (9.1)	16 (11.8)	
	TN	48 (46.6)	14 (42.4)	62 (45.6)	
Grade	2	34 (33.0)	8 (24.2)	42 (30.9)	0.464
	3	69 (67.0)	25 (75.8)	94 (69.1)	
Ki67 (%)	Median (IQR)	60.0 (32.5 to 75.0)	40.0 (30.0 to 60.0)	50.0 (30.0 to 75.0)	0.231
TILFactor	High	44 (42.7)	18 (54.5)	62 (45.6)	0.324
	Low	59 (57.3)	15 (45.5)	74 (54.4)	
Response	npCR	54 (52.4)	18 (54.5)	72 (52.9)	0.991
	pCR	49 (47.6)	15 (45.5)	64 (47.1)	

Continuous variables are represented by their median and IQR . Wilcoxon rank sum test and Pearson's Chi-square test were performed respectively for continuous and categorical variables between training and test sets. In circumstances where Chi-square test could not be used due to too few observations, Fisher's exact test was carried out.

Table 3.4: BI-RADS classification for training and test patients.

Label	Levels	Training	Testing	Total	p
Multifocality	No	76 (73.8)	26 (78.8)	102 (75.0)	0.729
	Yes	27 (26.2)	7 (21.2)	34 (25.0)	
Depth location	Anterior third	7 (6.8)	0 (0.0)	3 (2.2)	0.740
	Middle third	43 (41.7)	18 (54.5)	61 (44.9)	
	Posterior third	53 (51.5)	15 (45.5)	68 (50.0)	
Breast composition	A	3 (2.9)	0 (0.0)	3 (2.2)	0.740
	B	43 (41.7)	17 (51.5)	60 (44.1)	
	C	40 (38.8)	12 (36.4)	52 (38.2)	
	D	17 (16.5)	4 (12.1)	21 (15.4)	
Margins	Circumscribed/Irregular	55 (53.4)	22 (66.7)	77 (56.6)	0.256
	Spiculated	48 (46.6)	11 (33.3)	59 (43.4)	
Shape	Irregular	76 (73.8)	28 (84.8)	104 (76.5)	0.256
	Oval/Round	27 (26.2)	5 (15.2)	32 (23.5)	
Background parenchymal enhancement (BPE)	Marked	61 (59.2)	11 (33.3)	72 (52.9)	**0.017**
	Minimal	42 (40.8)	22 (66.7)	64 (47.1)	
Central Necrosis T2	No	65 (63.1)	22 (66.7)	87 (64.0)	0.871
	Yes	38 (36.9)	11 (33.3)	49 (36.0)	
Associated non-mass enhancement	Absent	82 (79.6)	27 (81.8)	109 (80.1)	0.979
	Present	21 (20.4)	6 (18.2)	27 (19.9)	
Peritumoral edema T2	No	31 (30.1)	9 (27.3)	40 (29.4)	0.928
	Yes	72 (69.9)	24 (72.7)	96 (70.6)	
Prepectoral edema T2	No	61 (59.2)	24 (72.7)	85 (62.5)	0.235
	Yes	42 (40.8)	9 (27.3)	51 (37.5)	
Subcutaneous edema T2	No	86 (83.5)	29 (87.9)	115 (84.6)	0.782
	Yes	17 (16.5)	4 (12.1)	21 (15.4)	
Breast edema score (BES)	1	27 (26.2)	8 (24.2)	35 (25.7)	0.402
	2	28 (27.2)	14 (42.4)	42 (30.9)	
	3	31 (30.1)	7 (21.2)	38 (27.9)	
	4	17 (16.5)	4 (12.1)	21 (15.4)	
Internal enhancement type	Heterogeneous	45 (43.7)	16 (48.5)	61 (44.9)	0.732
	Homogeneous	29 (28.2)	7 (21.2)	36 (26.5)	
	Rim enhancement	29 (28.2)	10 (30.3)	39 (28.7)	
Kinetic curve initial enhancement	Fast	98 (95.1)	30 (90.9)	128 (94.1)	0.401
	Slow/Intermediate	5 (4.9)	3 (9.1)	8 (5.9)	
Delayed phase enhancement	Persistent/Plateau	23 (22.3)	18 (54.5)	41 (30.1)	**0.001**
	Wash-out	80 (77.7)	15 (45.5)	95 (69.9)	
Combined size (mm)	Median (IQR)	28.0 (23.0 to 40.0)	32.0 (20.0 to 40.0)	30.0 (22.0 to 40.0)	0.841
Maximal size of lesion (mm)	Median (IQR)	25.0 (21.5 to 33.5)	30.0 (19.0 to 37.0)	25.0 (21.0 to 35.0)	0.984

Continuous variables are represented by their median and IQR . Wilcoxon rank sum test and Pearson's Chi-square test were performed respectively for continuous and categorical variables between training and test sets. In circumstances where Chi-square test could not be used due to too few observations, Fisher's exact test was carried out.

3.2 Association of clinical, biological and MRI features with pCR

3.2.1 Methods

The potential association of the previously defined clinical, biological and imaging features with the response to NAC was tested on the training set, to be used later on for the feature selection process. The p-values were calculated with Wilcoxon rank sum test, Pearson's Chi-square test or Fisher's exact test depending on the nature and size of the data. The p-values were corrected for multiple comparisons with the Benjamini and Hochberg (BH) [215] method using clinical, biological and imaging features. A significance level of 0.05 was chosen.

3.2.2 Results

Statistical analyses confirmed results found in the literature: molecular subtype, Ki67 and TILs were significantly associated with pCR (before correction) and close to the significance level after BH correction (Table 3.5). Amongst BI-RADS and other imaging features, the margins were found to have a significant association with the response before correction whereas the multifocality parameter was at the limit of significance ($p = 0.051$) (Table 3.6).

Table 3.5: Association of clinical & biological data with pCR on the training set.

Label	Levels	npCR	pCR	Total	p	Corrected p
Age	Median (IQR)	49.0 (43 to 59.8)	47.0 (38.0 to 53.0)	48 (39.5 to 56.5)	0.160	0.480
BMI	Median (IQR)	23.6 (21.5 to 26.0)	23.4 (21.3 to 24.8)	23.4 (21.4 to 25.7)	0.616	0.978
Menopause	Postmenopausal	25 (46.3)	17 (34.7)	42 (40.8)	0.5316	0.702
	Premenopausal	29 (53.7)	32 (65.3)	61 (59.2)		
T stage	0/I/II	47 (87.0)	44 (89.8)	91 (88.3)	0.764	1.000
	III/IV	7 (13.0)	5 (10.2)	12 (11.7)		
NStage	0	30 (55.6)	28 (57.1)	58 (56.3)	1.000	1.000
	I/II	24 (44.4)	21 (42.9)	45 (43.7)		
MStage	0	54 (100.0)	48 (98.0)	102 (99.0)	0.476	0.918
	I	0 (0.0)	1 (2.0)	1 (1.0)		
Histological Type	Ductal NOS	51 (94.4)	48 (98.0)	99 (96.1)	1.000	1.000
	Lobular	1 (1.9)	0 (0.0)	1 (1.0)		
	Mixt	1 (1.9)	1 (2.0)	2 (1.9)		
	Other	1 (1.9)	0 (0.0)	1 (1.0)		
Molecular subtype	HER2+	3 (5.6)	9 (18.4)	12 (11.7)	**0.006**	0.054
	Luminal B/HER2-	23 (42.6)	7 (14.3)	30 (29.1)		
	Luminal B/HER2+	7 (13.0)	6 (12.2)	13 (12.6)		
	TN	21 (38.9)	27 (55.1)	48 (46.6)		
Grade	2	21 (38.9)	13 (26.5)	34 (33.0)	0.212	0.572
	3	33 (61.1)	36 (73.5)	69 (67.0)		
Ki67	Median (IQR)	50.0 (30.0 to 70.0)	70.0 (40.0 to 80.0)	60.0 (32.5 to 75.0)	**0.005**	0.054
TILFactor	High	16 (29.6)	28 (57.1)	44 (42.7)	**0.006**	0.054
	Low	38 (70.4)	21 (42.9)	59 (57.3)		

Continuous variables are represented by their median and IQR.

Table 3.6: Association of BI-RADS features with pCR on the training set.

Label	Levels	Training	Testing	Total	p	Corrected p
Multifocality	No	35 (64.8)	41 (83.7)	76 (73.8)	0.051	0.275
	Yes	19 (35.2)	8 (16.3)	27 (26.2)		
Breast composition	A	1 (1.9)	2 (4.1)	3 (2.9)	0.842	1
	B	24 (44.4)	19 (38.8)	43 (41.7)		
	C	21 (38.9)	19 (38.8)	40 (38.8)		
	D	8 (14.8)	9 (18.4)	17 (16.5)		
Margins	Circumscribed/Irregular	22 (40.7)	33 (67.3)	55 (53.4)	**0.012**	0.081
	Spiculated	32 (59.3)	16 (32.7)	48 (46.6)		
Shape	Irregular	44 (81.5)	32 (65.3)	76 (73.8)	0.101	0.390
	Oval/Round	10 (18.5)	17 (34.7)	27 (26.2)		
Background parenchymal enhancement (BPE)	Marked	30 (55.6)	31 (63.3)	61 (59.2)	0.552	0.978
	Minimal	24 (44.4)	18 (36.7)	42 (40.8)		
Central Necrosis T2	No	33 (61.1)	32 (65.3)	65 (63.1)	0.813	1.000
	Yes	21 (8.9)	17 (34.7)	38 (36.9)		
Associated non-mass enhancement	Absent	39 (72.2)	43 (87.8)	82 (79.6)	0.087	0.390
	Present	15 (27.8)	6 (12.2)	21 (20.4)		
Peritumoral edema T2	No	18 (33.3)	13 (26.5)	31 (30.1)	0.592	0.978
	Yes	36 (66.7)	36 (73.5)	72 (69.9)		
Prepectoral edema T2	No	36 (66.7)	25 (51.0)	6 (59.2)	0.158	0.480
	Yes	18 (33.3)	24 (49.0)	42 (40.8)		
Subcutaneous edema T2	No	44 (81.5)	42 (85.7)	86 (83.5)	0.755	1.000
	Yes	10 (18.5)	7 (14.3)	17 (16.5)		
Breast edema score (BES)	1	16 (29.6)	11 (22.4)	27 (26.2)	0.338	0.702
	2	16 (29.6)	12 (24.5)	28 (27.2)		
	3	12 (22.2)	19 (38.8)	31 (30.1)		
	4	10 (18.5)	7 (14.3)	17 (16.5)		
Internal enhancement type	Heterogeneous	24 (44.4)	21 (42.9)	45 (43.7)	0.859	1.000
	Homogeneous	14 (25.9)	15 (30.6)	29 (28.2)		
	Rim enhancement	16 (29.6)	13 (26.5)	29 (28.2)		
Kinetic curve initial enhancement	Fast	51 (94.4)	47 (95.9)	98 (95.1)	1.000	1.000
	Slow/Intermediate	3 (5.6)	2 (4.1)	5 (4.9)		
Delayed phase enhancement	Persistent/Plateau	12 (22.2)	11 (22.4)	23 (22.3)	1.000	1.000
	Wash-out	42 (77.8)	38 (77.6)	80 (77.7)		
Combined size (mm)	Median (IQR)	30.5 (24.0 to 41.0)	27.0 (22.0 to 35.0)	28.0 (23.0 to 40.0)	0.268	0.658
Maximal size of lesion (mm)	Median (IQR)	25.0 (21.2 to 33.0)	25.0 (22.0 to 34.0)	25.0 (21.5 to 33.5)	0.9992	1.000

Continuous variables are represented by their median and IQR.

3.3 Clinical, biological and BI-RADS feature-based predictive models

3.3.1 Introduction

Several studies proposed predictive models using only clinical & biological data or BI-RADS descriptors without advanced texture or shape features requiring a segmentation of the tumors [18, 134, 193]. Amongst them, the study by Granzier et al. [18] stood out as they built two different models, performed their analysis before the beginning of NAC and disclosed precisely the composition of the database and the hyperparameters and features of the clinical & biological models which were all included in our own database. No weights were however provided. Granzier et al. [18] had two independent multi-scanner cohorts acquired in two different hospitals. They developed two different models, each trained on one cohort and tested on the other one. The first model (that we refer as "Granzier") achieved an AUC of 0.71, 95% confidence interval (CI) [0.62, 0.79] on the first test cohort of 152 lesions (129 patients) while the second model "Granzier2" achieved an AUC of 0.77, 95% CI [0.70, 0.85] when tested on the second cohort with 168 tumors (161 patients).

Granzier et al. [18] predicted the response using random forest models trained on a subset of features selected by the Boruta algorithm (described in Chapter 2). In the study, the collected clinical & biological data, on which to perform feature selection, included the age

of patients, the TNM staging, the grade, the histological type and the molecular subtype of tumors. Response was assessed using the Miller-Payne system which differs from the RCB score we used as it focuses on changes in cellularity of the primary tumor whereas RCB also takes into account lymph node metastases [216]. No information on TILs and Ki67 were included and no BI-RADS or other imaging descriptors were defined. However, as all the features they used were included in our database, it was possible to build a model based on these features on our training set and to test it on our test set. We then developed and tested our own models first using only clinical & biological data or only BI-RADS and imaging features and then combining the features.

3.3.2 Methods

The molecular subtype feature was divided into four different classes (HER2-enriched, Luminal B/HER2-, Luminal B/HER2+ and TN) as it was done in [18].

We then trained the two models proposed in [18] and developed new models. Several selection methods (a simple filter method, MRMR , Boruta and recursive feature elimination (RFE)) and types of classifiers (logistic regression, SVM with linear or radial kernels, random forests) were compared. The filter method consisted of a simple threshold cut-off on the Wilcoxon rank sum test p-values (<0.05). RFE is an iterative wrapper method that can be associated with different types of predictive models. RFE starts will all the available features in the dataset and at each iteration, fits a model, ranks features according to their importance and discards the least relevant ones. This process is repeated until a desired number of features, which is an hyperparameter of the model, is reached. The method used by Granzier et al. in their study (selection by Boruta and random forest model) was specifically included in our experiments. Feature selection was performed on the training set only.

As it gathers 49 patients that achieved pCR and 54 that did not, the maximal number of features to build models was set to five to avoid overfitting, with the idea of having, as a rule of thumb, one feature for every ten pCR patients. Models could nevertheless have less than five features as some selection methods found an optimum set of features on the training set containing fewer features. In case the Boruta algorithm selected more than five features, the five best features according to their importance calculated by the algorithm were kept.

As the training and test sets were evenly balanced (49 pCR/54 npCR vs 15 pCR/18 npCR), no particular resampling of the two classes was performed. Features were standardized and models were trained and tuned with repeated cross-validation and evaluated on the independent test set using AUC metric with 95% confidence interval. Models were first trained with only clinical & biological data or only BI-RADS and imaging features, and then with both types of features to ponder how they complement one another.

To complete our experiments and evaluate the robustness of the models, our original training and test sets were gathered and reshuffled to create a new training and test sets with a 75% split (102/34 patients with 48 pCR/ 54 npCR in the training and 16 pCR/ 18 npCR in the test sets). Granzier's models and the best models we developed previously were trained and tested on these new datasets.

3.3.3 Results

Tables 3.7, 3.8 and 3.9 report the features selected by each method when using clinical & biological data, imaging features or both of them. As RFE requires a model to perform feature selection, different subsets of features were thus selected depending on models. When using clinical & biological data (Tables 3.8 & 3.9), all methods selected the molecular subtype feature and the TILs factor. When using imaging descriptors, the margins feature was always selected.

Figure 3.3 depicts the ROC-curves for all types of models and selection methods on the original training and test sets when using only clinical & biological data. Figure 3.4 shows the ROC-curves of the Granzier models while Figure 3.5 displays ROC-curves when using both clinical & biological and other imaging features. The different models trained using only BI-RADS features achieved poor performances on the training set (AUCs close to 0.5) and the idea of using only BI-RADS and imaging features was thus discarded.

Best performances on the original training set were obtained with the random forest model trained on clinical & biological and BI-RADS features selected by the Boruta algorithm with AUC=0.75, 95% CI [0.65-0.85] (Figure 3.5g). This model also achieved the best results on the test set amongst all the experiments (AUC=0.76, 95% CI [0.59-0.93] (Figure 3.5h). The best model using only clinical & biological data was the logistic regression trained on features filtered by the threshold cut-off (AUC=0.73, [0.63-0.83] on the training set in Figure 3.3a, AUC=0.71, [0.52-0.80] on the test set in Figure 3.3b). There was no significant difference between the AUC obtained on the test set by these two models using Delong tests [217]. Table 3.10 compares the performances of the best model obtained in the experiments and the two Granzier models trained on our data.

Figure 3.6 depicts the ROC-curves obtained on the reshuffled data, highlighting a drop in performances of the Granzier's models on the test set. We also retrained to avoid data leakage on the reshuffled sets the two pipelines that gave the best results: filters with logistic regression or Boruta selection with random forest classifier. The filter selection proved very robust, selecting again the molecular subtype, TILs and Ki67 and the logistic regression achieved on the test set an AUC=0.70, [0.51-0.89] (Figure 3.6f). There was a slight difference in features selected by the Boruta algorithm with the presence of non-mass selected instead of the shape parameter. Random forest model achieved an AUC=0.72, [0.54-0.92] on the reshuffled test set (Figure 3.6f).

Table 3.7: Feature selection using only BI-RADS and imaging features.

Selection method	Selected features
Filter method	Margins
mRMR	Margins, multifocality, shape, BPE, BES
RFE-logistic regression	Margins, multifocality, shape
RFE-linear SVM	Margins, multifocality
RFE-radial SVM	Margins, multifocality, shape, BES
RFE-random forest	Margins, shape
Boruta	Margins, shape

Table 3.8: Feature selection using only clinical & biological data.

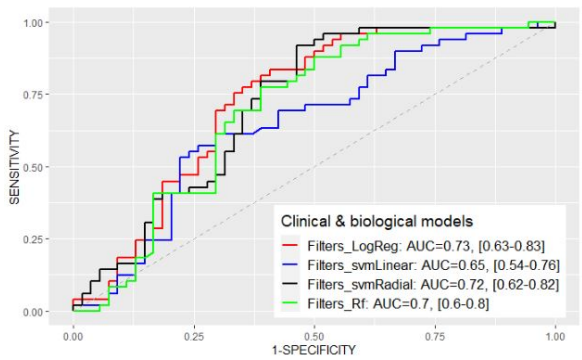
Selection method	**Selected features**
Filter method	Molecular subtype, TILs, Ki67, margins
mRMR	Molecular subtype, TILs, Ki67, age, N stage
RFE-logistic regression	Molecular subtype, TILs, Ki67, age, menopause
RFE-linear SVM	Molecular subtype, TILs, Ki67, age, menopause
RFE-radial SVM	Molecular subtype, TILs, Ki67
RFE-random forest	Molecular subtype, TILs, Ki67
Boruta	Molecular subtype, TILs, Ki67, grade
Literature	**Reported**
Granzier	Molecular subtype, age, T stage, grade
Granzier2	Molecular subtype, T stage, N stage, grade

Table 3.9: Feature selection using both clinical & biological and imaging features.

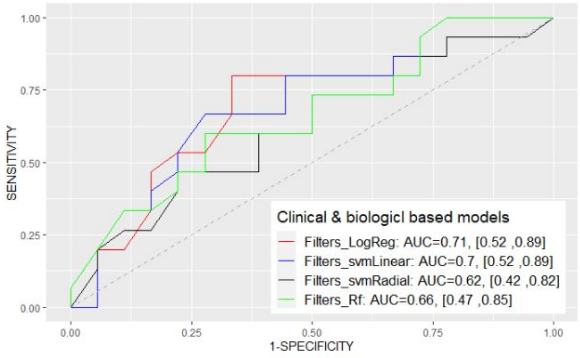
Selection method	Selected features
Filter method	Molecular subtype, TILs, Ki67
mRMR	Molecular subtype, TILs, margins, age, shape
RFE-logistic regression	Molecular subtype, TILs, Ki67, margins
RFE-linear SVM	Molecular subtype, TILs, Ki67, margins
RFE-radial SVM	Molecular subtype, TILs, Ki67, margins, maximal size
RFE-random forest	Molecular subtype, TILs, margins, BES, index lesion size
Boruta	Molecular subtype, TILs, margins, shape

Table 3.10: Performances summary.

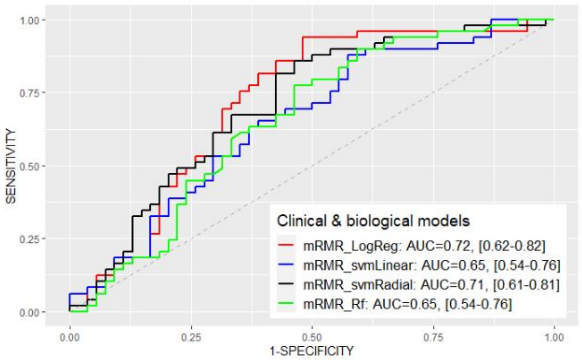
Model	Features	AUC [95%CI] (training)	AUC [95%CI] (test)
Best model built (RF with Boruta selection)	Molecular subtype, TILs, margins, shape	0.75 [0.65, 0.85]	0.76 [0.59, 0.93]
Granzier logistic regression	Molecular subtype, age, T stage, grade	0.65 [0.54, 0.76]	0.75 [0.64,0.90]
Granzier2 logistic regression	Molecular subtype, T stage, N stage, grade	0.61 [0.50, 0.72]	0.74 [0.57,0.92]



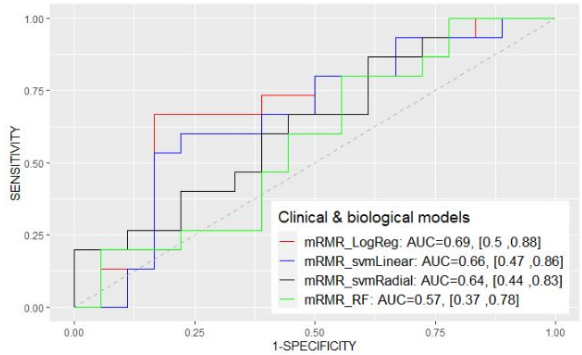
(a) Filter methods: Training



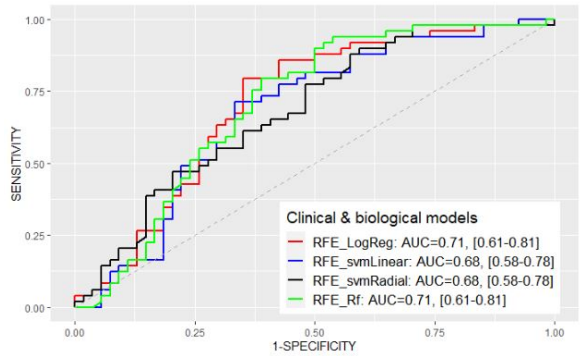
(b) Filter methods: Testing



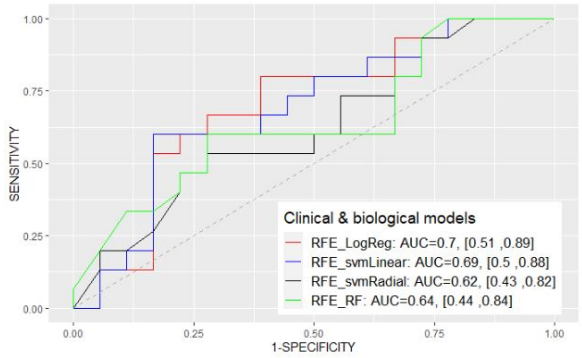
(c) mRMR: Training



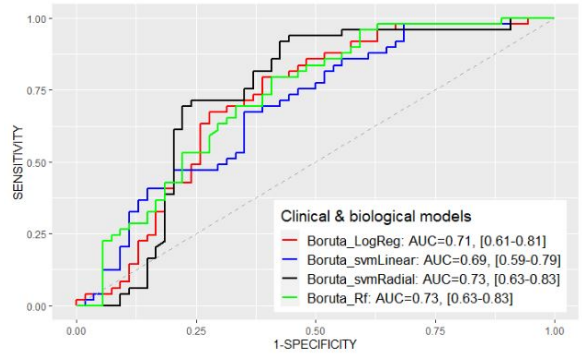
(d) mRMR: Testing



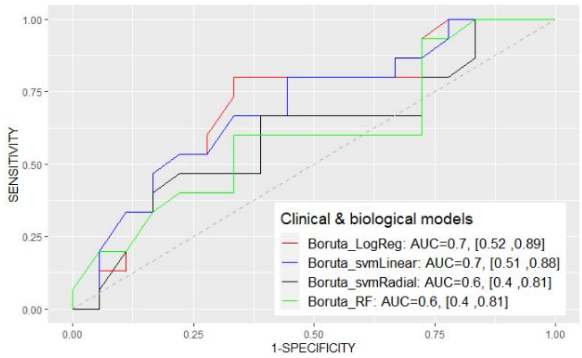
(e) RFE: Training



(f) RFE: Testing

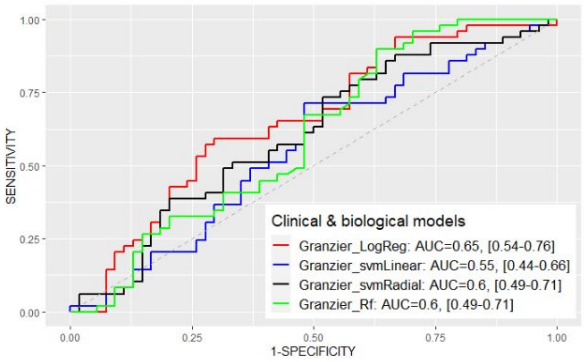


(g) Boruta: Training

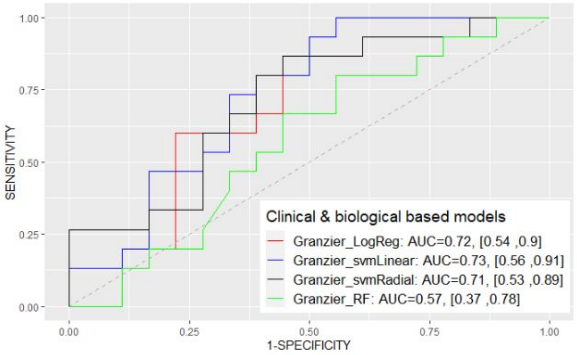


(h) Boruta: Testing

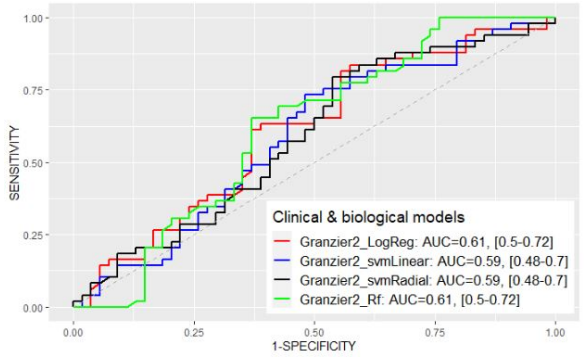
Figure 3.3: ***Clinical & biological models:*** *ROC-curves of the four types of models (**LogReg**: logistic regression; **svmLinear**: SVM with linear kernel; **svmRadial**: SVM with radial kernel and **RF**: random forest) obtained on the training set using to select features **(a)** a simple filter method (threshold cut-off); **(c)** the mRMR approach; **(e)** RFE; **(g)** Boruta. Their respective ROC-curves on the test set were illustrated in **(b)**, **(d)**, **(f)** and **(h)**.*



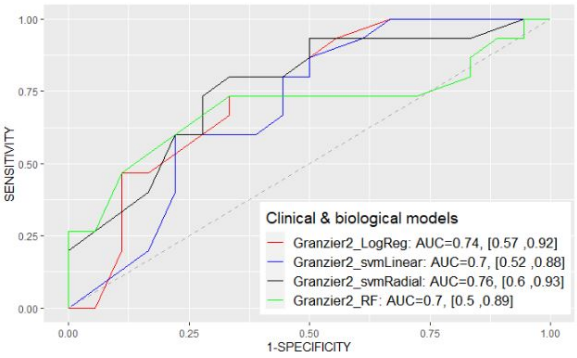
(a) Granzier: Training



(b) Granzier: Testing

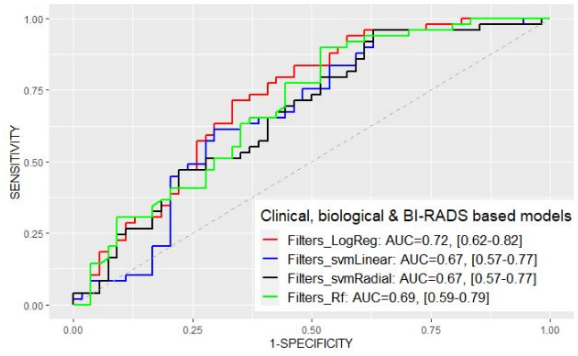


(c) Granzier2: Training

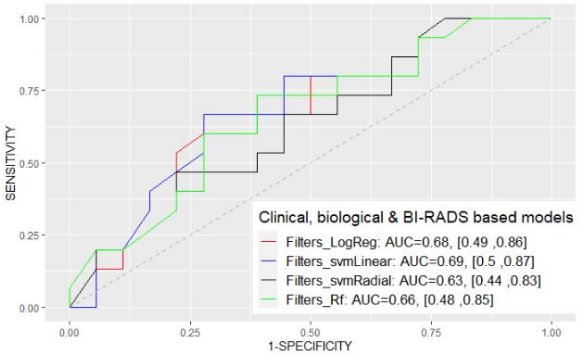


(d) Granzier2: Testing

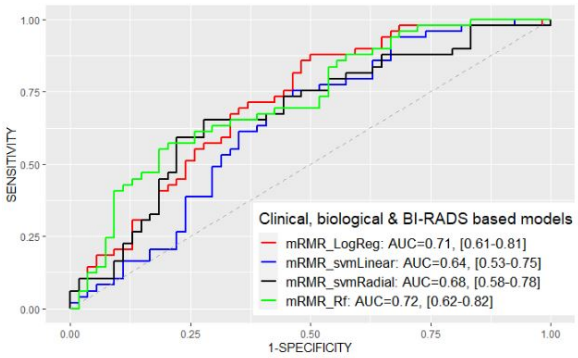
Figure 3.4: ***GRANZIER models:*** *ROC-curves of the four types of models (**LogReg**: logistic regression; **svmLinear**: SVM with linear kernel; **svmRadial**: SVM with radial kernel and **RF**: random forest) obtained on the training set using **(a)** the features of the "Granzier" model; **(c)** the features of the "Granzier2" model. Their respective ROC-curves on the test set were illustrated in **(b)** and **(d)**.*



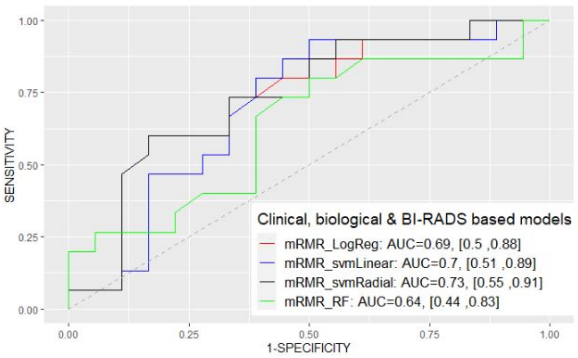
(a) Filter methods: Training



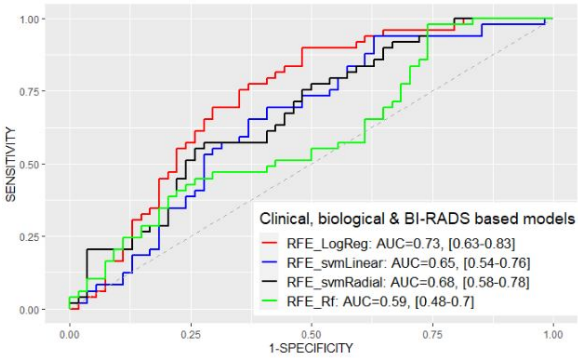
(b) Filter methods: Testing



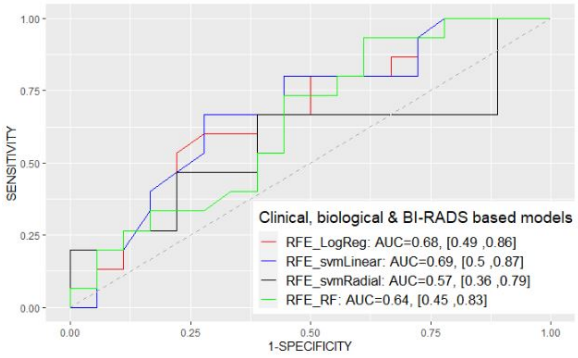
(c) mRMR: Training



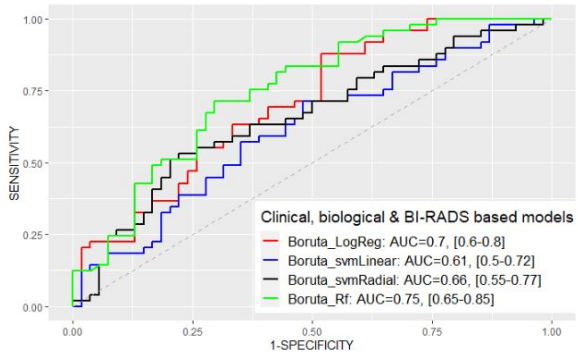
(d) mRMR: Testing



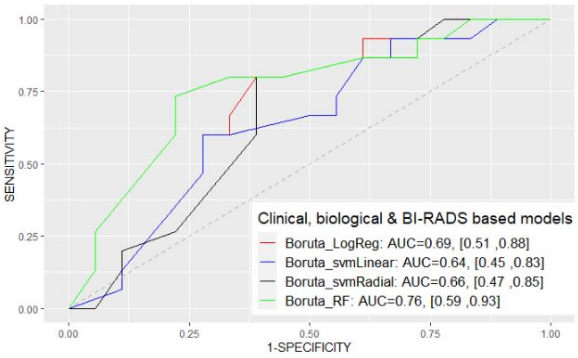
(e) RFE: Training



(f) RFE: Testing

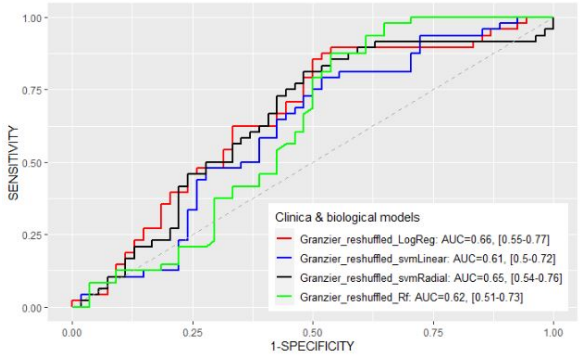


(g) Boruta: Training

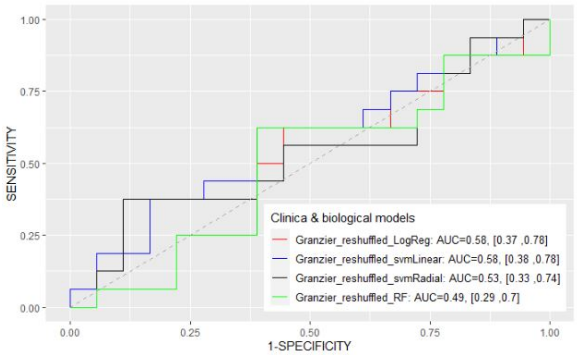


(h) Boruta: Testing

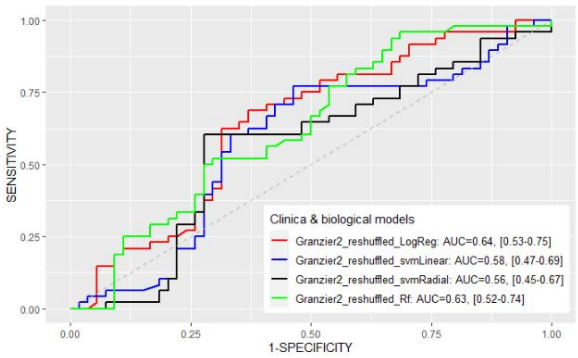
Figure 3.5: ***Clinical, biological & BI-RADS Models:*** *ROC-curves of the four types of models (**LogReg**: logistic regression; **svmLinear**: SVM with linear kernel; **svmRadial**: SVM with radial kernel and **RF**: random forest) obtained on the training set using to select features **(a)** a simple filter method (threshold cut-off); **(c)** the mRMR approach; **(e)** RFE; **(g)** Boruta. Their respective ROC-curves on the test set were illustrated in **(b)**,**(d)**, **(f)** and **(h)**.*



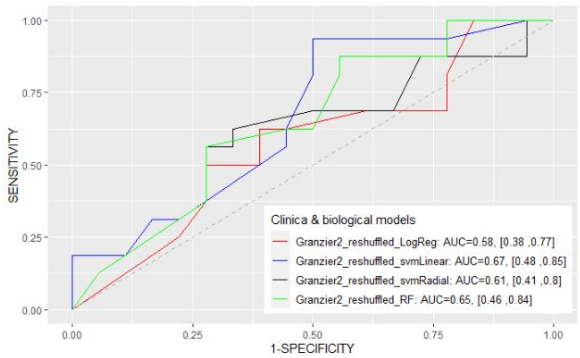
(a) Granzier: Training



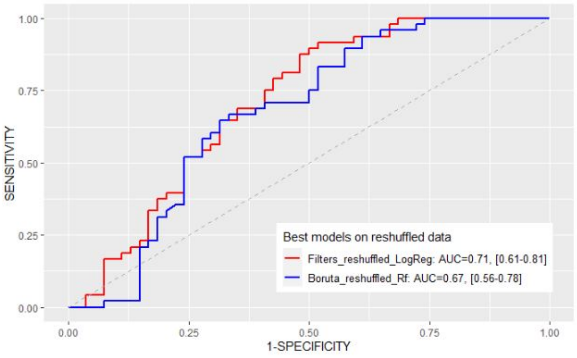
(b) Granzier: Testing



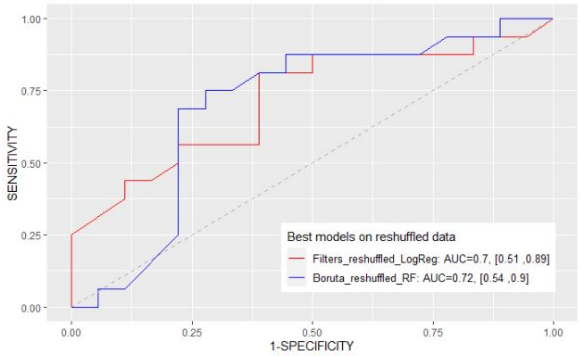
(c) Granzier2: Training



(d) Granzier2: Testing



(e) Best models: Training



(f) Best models: Testing

Figure 3.6: ***GRANZIER models and best models previously defined on reshuffled data:*** *ROC-curves of the four types of models (**LogReg**: logistic regression; **svmLinear**: SVM with linear kernel; **svmRadial**: SVM with radial kernel and **RF**: random forest) obtained on the reshuffled training set using **(a)** the features of the "Granzier" model; **(c)** the features of the "Granzier2" model; **(e)** the best models previously obtained. Their respective ROC-curves on the reshuffled test set were illustrated in **(b)**, **(d)** and **(f)**.*

3.3.4 Discussion

The two models developed by Granzier et al. [18] achieved globally average performances when evaluated on the original training set (best AUC=0.65, [0.54, 0.76]) but better performances on the original test set (best AUC 0.74, [0.57, 0.92]), equivalent to those of the best models built in our experiments and the performances reported in [18]. This trend goes unlike what is observed in the rest of the experiments where there is a slight drop (or sometimes a very slight increase) of the performances between training and testing. This could be explained by the presence of the T stage parameter in the models created by Granzier et al. In their cohorts, T stage and tumor grade are indeed strongly associated with the response ($p < 0.01$). In our database, T stage is not statistically associated with pCR on the training set but its distribution is very different on the test set (see Table 3.3) where it is strongly associated with pCR (Wilcoxon rank sum test p for the test set is $p < 0.005$). None of our feature selection methods selected the T stage parameter when performed on the training set. After reshuffling the data, a drop of performances of the Granzier models is observed. This suggests that the composition of the database and the distribution of features like the T stage parameter, can have an important impact on model performances and reduce their exportability.

Globally, clinical, biological and BI-RADS feature-based models achieve AUC performances in the range of [0.65-0.76] on our original test set. Best performances are thus in par with the results of the clinical models reported in the literature review of radiomic analyses. Using the Boruta algorithm to select features and then a random forest classifier, as did Granzier et al., achieved the best results though other models offered equivalent performances. There was a trend showing that adding BI-RADS features to clinical & biological data improved models' performances but further testing on larger cohorts are needed to confirm these results.

Conclusion

This chapter introduced the cohort that is going to be used in the rest of the manuscript and the inclusion process that took place. The training/testing split was presented and the different imaging acquisition parameters described, highlighting the heterogeneous and multicentric characteristics of the datasets.

A first set of features including clinical & biological data from patient records, BI-RADS and other simple imaging features visually assessed by radiologists without segmenting the lesions was gathered. Univariate analyses were performed to ponder the association of features with the response. First predictive models were then built, comparing several feature selection and model building approaches. Two models from the literature were tested to validate them. Best performances of clinical & biological and BI-RADS based models achieved AUCs as high as 0.75 on the independent test sets. Ultimately, the difficulty of exporting models on different databases was underscored.

Chapter 4

Phantom experiments

Preface

This chapter will present experiments and analyses conducted on breast phantoms scanned at Institut Curie using routine clinical protocols. These experiments aim to develop a correction pipeline to tackle inhomogeneities that affect MR images and radiomic features, with the hope to apply it in a later stage on patient images. Main results have been published in an article in *Magnetic Resonance Materials in Physics, Biology and Medicine* [19].

4.1 Introduction

The previous chapter described the inclusion and image acquisition process that was undertaken to gather the patient cohort and their images.

In view of carrying robust radiomic analyses, several characteristics of the dataset raise issues. The choice of the MR modality to monitor the response to treatment is well-adapted according to the literature [9, 10] but two main drawbacks inherent to MR imaging must be taken into account. First, MR signal is measured in arbitrary units which vary between acquisitions and device settings, unlike in CT or PET imaging respectively measured in Hounsfield units and kBq/mL. This makes comparisons between different acquisitions difficult, even between repeated acquisitions of the same patient using the same imaging set-up.

Besides, as previously mentioned in Chapter 2, MR images can be affected by the bias-field non-uniformity that creates local intensity inhomogeneities within tissues. Sources of the bias field have been linked to two main causes. First, effects associated with the properties of the MR devices including radio frequency emission and transmission inhomogeneities, edgy currents created by field gradients or static field inhomogeneities, have been pointed at. The shape, position and orientation of the scanned object within the magnet, its dielectric properties and magnetic permeability also have an impact [218, 219]. The first designated source of the bias field can be mitigated by improving the design of the coils, calibrating machines using dedicated phantoms or by shimming techniques, to make the main magnetic field more homogeneous [220]. Cohorts acquired prospectively can thus attempt to reduce bias field gain by carefully planning acquisitions on selected devices. This is however not possible in retrospective studies as patients were collected with this bias. Regardless, inhomogeneities due to the imaged object or patient always need to be corrected retrospectively using image

processing techniques. Numerous studies have focused on the correction of MR bias field gain but most of them were designed very specifically for brain datasets [218].

Furthermore, as the training and test set include patients imaged on different scanners, coils and imaging centers, the "scanner effect" affecting radiomic feature values must also be corrected. The correction of this effect has two main purposes: allowing the true biological effects that could be hidden by a misalignment of feature distributions to stand out and improving the exportability of decisions made on feature values like simple threshold levels or more advanced radiomic signatures.

As a majority of the patients (110/136) were acquired in one of the three MR imaging settings of Institut Curie, where experiments could be carried out, a phantom experiment was designed to study the bias field, inter-acquisition variabilities and the "scanner effect". Dedicated breast phantoms mimicking real breast tissue were chosen to get the closest idea as possible of the bias field gain impacting patient images. Phantoms are free of biological effects or artefacts due to tissue interference and their simple composition makes it easier to model what the ground truth image without bias field gain would look like. Imaging phantoms according to the clinical imaging routine of Institut Curie offered the opportunity to test and adapt pre-processing pipelines for radiomic analyses that have later been carried out on patient images (Chapters 5 and 6).

4.2 Article - Saint Martin et al., MAGMA, 2021

A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study.

PUBLISHED in *Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA)* [19].

Marie-Judith Saint Martin[1], Fanny Orlhac[1], Fahad Khalid[1], Pia Akl[2], Christophe Nioche[1], Irène Buvat[1], Caroline Malhaire[1,3] and Frédérique Frouin[1]

[1] U1288-LITO, Inserm, Centre de Recherche de l'Institut Curie, Université Paris-Saclay, Orsay, France
[2] Department of Radiology, Hôpital Femme Mère Enfant, Hospices civils de Lyon, Lyon, France
[3] Department of Radiology, Ensemble Hospitalier de l'Institut Curie, Paris, France

Abstract

Object: Quantitative analysis in MRI is challenging due to variabilities in intensity distributions across patients, acquisitions and scanners and suffers from bias field inhomogeneity. Radiomic studies are especially impacted by these effects that affect radiomic feature values. This paper describes a dedicated pipeline to increase reproducibility in breast MRI radiomic studies.
Materials and Methods: T1, T2, and T1-DCE MR images of two breast phantoms were acquired using two scanners and three dual breast coils. Images were retrospectively corrected for bias field inhomogeneity and further normalized using Z-score or histogram matching. Extracted radiomic features were harmonized between coils by the ComBat method. The whole pipeline was assessed qualitatively and quantitatively using a statistical comparison of radiomic feature values.
Results: Intra and inter-acquisition variabilities were strongly reduced by the standardization pipeline. Harmonization by ComBat lowered the percentage of radiomic features significantly different between the three coils from 87% after bias field correction and MR normalization to 3%, while preserving or improving performances of lesion classification on the phantoms.
Discussion: A dedicated standardization pipeline was developed to reduce variabilities in breast MRI, which paves the way for robust multi-scanner radiomic studies but needs to be assessed on patient data.

Introduction

Radiomics is a recent field of study involving the extraction of large amounts of quantitative imaging features from radiological images [8]. These radiomic features can then feed machine learning methods to build predictive models that might assist diagnosis and patient monitoring.

Radiomic studies in breast cancer patients have shown promises, for instance in assessing the risk of breast cancer recurrence [221], detecting malignant from benign lesions [13, 112], or estimating disease free survival [117]. Over the last few years, several attempts to predict the response to neoadjuvant chemotherapy using radiomics have been reported, but this remains a challenging task [134, 144, 145, 169, 222].

Radiomic studies and subsequent machine learning approaches require a substantial number of images to achieve relevant performance, which encourages the use of multicentric and retrospective data. However, many articles have highlighted the influence of scanner parameters on radiomic features in PET and CT imaging (see for instance the review [135]) and in MRI [136–138, 223]. This so-called "scanner effect" requires standardization and harmonization procedures. In particular, MRI radiomic feature values have been shown to depend on magnetic field strengths, voxel size, pulse sequence parameters or receiver coils [139, 140]. The standardization process is especially important in MR as images are expressed in arbitrary units that vary between patients, acquisitions and scanners. MR images also suffer from MR bias field non-uniformity, generating regional and local spatial inhomogeneities. Since the impact of this latter effect goes beyond the field of radiomics and affects tasks such as segmentation, an abundant literature already addresses this issue, but studies are mainly oriented towards brain MRI. Several methods have been developed to correct bias field inhomogeneity retrospectively [141]. Frackiewicz et al. [224] compared a subset of these approaches on breast phantoms. They found that the N4 algorithm [20] gave the most uniform results, slightly outperforming F3CM [225], but hinted that adapting the parameters of the method specifically for breast imaging could improve the correction. Following bias field correction, MR normalization techniques have been applied to reduce inter-patient variabilities, the most frequent being the Z-score standardization [134]. Shinohara et al. [226] designed a new linear approach, the hybrid White-Stripe, using white matter as a reference tissue in the brain to normalize images, from which Fortin et al. [227] derived the voxel-based RAVEL method. Other non-linear normalization techniques have been proposed such as histogram matching (referred as HM) by Nyul et Udupa [21], further adapted in a multiple sclerosis study by Sha et al. [22]. Bias field correction and intensity normalization have been shown to improve the radiomic characterization of tumors from single center MR images of paediatric brain tumors [228] and lung cancer [229].

Recent works in glioblastoma [230, 231] and prostate cancer [140, 232] patients specifically investigated the influence of bias field correction, noise reduction and histogram normalization on the "scanner effect" affecting MR radiomic features. Authors identified small subsets of features that were reproducible across scanners after standardization but did not manage to successfully harmonize all features. A harmonization method called ComBat initially developed to mitigate batch effects in genomic studies [23] was successfully applied to compensate for the "scanner effect" in PET [233], CT [234] and MR [174]. Few breast radiomic studies mention bias field correction [113] or MR normalization [134], before computing features.

In this study, we propose and validate a radiomics pipeline dedicated to breast MRI. First, a bias field correction method was adapted for breast MR images to overcome the limitations of the conventional approaches. Second, two MR image standardization techniques (Z-score, HM) were investigated to study their impact on MR intensity distribution. Third, the ComBat method was proposed to further reduce the "scanner effect" affecting radiomic features. Our experimental study was conducted using two breast phantoms designed for biopsy training in

order to monitor the effects of standardization and harmonization without any biological or tissue interference. MR acquisitions following the standard clinical protocol in our institution were performed using two MR scanners and three dedicated breast coils. Our pipeline efficacy was assessed by studying the reproducibility of each radiomic feature across thirty regions mimicking normal tissue and by comparing the performances of lesion classification on the phantoms before and after harmonization by ComBat.

Materials & Methods

Phantom

In order to remain as close as possible to clinical settings and to better investigate MR bias field considering the symmetry inherent to breast imaging, all the experiments described were carried out using two phantoms simultaneously. Two Multi-Modality Breast Biopsy and Sonographic phantoms, CIRS reference 073 (Norfolk, VA, USA), were used (Figure 4.1a). They consist of an elastomer membrane simulating the skin and subcutaneous fat layer of breast in patients and are filled with a branded gel (Zerdine®). Five to ten cystic lesions (5-10 mm) and ten to fifteen dense lesions (5-10 mm) are included in the gel. Half of the dense lesions are spheres including microcalcifications (Figure 4.1c) while the other half are spiculated (Figure 4.1d). This model is dedicated to biopsy training and accurately reproduces breast tissue with lesions for MR imaging.

Image acquisition

Images were acquired in the three clinical imaging settings in which patients can be imaged at our institution. The two phantoms were scanned in a first setting, using a 1.5 T magnet, Optima MR450w (GE, MA, USA) with an 8-channel breast coil, further referred as "Coil 1". They were also scanned in a second setting, using a 1.5T magnet, MAGNETOM Aera (Siemens, Munich, Germany) with an 18-channel breast coil ("Coil 2"). The third setting consisted of using a 16-channel Sentinelle breast coil, dedicated to diagnosis and MR-guided biopsy, on the 1.5T MAGNETOM Aera (Siemens, Munich, Germany) ("Coil 3"). For each setting, two acquisitions were acquired (Acq. A and Acq. B), between which the positions of the two phantoms on the dual breast coils were switched.

The phantoms were scanned with three sequences routinely used in breast clinical imaging protocol to get T1-weighted, fat-saturated T2-weighted and T1-weighted DCE images, with parameters listed in Table 4.1. T1-DCE images and T1 images on the GE scanner used spoiled gradient recalled techniques whereas T2 and other T1 images used turbo spin echo. The T1 sequence was acquired for anatomical purposes and does not include fat saturation, while T1-DCE was acquired for functional imaging and includes high resolution voxels and fat saturation. Parallel imaging techniques such as ARC for the GE machine and GRAPPA for Siemens were used in T1-DCE. Global acquisition time was around 18 minutes for every coil. Scanning parameters were determined by the clinical protocols set up in our institution and may differ for each coil, adding another variability to the settings. Only the first dynamic of T1-DCE, acquired 90s after what would have been in patients the injection of a contrast product was presented throughout this work. This study consists of six acquisitions (two acquisitions per

coil) with three sequences each (T1, fat-saturated T2 or T1-weighted DCE), yielding eighteen raw 3D images.

Table 4.1: Scanning parameters of routine sequences of imaging devices of Institut Curie.

	T1			fat-saturated T2			T1-weighted DCE		
	Coil 1	Coil 2	Coil 3	Coil 1	Coil 2	Coil 3	Coil 1	Coil 2	Coil 3
TR (ms)	6.9	592	545	5544	3310	6400	6.81	5.2	5.2
TE (ms)	4.2	13	13	90	88	88	3.3	2.4	2.4
Slice thickness (mm)	1.6	3.5	3.0	3.0	3.5	3.0	1.0	0.9	0.9
Spacing between slices(mm)	0.8	4.2	3.6	3.3	4.2	3.6	1.0	0.9	0.9
Pixel spacing (mm)	0.68x0.68	0.71x0.71	0.68x0.68	0.70x0.70	0.70x0.70	0.70x0.70	0.82x0.82	0.91x0.91	0.91x0.91
Pixel bandwidth (Hz/pixel)	244	130	130	558	315	375	434	355	355
Flip angle	20	148	148	160	150	180	15	10	10

***Coil 1**: Optima MR450w with 8-channel coil; **Coil 2**: Magneton Aera with 18-channel brest coil; **Coil 3**: Magneton Aera with Sentinelle breast coil.*

Bias field correction

Images were corrected for bias field inhomogeneity using the SimpleITK N4BiasFieldCorrection Image filter class adapted for python from the implementation of the N4 algorithm [20] in the ITK library. The N4 method is based on the following image model:

$$I_{cor}(x) = I(x)B(x) + \eta(x) \tag{4.1}$$

where x is a voxel, I_{cor} is the corrupted image, B the bias field, I the bias-free image and η an independent Gaussian noise.
In a noise-free case, using logarithmic transformation, with $\hat{I} = logI$:

$$\hat{I}_{cor} = \hat{I} + \hat{B} \tag{4.2}$$

The N4 method uses an iterative multi-scale optimisation approach, at iteration n:

$$\hat{I}^n = \hat{I}^{n-1} - \hat{B}^n_{res} \tag{4.3}$$

$$\hat{I}^n = \hat{I}^{n-1} - S^*\{\hat{I}^{n-1} - E[\hat{I}|\hat{I}^{n-1}]\} \tag{4.4}$$

where S^* is an adapted B-spline approximation, $\hat{I}^0 = \hat{I}_{cor}$ and B_{res} is the residual bias field at step.

The number of resolution levels and the number of iterations at each level are set by default to four levels and fifty iterations but can be changed. A default mask to select the pixels used to estimate the bias field is defined using Otsu thresholding unless a specific mask is provided.

The impact of the hyper parameters mentioned above were investigated by running several trials using three, four, five or six resolution levels, fifty or a hundred iterations, combined with either the default mask or a full mask of the phantom.

To assess the ability of the N4 correction to reduce intensity non-uniformities within similar tissue types, voxels corresponding to the background gel and embedded masses were clustered using the k-means algorithm. The clustering results were compared before and after the bias field correction. As the inner part is made of three different materials (background gel, dense masses and cyst masses), that have different physical properties, the number of clusters was set to three.

The performance of N4 algorithm was also assessed by comparing the coefficients of variation of the mean intensity of small regions drawn in the background gel for the different corrections: fifteen 3D spherical regions of 600 voxels each were drawn using the LIFEx freeware [235] (www.lifexsoft.org) on every raw acquisition. These spheres were located in the background neutral gel of the phantom, avoiding any cyst or dense masses. As there were two acquisitions per coil, regions from the same coil were pooled to get thirty regions per coil.

MR normalization

MR images were normalized after bias field correction as it has been shown that pre-correcting intensity non-uniformities leads to an improved standardization [236]. Two types of normalization were performed separately and compared: 1) Z-score standardization using a mask of the phantom to compute the mean and standard deviation of intensities (linear transform); 2) piecewise linear histogram matching [21, 22]. Histogram matching includes two stages: first, HM learns landmarks of a standard histogram and then landmarks of the image histograms are non-linearly mapped to the ones of the standard histogram to align the intensity distributions. HM was applied independently on the three sequences with codes adapted from Reinhold et al. [237], using the decile landmarks and standard scale defined by Shah et al. [22]. The impact of normalization in correcting inter-subject and inter-coil variabilities was evaluated by qualitatively comparing intensity histogram alignment and by using texture analyses.

Texture analysis

After MR normalization, four MR volumes of the same acquisition were available, corresponding to raw data, N4-corrected data, Z-score normalized-N4-corrected data, and HM normalized-N4-corrected data.

All 18x4 (3 sequences x 3 coils x 2 acquisitions x 4 normalizations) MR volumes were resampled, as recommended by Image Biomarker standardization Initiative guidelines [143] before extracting features, using nearest neighbour interpolation: T1 and fat-saturated T2 images were resampled to 0.7x0.7x4 mm^3 voxels and T1-weighted DCE images to 1x1x1 mm^3 voxels.

For each MR volume and each of the fifteen regions described previously, forty-two radiomic features were computed with LIFEx v5.79 [235] in compliance with the Image Biomarker standardization Initiative guideline [143]. Besides first order features, features included indices from the grey-level co-occurrence matrix (GLCM), the grey-level run length matrix (GLRLM), the grey-level zone length matrix (GLZLM) and the neighbourhood grey-level different matrix (NGLDM). The list of radiomic features is provided in Supplemental Table 1. For texture

calculations, absolute discretization was chosen [228]. For each sequence and each step of the standardization pipeline separately, the minimum and maximum intensities inside the regions were calculated to determine the range of intensities and the average standard deviation of intensities over the regions was defined as the fixed bin size.

Harmonization of radiomic features

As radiomic features are computed inside similar regions, they should be comparable. Since radiomic feature values might differ between the three experimental settings even after the different processing steps, for every sequence separately, the distributions of the radiomic features extracted after normalization were harmonized across the three coils using the ComBat method [233, 234]. The ComBat method intends to correct any underlying differences that could be due to coils, scanners and/or scanning parameters [40, 41]. For feature y measured in region j in center i, feature y_{ij} can be modelled as:

$$y_{ij} = \alpha + \gamma_i + \delta_i \epsilon_{ij} \tag{4.5}$$

where α is the average value of y, γ_i is an additive center effect and $\delta_i \epsilon_{ij}$ a multiplicative center effect associated to an error term.

The ComBat method corrects the distributions by calculating $\hat{\alpha}$, $\hat{\gamma_i}$ and $\hat{\delta_i}$ as estimators of α, γ_i and δ_i using maximum likelihood estimation so that:

$$y_{ijcorrected} = \frac{y_{ij} - \hat{\alpha} - \hat{\gamma_i}}{\hat{\delta_i}} + \hat{\alpha} \tag{4.6}$$

The non-parametric form of the method was used without any empirical Bayes assumption. A specific transformation was determined for every feature independently, using R codes by Fortin et al. [238, 239].

Statistical analysis

Statistical analyses were performed in R. p-values less than 0.05 were interpreted as statistically significant. For each step of the pipeline and after the harmonization by ComBat, differences in statistical distributions of radiomic features between coils were assessed with the Kruskal-Wallis test. 3 (sequences) x 5 (raw, N4 correction, Z-score, HM, HM & ComBat) x 42 (radiomic features) Kruskal-Wallis tests were performed. To provide a synthetic view of the test results, five ranges of p-values were defined: $p < 10^{-5}$, $10^{-5} \leq p < 10^{-3}$, $10^{-3} \leq p < 0.01$, $0.01 \leq p < 0.05$ and $0.5 \leq p$. Radiomic features were then put into the five classes defined by the previous ranges of p-values, depending on the p-value of their Kruskal-Wallis test. For each sequence at each step of the pipeline, the number of features in every class was calculated and reported in a table.

For there could be concerns that the ComBat method harmonized data too much thus reducing the discriminative power of features, we proposed to evaluate the impact of ComBat on the task of separating two types of dense lesions (Figure 4.1c and Figure 4.1d) on the different sequences. Lesions were segmented on the three coils semi-automatically with the k-means algorithm and corrected by hand. Features were extracted from the lesions on the HM-normalized-N4-corrected images and compared before and after harmonization by ComBat using Wilcoxon tests.

Results

Bias field correction

Default parameters of the N4 algorithm (four resolution levels, fifty iterations per level and the use of a mask defined by Otsu thresholding) proved suboptimal for breast MR images. From a qualitative point of view, the corrected images showed little improvement when compared to the raw images (Figure 4.1b and 1g). The bias field estimated with the default parameters was almost flat on the upper half of the images (Figure 4.1f). These upper regions are the regions of interest where clinical information will be looked for, whereas lower regions have less clinical relevance (corresponding to zones posterior to the thorax in patients). Increasing the number of iterations per level did not improve the corrections as the results seemed to stabilise after fifty iterations. Running the algorithm with five resolution levels instead of four yielded a bias field with much stronger variations in the upper zones resulting in a greater correction of the images. Combining it with a full mask of the phantoms to estimate the bias field further improved the correction in the upper regions of interest (Figure 4.1m).

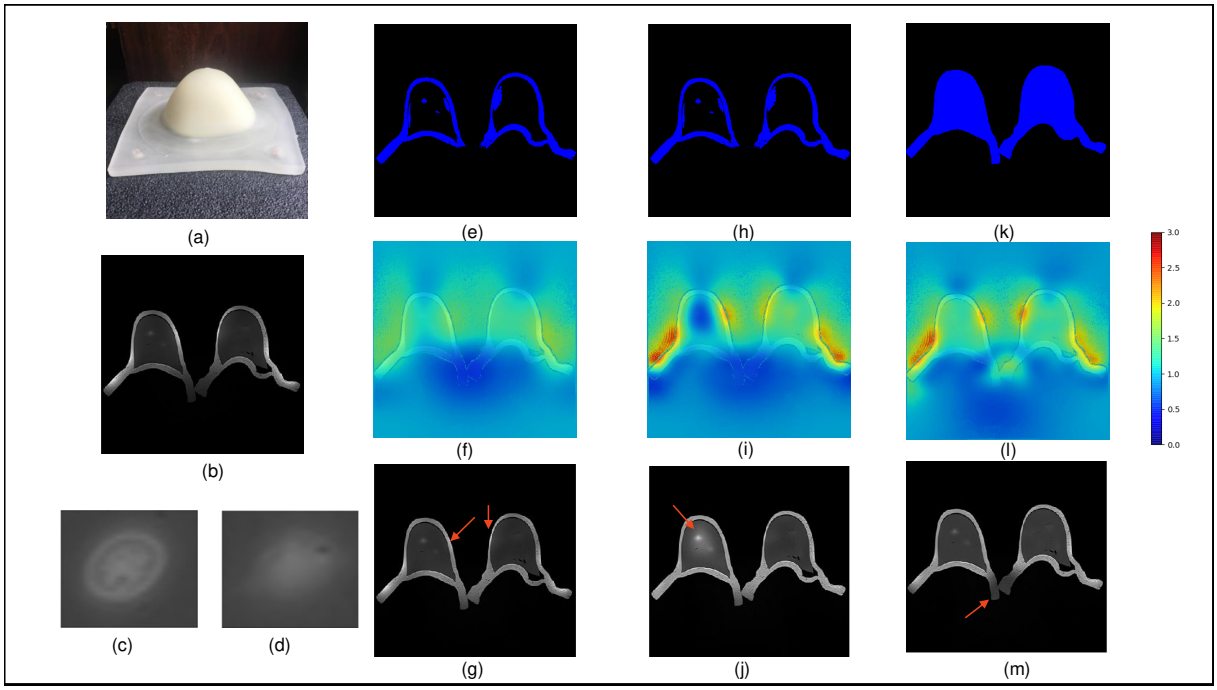


Figure 4.1: ***(a)*** *Phantom.* ***(b)*** *Raw T1 image from Coil 3.* ***(c)*** *Dense lesion with microcalcification.* ***(d)*** *Dense spiculated lesion.* ***(e)*** *Default mask.* ***(f)*** *Bias field estimated with mask e and 4 fitting levels.* ***(g)*** *Corrected image obtained from bias field f.* ***(h)*** *Default mask.* ***(i)*** *Bias field estimated with mask h and 5 fitting levels.* ***(j)*** *Corrected image from bias field i.* ***(k)*** *Full mask.* ***(l)*** *Bias field estimated with mask k and 5 fitting levels.* ***(m)*** *Corrected image from bias field k. Red arrows point at regions with residual intensity non-uniformity*

The impact of the N4 correction on the coefficients of variation of the means over the regions across coils in different correction scenarios is shown in Figure 4.2.

The effect of N4 correction with the full mask, 5 levels (50 Iterations) on k-means segmentation results using three clusters on a raw image versus the image after bias field correction is shown in Figure 4.3.

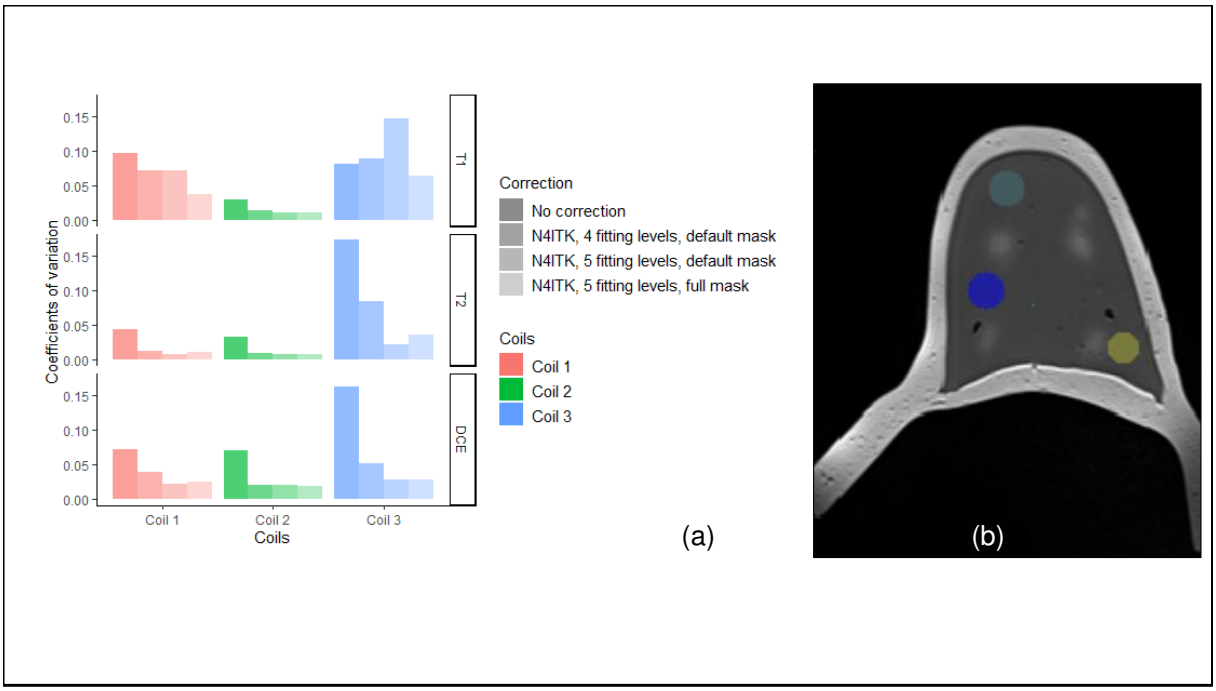


Figure 4.2: ***(a)*** *Coefficients of variation of the means over thirty regions across settings with different corrections.* ***(b)*** *Example of 3 regions (in blue, light blue and yellow) drawn in LIFEx.*

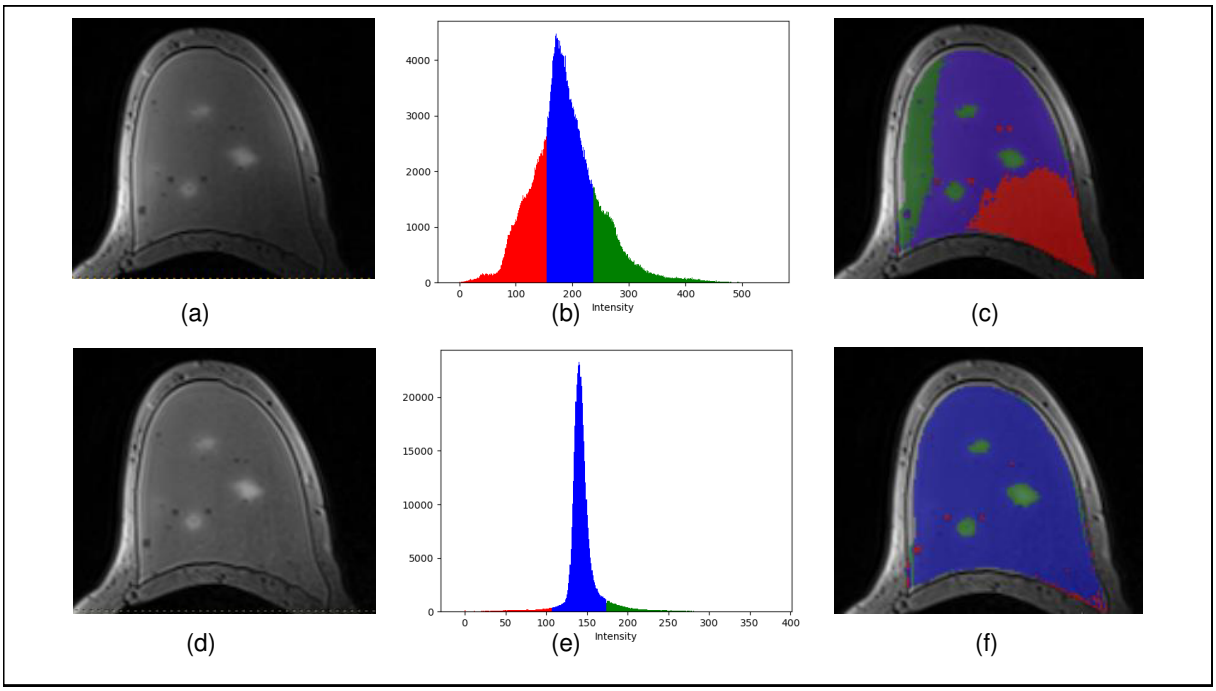


Figure 4.3: ***(a)***, ***(d)*** *T1-weighted DCE image from coil 3.* ***(b)***, ***(e)*** *Histogram of the inner layer voxels of image coloured by the results of k-means clustering.* ***(c)***, ***(f)*** *k-means clustering results overlaid on image. First line: raw image. Second line: N4 corrected (full mask, 5 levels, 50 iterations) image.*

Figure 4.4 presents all bias field estimations across sequences and acquisitions, with images normalized so that the mean intensity in the mask used to estimate the bias field is 1.

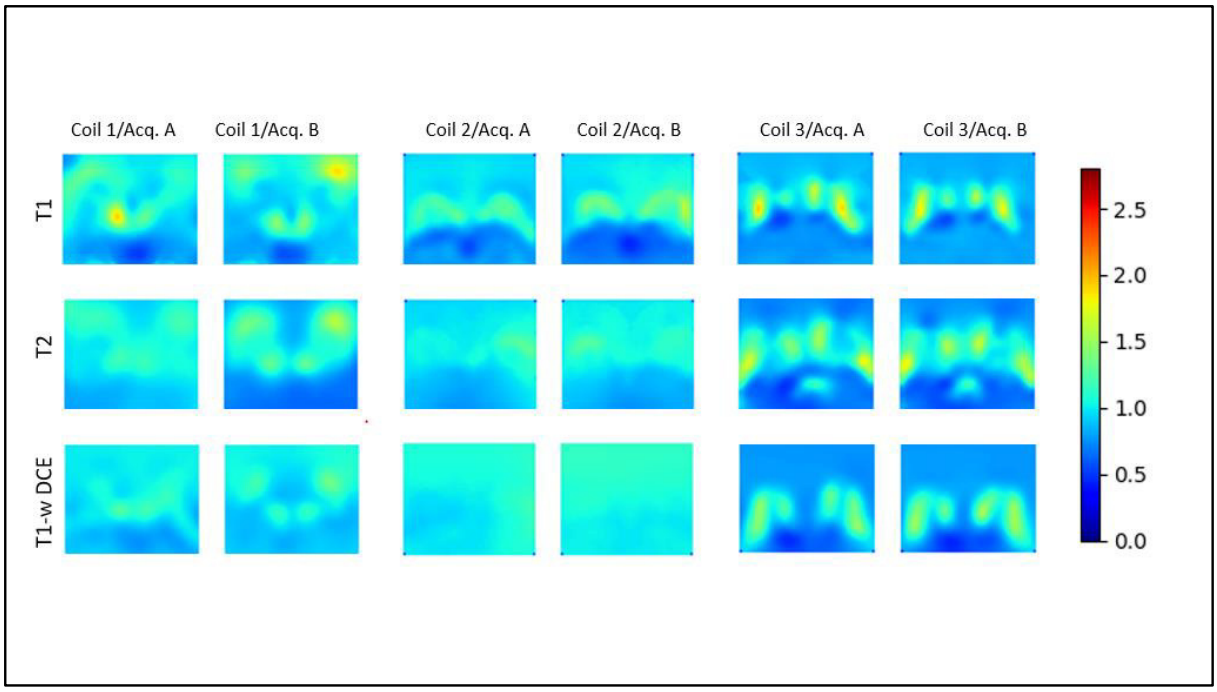


Figure 4.4: Examples of estimated bias fields across sequences and acquisitions.

MR normalization

Histograms of image intensities within the phantoms (as defined by the mask used for N4 correction) are shown in Figure 4.5 for the four stages of the post-processing pipeline. The different post-processing methods had a similar behaviour across the three sequences. In raw images, the peaks of the histograms were not aligned before any correction. Intensities from coil 1 (the GE machine), in particular, spread on a significantly greater range than the intensities from the two Siemens coils. N4 correction sharpened the peaks but did not align them. Z-score normalization combined with N4 correction realigned perfectly acquisitions from the same coils and managed to align the peaks of different coils around the same value. The alignment was nevertheless not optimal, especially in high intensities in T2 images. Histogram matching produced the best alignments whatever the coils.

Harmonization of radiomic features

To illustrate the impact of the pipeline on radiomic feature values, Figure 4.6 shows the statistical distributions of the Short-Zone High Gray-level Emphasis (GLZLM-SZHGE) feature extracted from regions on fat-saturated T2 images across coils for the four stages of the standardization pipeline, and after further harmonization using ComBat. This example shows that the Z-score and HM normalization contributed to realigning the distributions across coils (Figure 4.6c, 6d) but that further harmonization using ComBat was needed to co-align all three coils distributions (Figure 4.6e). Figure 4.6f presents a plot of the ComBat corrected GLZLM-SZHGE against the feature before its harmonization to emphasize the action of ComBat and illustrate the different transformations applied to the features depending on the coil.

Table 4.2 reports the number of features for which Kruskal-Wallis p-values are inside a

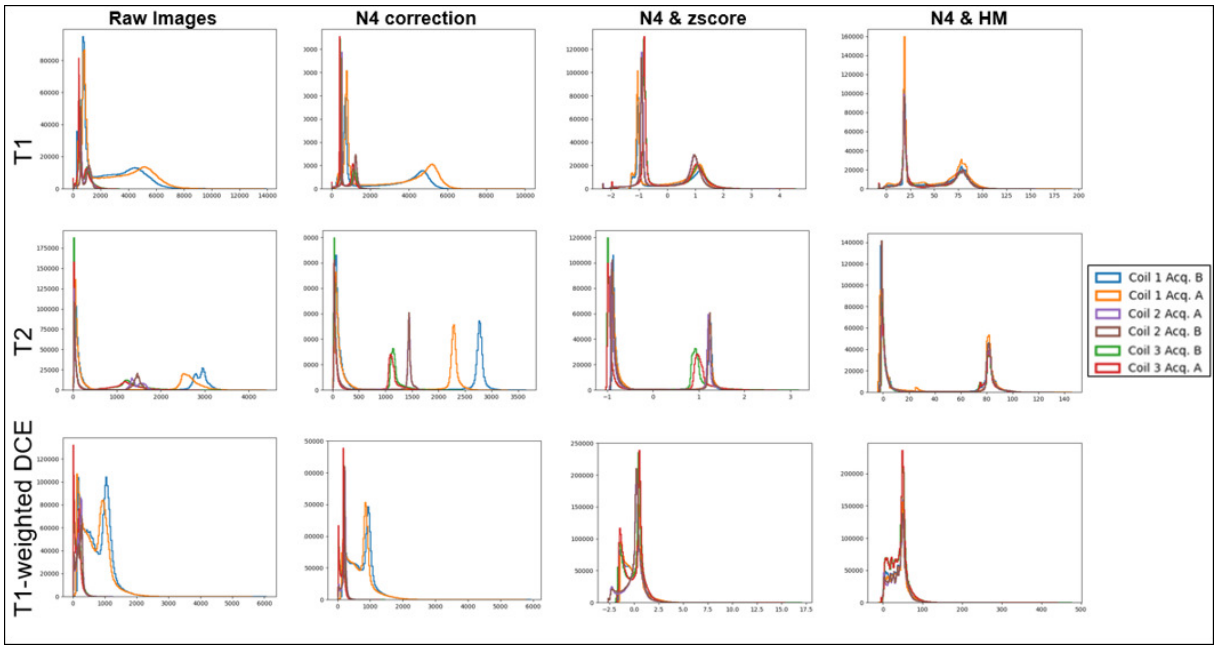


Figure 4.5: Image intensity histograms of the six acquisitions (Acq.) for the four steps of the standardization pipeline across the three sequences. Each row represents a sequence and each column a step of the pipeline.

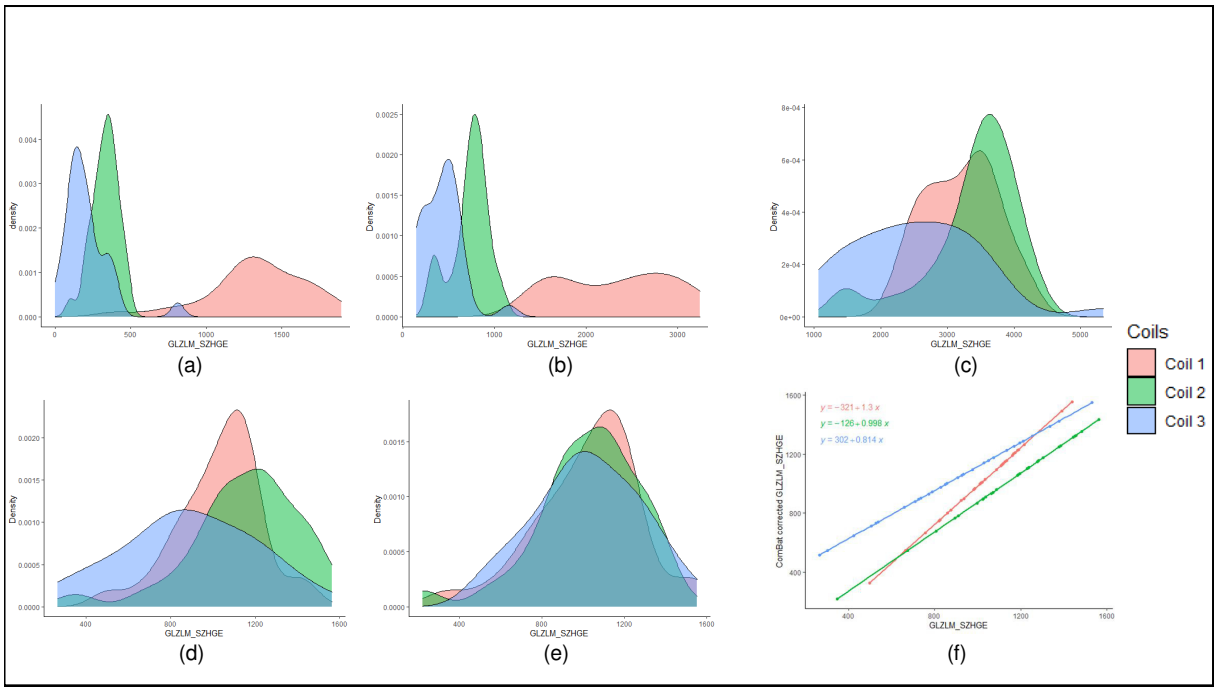


*Figure 4.6: Statistical distributions across coils of the GLZLM-SZHGE texture feature extracted from **(a)** raw T2 images. **(b)** N4 corrected (full mask, 5 levels) T2 images. **(c)** Z-score normalized-N4 corrected T2 images. **(d)** Histogram-matched-N4 corrected T2 images. **(e)** Histogram-matched-N4 corrected T2 images and harmonized by ComBat. **(f)** Transformation carried out by ComBat on the GLZLM-SZHGE extracted from Histogram-matched-N4 corrected T2 images depending on coils.*

specific range: in T1 raw images, 37 out 42 features were significantly different between the 3 coils with a p-value $p < 10^{-5}$, 3 features were significantly different with a p-value $10^{-5} \leq p < 10^{-3}$, 1 with a p-value $10^{-3} \leq p < 0.01$ and the last one with a p-value $0.01 \leq p < 0.05$. The ranges of p-values of the tests run on the forty-two radiomic features for each step and each sequence are provided in Supplemental Tables 2, 3 and 4.

The same pattern was observed across sequences: the number of features that were significantly different ($p < 0.05$) decreased gradually when they were computed from N4-corrected data, Z-score-normalized N4-corrected data, HM-normalized N4-corrected data, ComBat harmonized HM-normalized N4-corrected data and the number of small p-values was reduced accordingly. Harmonization by ComBat was essential to reduce drastically the number of significantly different features in all three settings, especially for T1-weighted DCE features.

P value	Raw images	N4 correction	N4 & Z-score	N4 & HM	N4 & HM & ComBat
T1					
$p<10^{-5}$	37	37	24	5	0
$10^{-5}\leq p<10^{-3}$	3	1	6	7	0
$10^{-3}\leq p<0.01$	1	3	3	11	0
$0.01\leq p<0.05$	1	0	3	9	2
$0.05\leq p$	0	1	6	10	40
T2					
$p<10^{-5}$	38	36	26	13	0
$10^{-5}\leq p<10^{-3}$	4	2	5	6	0
$10^{-3}\leq p<0.01$	0	1	5	10	0
$0.01\leq p<0.05$	0	2	2	4	1
$0.05\leq p$	0	1	4	9	41
T1-weighted DCE					
$p<10^{-5}$	39	39	39	40	0
$10^{-5}\leq p<10^{-3}$	1	0	1	1	0
$10^{-3}\leq p<0.01$	1	1	0	0	0
$0.01\leq p<0.05$	0	0	0	0	1
$0.05\leq p$	1	2	2	1	41

Table 4.2: Number of radiomic features in the 5 different ranges of p-values. The p-values correspond to Kruskal-Wallis tests between the three coils, radiomic features being extracted from 30 similar regions. Results are given for the three MR sequences and each main step of the processing pipeline.

Wilcoxon tests were performed between the two dense lesion types segmented on the phantoms across sequences. Before harmonization, on T1 (respectively T2, T1-DCE) images, 10 (respectively 39, 7) features out of 42 were significantly different between the two lesion types, whereas after ComBat harmonization, 32 (respectively 39, 21) features were significantly different. Figure 4.7 shows the impact of ComBat on the mean intensity.

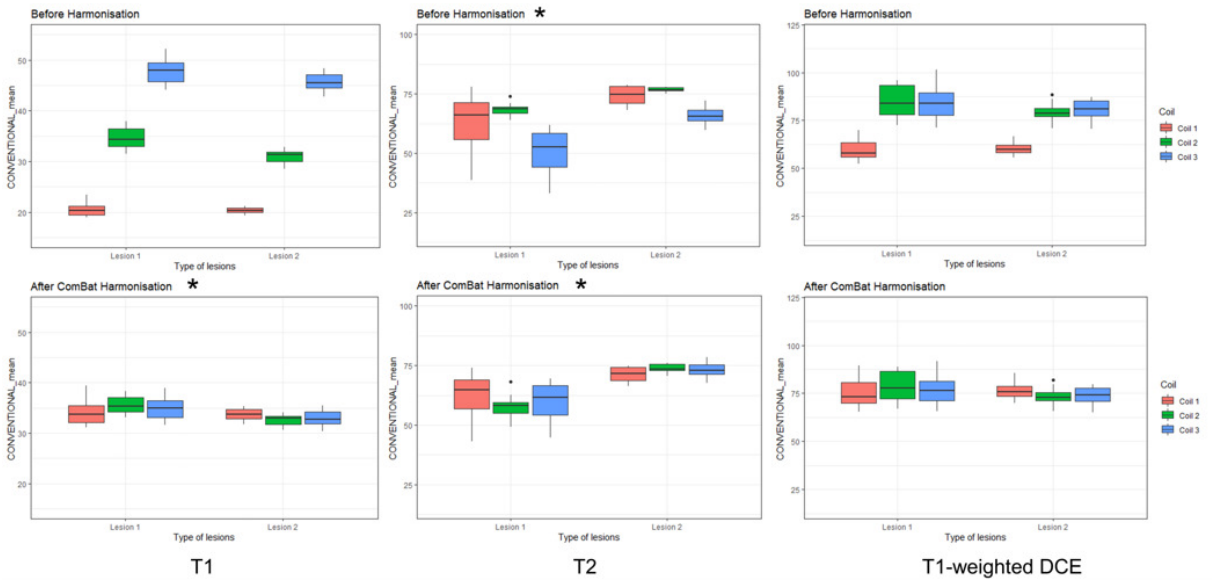


Figure 4.7: Mean intensity before and after ComBat harmonization across sequences and coils of Lesion 1 (dense lesion with microcalcification) and Lesion 2 (dense spiculated lesion). Asterisks denote cases where the difference between the two lesion types is significant.

Discussion

The present study suggests that standardization methods developed for brain or lung MRI should be adapted specifically to breast MR images. The whole process includes bias field correction to reduce local/regional inhomogeneities in similar regions (intra-image variabilities), intensity normalization to lessen inter-acquisitions variabilities, and statistical harmonization to make results across coils comparable. We have shown that the three steps, each tackling a different kind of variability, are all needed and complementary. They pave the way towards an efficient standardization pipeline for multi-scanner radiomic studies of patients' acquisitions.

To enable retrospective patients' studies, bias field correction was based on an a posteriori method. Comparisons of bias fields with different settings of the N4 algorithm led to a set of parameters appropriate for breast MRI when using dual breast coils. Based on our study, we recommend using a mask including the internal part of the breast phantoms (unlike the mask defined by Otsu's threshold) and performing the optimisation across five resolution levels (instead of four) with fifty iterations per level. Using the default parameters optimised for brain MRI underestimated the variations in the bias field, even when using the mask including the phantom inner part. It resulted in intensity non-uniformity inside this mask, where MR information is of prime importance in a clinical context (Figure 4.1g). The drawback of the default mask and five resolution levels was also illustrated in Figure 4.1. The bias field was indeed underestimated in the central part of the phantom, yielding a hypersignal effect in the corrected image (Figure 4.1j) and thus increased heterogeneity in the background gel hence an increase in the coefficient of variation of the mean intensity (Figure 4.2). The analysis of the quantitative assessments of all experiments (Figure 4.2) showed that the proposed breast specific N4 parameters led to the greatest decrease, across coils, of the coefficients of variation of the mean intensity over homogeneous regions inside the phantom. The k-means clustering performed on the inner part of the phantom clearly shows how N4 correction

reduced intensity variations across tissue types. In addition, the intensity histogram of the inner part of the phantom on N4 corrected images showed a strong sharpening of its peak around the mean value of the largest structure, i.e. the background gel (Figure 4.3e). The overlay of the segmented regions demonstrated a clear improvement in the identification of masses on N4 corrected images (Figure 4.3f). Bias field correction thus appeared essential to improve homogeneity inside the breast MR images and is crucial for a correct segmentation of abnormalities in the breast. It should be underlined that the estimated bias fields depend not only on the MR scanner, but also on coils, type of sequence (T1, T2, T1-DCE), and on the positioning of the phantoms inside the breast coils. As shown in Figure 4.4, the coil has a high impact. Indeed, bias field images from coils 2 and 3 originating from the same MR scanner were quite different. Using a same coil, bias field also showed large fluctuations across sequences. This work only presents the first dynamic of the T1-DCE sequence, but should several dynamics be studied, the N4 correction would have been applied separately on each dynamic.

To reduce inter-subject and multi-scanner variabilities, MR normalization was performed after bias field correction. Linear approaches using a reference tissue, similar to Shinohara et al. method [226] involving white matter in the brain were not reported as no satisfying reference tissue could be found in breast for all sequences, despite attempts with the subcutaneous fat layer of the breast. Studying the co-alignment of intensity histograms across acquisitions and coils highlighted the impact of intensity normalization, and the good performance of the histogram matching approach. Results from Figure 4.5 supported the idea that it was necessary to go beyond linear normalization and Z-score standardization [134], confirming findings by Nyul et al. [21] and Fortin et al. [227]. Z-score normalization indeed squashed all intensities inside a range of values but did not succeed at aligning tissue-specific peaks. As observed by Isaksson et al. [24] (though with different types of landmarks) in the normalization of prostate radiomics, the piecewise linear histogram matching gave excellent results in realigning intensity distributions. However, histogram matching depends on the set of images selected to extract a standard histogram. In a clinical setting, inclusion of new patients in a study often goes on after the beginning of processing work on the original database. To avoid recalculating a standard histogram every time new patients are added to the database, it is thus important to select the images from which to extract it across a wide range of scanner and biological variabilities to identify robust landmarks [22]. Considering radiomic features computed inside thirty similar regions, statistical tests showed that N4 correction combined with histogram matching normalization could not completely remove the "scanner effect". Each stage of the pipeline decreased the number of features that were significantly different between the three coils, but it was not sufficient to harmonize all radiomic features. This result agrees with the trends reported in glioblastoma [230, 231] and prostate [140, 232] cancer patients. Further harmonization of the radiomic features is needed and ComBat succeeded in realigning feature distributions across scanners. Some studies normalize the features using scaling or Z-score [240, 241] separately for each center but unlike ComBat, these methods cannot model possible covariates that could affect the features leading to skewed data [23, 238, 239]. Performances in separating the two dense lesion types were preserved or improved by the ComBat harmonization, therefore suggesting that ComBat was able to successfully harmonize the features across coils while preserving their biological variabilities and their predictive powers. Though the effect of ComBat is prevailing in reducing the "scanner effect", the N4 correction

and the normalization are paramount to reduce intra-image and inter-acquisition variabilities on which the affine transformation applied by ComBat has no effect. Combining the corrections is thus essential to lessen all types of variabilities.

The present study has several limitations. First, the CIRS model was built to be usable in multiple imaging modalities and not specifically in MRI. The phantom was also aimed at biopsy training providing lesions that could be biopsied multiple times and was therefore not designed for radiomic studies unlike phantoms used in normalization [136, 138, 242]. The phantom was made from simple materials to capture the global breast heterogeneous appearance but not to mimic the very fine heterogeneity that could be observed in tumors and modelled in other phantoms [243]. Another limitation is that our experiments were performed using two MR scanners and three coils at the same institution, but we are confident in the possibility to extend our results to other scanners, centers and acquisitions protocols. Finally, there is always an inherent limit in using a phantom to assess performances of methods that we want to apply in clinical settings. Nevertheless, phantoms offered the opportunity to properly monitor the effects of standardization without any interference of biological covariates.

Conclusion

This study shows the necessity to use a standardization pipeline before performing radiomic studies involving MR breast images acquired using multiple settings. A retrospective bias field correction dedicated to dual breast coils and non-linear MR intensity normalization reduced the "scanner effect" for subsets of radiomic features, but further statistical harmonization was needed to fully correct for it. The results were obtained on breast phantoms and future work will assess the pipeline on patient data, where biological and pathological variations increase the sources of MR intensity variations.

Supplemental data

Supplemental Data for article: A radiomics pipeline dedicated to Breast MRI: validation on a multicentre phantom study, in Magn Reson Mater Phy

Authors: Marie-Judith Saint Martin[1], Fanny Orlhac[1], Pia Akl[1,2,3], Fahad Khalid[1], Christophe Nioche[1], Irène Buvat[1], Caroline Malhaire [1,3], Frédérique Frouin[1]

[1] Université Paris-Saclay, Inserm, Institut Curie, Laboratoire d'Imagerie Translationnelle en Oncologie (LITO), Bât 101B rue Henri Becquerel, 91 401 Orsay, France.

[2] HCL, Radiologie du groupement hospitalier Est, Hôpital Femme Mère enfant, Unité Fonctionnelle: imagerie de la femme, 3 Quai des Célestins, 69002 Lyon, France.

[3] Institut Curie, Service de Radiodiagnostic, 26 rue d'Ulm, 75005 Paris, France.

Corresponding author: Marie-Judith Saint Martin, email: marie-judith-astrid.saint-martin@u-psud.fr

1	CONVENTIONAL_min
2	CONVENTIONAL_mean
3	CONVENTIONAL_std
4	CONVENTIONAL_max
5	CONVENTIONAL_Q1
6	CONVENTIONAL_Q2
7	CONVENTIONAL_Q3
8	HISTO_Skewness
9	HISTO_Kurtosis
10	HISTO_Entropy
11	HISTO_Energy (Uniformity)
12	GLCM_Homogeneity (Inverse Difference)
13	GLCM_Energy (Angular Second Moment)
14	GLCM_Contrast (Variance)
15	GLCM_Correlation
16	GLCM_Entropy
17	GLCM_Dissimilarity
18	GLRLM_SRE
19	GLRLM_LRE
20	GLRLM_LGRE
21	GLRLM_HGRE
22	GLRLM_SRLGE
23	GLRLM_SRHGE
24	GLRLM_LRLGE
25	GLRLM_LRHGE
26	GLRLM_GLNU
27	GLRLM_RLNU
28	GLRLM_RP
29	NGLDM_Coarseness
30	NGLDM_Contrast
31	NGLDM_Busyness
32	GLZLM_SZE
33	GLZLM_LZE
34	GLZLM_LGZE
35	GLZLM_HGZE
36	GLZLM_SZLGE
37	GLZLM_SZHGE
38	GLZLM_LZLGE
39	GLZLM_LZHGE
40	GLZLM_GLNU
41	GLZLM_ZLNU
42	GLZLM_ZP

Supplemental Table 1 List of radiomic features extracted using the LIFEx freeware

		Raw images	N4 correction	N4 & Z-score	N4 &HM	N4 & HM & ComBat
1	CONVENTIONAL_min					
2	CONVENTIONAL_mean					
3	CONVENTIONAL_std					
4	CONVENTIONAL_max					
5	CONVENTIONAL_Q1					
6	CONVENTIONAL_Q2					
7	CONVENTIONAL_Q3					
8	HISTO_Skewness					
9	HISTO_Kurtosis					
10	HISTO_Entropy					
11	HISTO_Energy (Uniformity)					
12	GLCM_Homogeneity (InverseDifference)					
13	GLCM_Energy (Angular SecondMoment)					
14	GLCM_Contrast (Variance)					
15	GLCM_Correlation					
16	GLCM_Entropy_log10					
17	GLCM_Dissimilarity					
18	GLRLM_SRE					
19	GLRLM_LRE					
20	GLRLM_LGRE					
21	GLRLM_HGRE					
22	GLRLM_SRLGE					
23	GLRLM_SRHGE					
24	GLRLM_LRLGE					
25	GLRLM_LRHGE					
26	GLRLM_GLNU					
27	GLRLM_RLNU					
28	GLRLM_RP					
29	NGLDM_Coarseness					
30	NGLDM_Contrast					
31	NGLDM_Busyness					
32	GLZLM_SZE					
33	GLZLM_LZE					
34	GLZLM_LGZE					
35	GLZLM_HGZE					
36	GLZLM_SZLGE					
37	GLZLM_SZHGE					
38	GLZLM_LZLGE					
39	GLZLM_LZHGE					
40	GLZLM_GLNU					
41	GLZLM_ZLNU					
42	GLZLM_ZP					

Supplemental Table 2 Ranges of *p*-values of Kruskal-Wallis tests on 42 radiomic features extracted from T2 images between the three coils at each step of the pipeline. Pink is $p < 10^{-5}$, orange corresponds to $10^{-5} \leq p < 10^{-3}$,yellow represents $10^{-3} \leq p < 0.01$, green is $0.01 \leq p < 0.05$ and white is for $p \geq 0.05$

		Raw images	N4 correction	N4 & Z-score	N4 &HM	N4 & HM & ComBat
1	CONVENTIONAL_min					
2	CONVENTIONAL_mean					
3	CONVENTIONAL_std					
4	CONVENTIONAL_max					
5	CONVENTIONAL_Q1					
6	CONVENTIONAL_Q2					
7	CONVENTIONAL_Q3					
8	HISTO_Skewness					
9	HISTO_Kurtosis					
10	HISTO_Entropy					
11	HISTO_Energy (Uniformity)					
12	GLCM_Homogeneity (InverseDifference)					
13	GLCM_Energy (Angular SecondMoment)					
14	GLCM_Contrast (Variance)					
15	GLCM_Correlation					
16	GLCM_Entropy_log10					
17	GLCM_Dissimilarity					
18	GLRLM_SRE					
19	GLRLM_LRE					
20	GLRLM_LGRE					
21	GLRLM_HGRE					
22	GLRLM_SRLGE					
23	GLRLM_SRHGE					
24	GLRLM_LRLGE					
25	GLRLM_LRHGE					
26	GLRLM_GLNU					
27	GLRLM_RLNU					
28	GLRLM_RP					
29	NGLDM_Coarseness					
30	NGLDM_Contrast					
31	NGLDM_Busyness					
32	GLZLM_SZE					
33	GLZLM_LZE					
34	GLZLM_LGZE					
35	GLZLM_HGZE					
36	GLZLM_SZLGE					
37	GLZLM_SZHGE					
38	GLZLM_LZLGE					
39	GLZLM_LZHGE					
40	GLZLM_GLNU					
41	GLZLM_ZLNU					
42	GLZLM_ZP					

Supplemental Table 3 Ranges of *p*-values of Kruskal-Wallis tests on 42 radiomic features extracted from T1 images between the three coils at each step of the pipeline. Pink is $p < 10^{-5}$, orange corresponds to $10^{-5} \leq p < 10^{-3}$,yellow represents $10^{-3} \leq p < 0.01$, green is $0.01 \leq p < 0.05$ and white is for $p \geq 0.05$

		Raw images	N4 correction	N4 & Z-score	N4 &HM	N4 & HM & ComBat
1	CONVENTIONAL_min					
2	CONVENTIONAL_mean					
3	CONVENTIONAL_std					
4	CONVENTIONAL_max					
5	CONVENTIONAL_Q1					
6	CONVENTIONAL_Q2					
7	CONVENTIONAL_Q3					
8	HISTO_Skewness					
9	HISTO_Kurtosis					
10	HISTO_Entropy					
11	HISTO_Energy (Uniformity)					
12	GLCM_Homogeneity (InverseDifference)					
13	GLCM_Energy (Angular SecondMoment)					
14	GLCM_Contrast (Variance)					
15	GLCM_Correlation					
16	GLCM_Entropy_log10					
17	GLCM_Dissimilarity					
18	GLRLM_SRE					
19	GLRLM_LRE					
20	GLRLM_LGRE					
21	GLRLM_HGRE					
22	GLRLM_SRLGE					
23	GLRLM_SRHGE					
24	GLRLM_LRLGE					
25	GLRLM_LRHGE					
26	GLRLM_GLNU					
27	GLRLM_RLNU					
28	GLRLM_RP					
29	NGLDM_Coarseness					
30	NGLDM_Contrast					
31	NGLDM_Busyness					
32	GLZLM_SZE					
33	GLZLM_LZE					
34	GLZLM_LGZE					
35	GLZLM_HGZE					
36	GLZLM_SZLGE					
37	GLZLM_SZHGE					
38	GLZLM_LZLGE					
39	GLZLM_LZHGE					
40	GLZLM_GLNU					
41	GLZLM_ZLNU					
42	GLZLM_ZP					

Supplemental Table 4 Ranges of *p*-values of Kruskal-Wallis tests on 42 radiomic features extracted from T1-weighted DCE images between the three coils at each step of the pipeline. Pink is $p < 10^{-5}$, orange corresponds to $10^{-5} \leq p < 10^{-3}$,yellow represents $10^{-3} \leq p < 0.01$, green is $0.01 \leq p < 0.05$ and white is for $p \geq 0.05$

4.3 Discussion

The study presents some limitations that are detailed in the section below.

Though the phantoms gave us an easy opportunity to compare the repeated acquisitions of the same objects over three imaging devices, it could not replace a test-retest experiment involving patients. Indeed, phantoms did contain several masses of different sizes and shapes. However, only the masses of one phantom could be really used as the second one was much older and its composition altered. As the phantom was not specifically designed for MRI, it was sometimes very difficult to spot masses and segment them correctly. Accurately identifying a designated mass from one acquisition to another one, was impossible as contrast and changes in the shape of the breast due to the different coils used, made it extremely difficult. A mass could only be identified as a cystic or dense mass.

In the article, bias fields in the three settings were visualized for each modality. However, in patient images, anatomy of the subjects and the injection of a contrast media agent in T1-DCE sequences and the subsequent enhancement of the heart and tumor regions could change the bias field patterns observed. Besides, the correction was tested only on two scanners that had the same magnetic field strength (1.5T). The two constructors (GE healthcare and Siemens) were used in the training and test sets. To make the results more robust, other constructors and devices, and scanners at 3T especially, should be investigated. Strong magnetic fields indeed increase the impact of the imaged object on bias field gain as increased radio-frequencies are needed in these conditions, leading to enhanced radio-frequency standing-waves [219].

Regarding the normalization step, we concluded that histogram matching was the best method to realign intensity peaks of the different materials in the phantoms between the three imaging settings. In the context of considering the phantom experiments as a first step towards pre-processing patient images, these results call for some nuances. Adaptations of the common method of White Stripe normalization, originally designed for brain images and using white matter as reference tissue to normalize them, could not be carried out on the phantom images. Indeed, a satisfactory reference tissue could not be found on fat-saturated T2-weighted images in phantom images but several approaches like using the sternum could be explored on patient images. Moreover, though these results were not included in the article as the point was to stress the importance of each of the correcting step, Z-score normalization followed by ComBat harmonization dramatically reduced the number of features affected by the "scanner effect" too (Table 4.3). This could be an alternative that would be easy to perform on both training and test sets and would not require to learn landmarks through another training process.

Finally, the harmonization step of the pipeline heavily relies on the use of the ComBat method but this approach requires at least 20 patients per imaging device (or batch) to estimate a distribution of features within it, that would be realigned with the distributions of other centers [11]. These conditions are met in the training set. The test set on the other hand gathers 33 patients coming from more than 15 imaging centers. Other strategies have been pursued to correct the "scanner effect" in the test set and will be introduced in Chapter 6.

Table 4.3: Number of radiomic features in the 5 different ranges of p-values.

p-value	After Z-score & ComBat	After HM & ComBat
T1		
$p < 10^{-5}$	0	0
$10^{-5} \leq p < 10^{-3}$	0	0
$10^{-3} \leq p < 0.01$	0	0
$0.01 \leq p < 0.05$	0	2
$0.05 \leq p$	42	40
T2		
$p < 10^{-5}$	0	0
$10^{-5} \leq p < 10^{-3}$	0	0
$10^{-3} \leq p < 0.01$	1	0
$0.01 \leq p < 0.05$	0	1
$0.05 \leq p$	41	41
T1-weighted DCE		
$p < 10^{-5}$	0	0
$10^{-5} \leq p < 10^{-3}$	0	0
$10^{-3} \leq p < 0.01$	1	0
$0.01 \leq p < 0.05$	0	1
$0.05 \leq p$	41	41

The p-values correspond to Kruskal-Wallis tests between the three coils, radiomic features being extracted from 30 similar regions. Results are given for the three MR sequences. HM:Histogram matching.

Conclusion

This chapter proposed a pipeline dedicated to breast MR imaging, highlighting the need for bias field reduction, normalization and further harmonization of features to reduce the different types of inhomogeneities affecting analyses. Exporting these results to patient images will however need to consider other factors, like the use of contrast agents or potential artefacts not visible on phantoms, and consider the ease of the implementation of the pipeline. It is also important to underscore that this pipeline, though intending to be exportable to all imaging centers, was designed with the imaging devices used in the training set. The following chapter will apply and adapt methods proposed for the phantom images to patient images.

Chapter 5

Handcrafted radiomic analysis pipeline for breast MRI

Preface

This chapter presents the work conducted to export the methods developed on the breast phantoms to patient images. Decisions taken at each step of the global handcrafted radiomic analysis pipeline (Figure 5.1) are first described. Specific points that needed further adaptations for patient images like bias field correction and image normalization are then detailed.

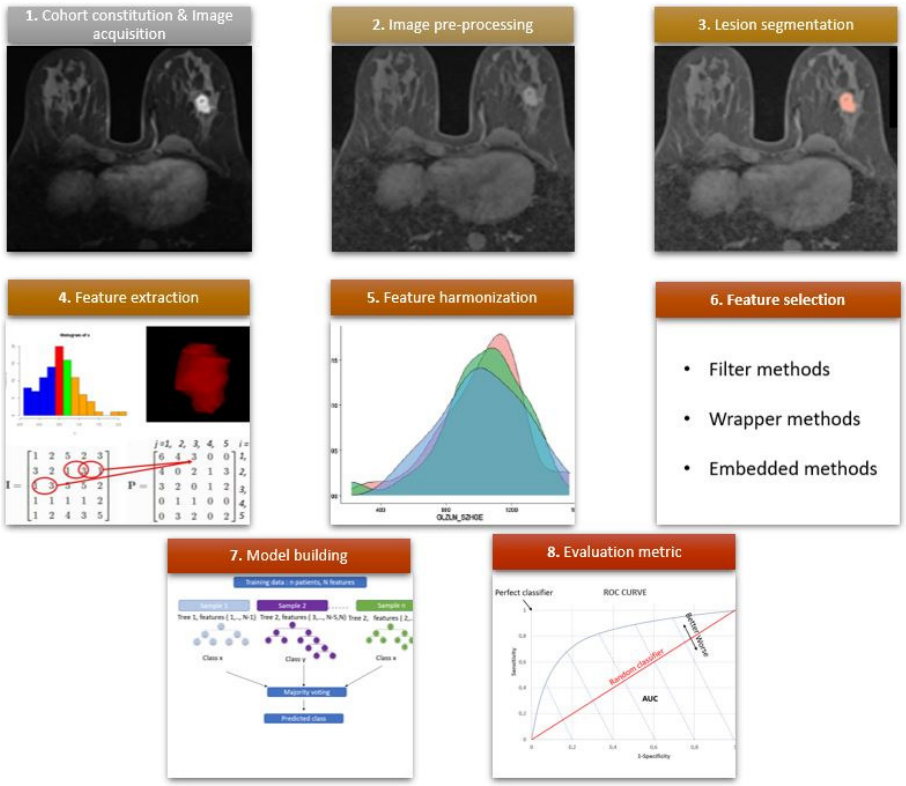


Figure 5.1: Main steps of the handcrafted radiomic analysis pipeline, introduced in Chapter 2.

5.1 General Pipeline

5.1.1 Cohort constitution & Image acquisition

As previously defined in Chapter 3, 136 patients were collected retrospectively, imaged using T1-weighted DCE and fat-saturated T2 sequences and separated between a training set (103/136 patients) and test set (33/136). Training patients were all scanned at Institut Curie while test patients were imaged in a variety of centers (see Table 3.2).

5.1.2 Image pre-processing

Bias field correction

Methods developed to correct bias field gain on breast phantoms as described in Chapter 4 need to be tested and potentially adapted to patient images. Experiments carried out and associated results are detailed in Section 5.2.

Spatial resampling

Spatial resampling is an important step of the radiomic pipeline as inhomogeneities in voxel size in a dataset affect radiomic feature values. The influence of spatial resampling has been the focus of several studies [130, 230, 244].

In this work, T1-weighted DCE images were resampled using B-Spline interpolation to have isotropic voxels (1 mm x 1 mm x 1 mm) as it is preferred to calculate some texture features and it is recommended by the IBSI guidelines [130, 143]. Isotropic resampling has also been associated with a decrease in the number of radiomic features dependent on the magnetic field strength of scanners (1.5 versus 3T) in a study by Um et al. [230].

Furthermore, T2-weighted images were resampled to have voxels of dimensions 0.7 mm x 0.7 mm x 4 mm. These dimensions were chosen to take the mean of the parameters of the different scanners of the training set (Table 3.1).

Normalization

Similarly to bias field correction, the normalization step required further testing on patient images, described in Section 5.3.

5.1.3 Lesion segmentation

Tumors were segmented on the first post-contrast image after injection. Two radiologists, Dr. Caroline Malhaire and Dr. Pia Akl, segmented separately half of the lesions in 3D using the LIFEx software (version 6.0, www.lifexsoft.org) [235]. Thirty tumors of the training set were segmented by both. Both radiologists segmented lesions in the same manner.

To get a more refined delineation of the lesion, segmentations were thresholded using 40% of the maximum tumor intensity value after bias correction as threshold level. These refined segmentations are referred as "Thresholded segmentations" by opposition to the "Full segmentations". Full segmentations were resampled using nearest neighbor interpolation to fit

onto T2-weighted images. Each patient has thus three tumor segmentations as illustrated in Figure 5.2.

Based on the 30 tumors segmented by both radiologists, a mean Dice similarity coefficient of 0.78 $\pm$ 0.10 was reached on the full segmentations. This score increased to 0.88 $\pm$ 0.12 on the thresholded segmentations, showing a good to excellent agreement between radiologists.

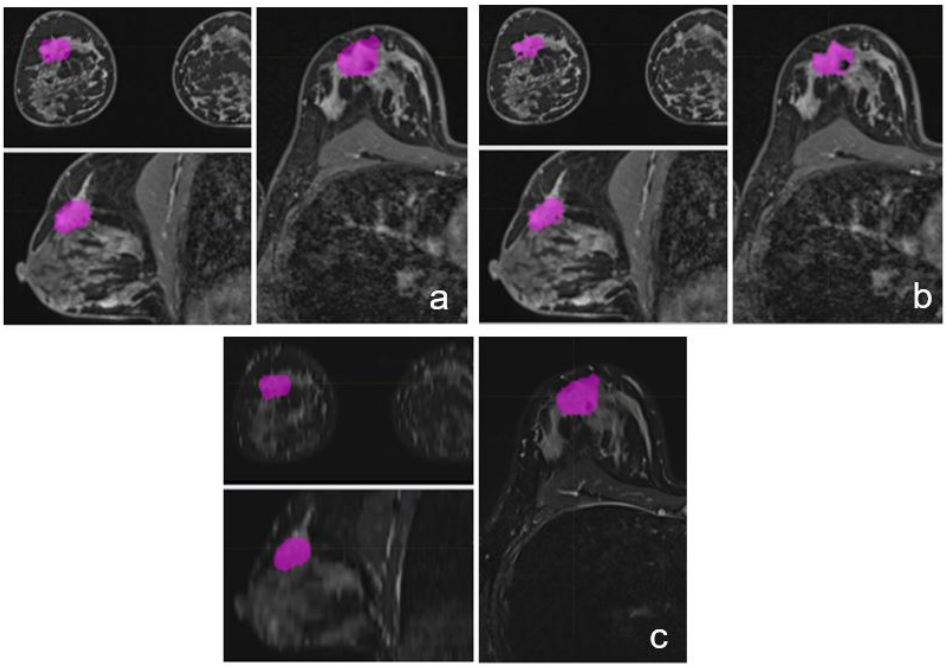


*Figure 5.2: Illustrations of **(a)** the full segmentation delineated by radiologists on the first T1-weighted DCE image after contrast injection; **(b)** the thresholded segmentation on the same T1-DCE image; **(c)** the full segmentation on the fat-saturated T2 image for the same patient.*

Chapter 7 will present an alternate way to get the "full segmentations".

5.1.4 Feature extraction

Shape, first-order statistics and texture features were extracted from the segmentations on the normalized images using the Pyradiomics software (v3.0). Native and wavelet-filtered images (9 images in total) were used conjointly to extract features. The same set of 107 features (14 shape features and 93 first-order and texture features) was extracted from the three segmentations (full and thresholded segmentations on T1-weighted DCE and full segmentations on T2 images) for each patient on the native images. On the wavelet-filtered images, for each of the three segmentations, only the 93 first-order and texture features were extracted. In total, 93x3x9 + 14x3 = 2553 features were extracted for every patient. As advised by Goya-Outi et al. [228], the extraction was carried out using absolute discretization with a fixed bin size of 1 due to the normalization.

5.1.5 Feature harmonization

Features were then harmonized to correct the "scanner effect". On the training set, the conditions to apply the ComBat method were met (minimum of 20 patients per imaging device) [11] and the method subsequently performed. The non-parametric version of ComBat without the empirical Bayes assumption and using three batches corresponding to the three image settings of Institut Curie was carried out. No covariate was added to the model. Table 5.1 records the number of features affected by the "scanner effect" (Kruskal-Wallis $p < 0.05$) before and after harmonization and underscores the reduction of the effect. After harmonization, features were standardized with Z-score.

Table 5.1: Impact of ComBat harmonization on features to reduce the "scanner effect".

Harmonization	Number of features significantly affected by the "scanner effect"
No Harmonization	1580/2553 (61.9%)
ComBat harmonization	99/2553 (3.9%)

As conditions to apply ComBat were not met on the test set since the 33 patients were imaged in 15 different imaging centers, an ad hoc harmonization needs to be developed, that will be presented in Chapter 6.

5.1.6 Feature selection

Feature robustness towards segmentation

Using the 30 tumors of the training set segmented by both radiologists, two-way random intraclass correlation coefficients of radiomic features were calculated. Following Granzier et al. [166] and Saha et al. [148] recommendations, "robust" features were defined as features with ICC > 0.8. Table 5.2 displays the number of "robust" features for each one of the three segmentation forms. Figure 5.3 reports the ICC values obtained for the features depending on the modality, the filter applied on the images and the type of features. No particular group of features could be easily identified as "robust" as the influence of the filter applied to the image was important but the features from the "GLZSM" matrix seemed particularly affected by segmentation variabilities. Features that did not qualify as "robust" were discarded and further selection applied only to "robust" features.

Table 5.2: Features deemed robust according to the modality.

Modality	Number of features with ICC>0.8
Full segmentation T1-weighted DCE	692/2553 (27.1%)
Thresholded segmentation T1-weighted DCE	668/2553 (26.2%)
Full segmentation fat-saturated T2	492/2553 (19.3%)

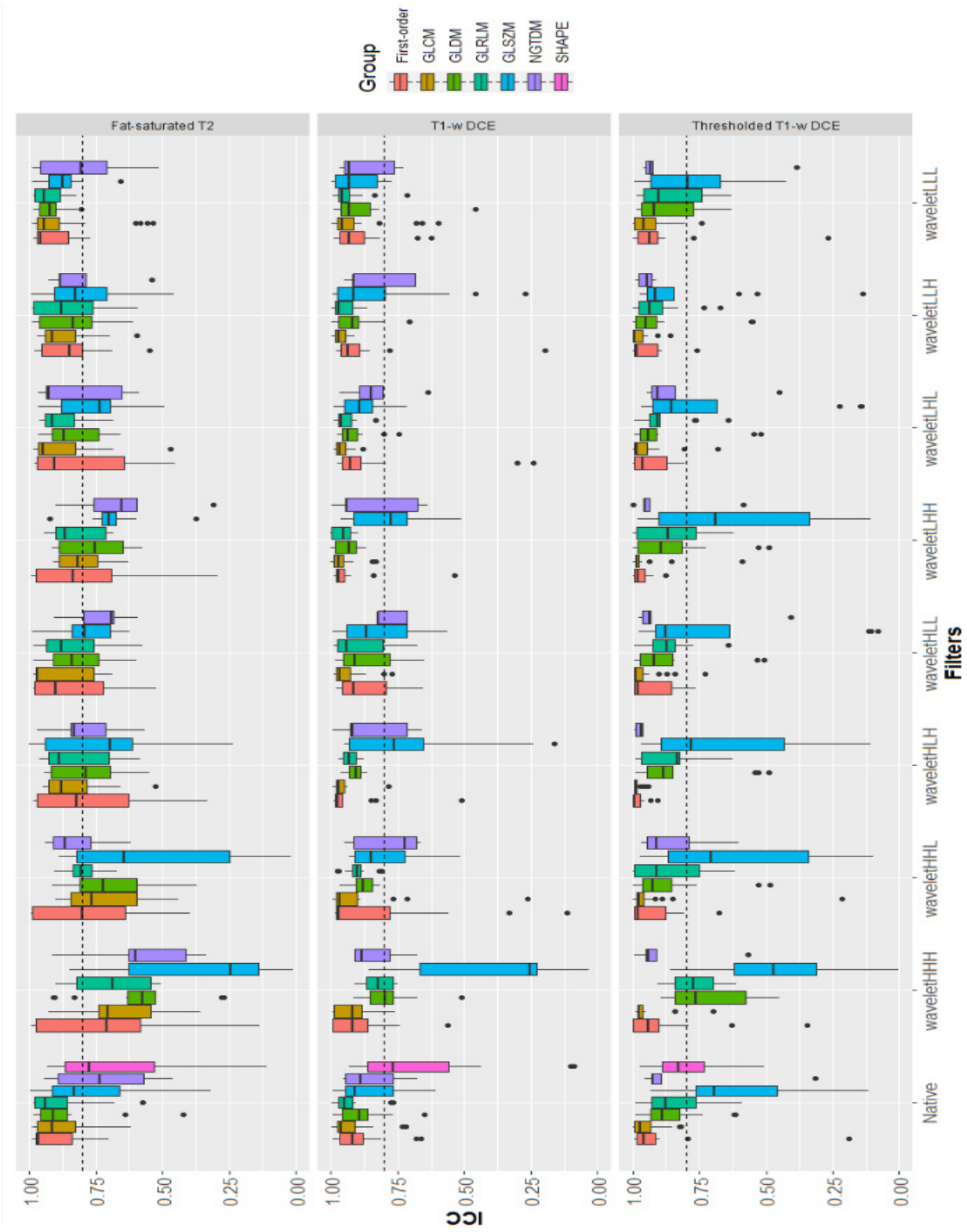


Figure 5.3: Intraclass correlation coefficients (ICC) for every group of features and filter applied on images. Results are represented separately for each segmentation (DCE, thresholded DCE, T2). Dashed lines represent the 0.8 cut-off that select features robust to the segmentation.

Feature reduction & further selection

As a large number of features remained after discarding features not robust to the segmentation, a first univariate filter step was carried out to select the most interesting features with respect to pCR in the training set. Thus, features whose lower AUC bound was inferior to 0.5 were discarded. Then to remove redundancy, highly correlated features, using a Spearman's rank correlation coefficient cut-off of 0.8, were removed.

Feature selection was finally performed on the decorrelated set of features where using a 100 rounds of the Boruta algorithm, the five most frequently selected features were used to build the predictive models.

5.1.7 Model building

Random forest models were predominantly used in our pipeline. They were tuned with leave-one-out cross-validation on the training set for each experiment.

5.1.8 Model evaluation

Performances were evaluated with two different metrics: the Youden index and the AUC. The AUC gives a certain leeway on the test set by not having a determined classification threshold. The Youden index, on the other hand, has a fixed threshold and can be used when the cost of wrongly predicting an observation as positive or negative is comparable.

5.2 Image pre-processing: Bias field correction

5.2.1 Introduction

The phantom experiments concluded that the N4 algorithm applied with five resolution levels, 50 iterations and the use of a full mask of the breast and posterior regions, achieved the best performances in reducing the bias field in the three experimental settings that we tested.

5.2.2 Methods

Images of the training and test sets were segmented to get a full mask of the breast region, using mean thresholding and morphological closing operations, and corrected with the N4 algorithm tuned with the parameters defined on the phantoms.

To measure quantitatively the effects of bias field correction, the coefficient of variation (CV) of the intensities inside a reference tissue was calculated for each modality. CV is defined as

$$CV = \frac{standardDeviation}{mean} \tag{5.1}$$

As lower CV values are associated with a more homogeneous tissue in terms of intensities, changes in CV before and after bias correction were analyzed.

In T1-weighted DCE images, the enhanced subcutaneous fat layer was selected as reference tissue and segmented in 3D. As images acquired on the GE scanner were particularly noisy, a

first denoising step using median filtering with a kernel size of 3 was performed to facilitate the segmentation process. The subcutaneous fat layer was then segmented using successively hysteresis thresholding, with high and low thresholds respectively set to the mean of the image and 40% of the mean of the image, morphological closing operations using the default kernel with square connectivity equal to 1, and then distance transforms selecting all voxels whose values were inferior to 4. Figure 5.4 depicts the segmentation of the fat layer in T1-weighted DCE images.

In fat-saturated T2 images, several slices of the sternum were segmented manually by radiologists (Figure 5.5) to be used as reference tissue. Paired Wilcoxon tests were carried out to compare distributions of CV before and after correction.

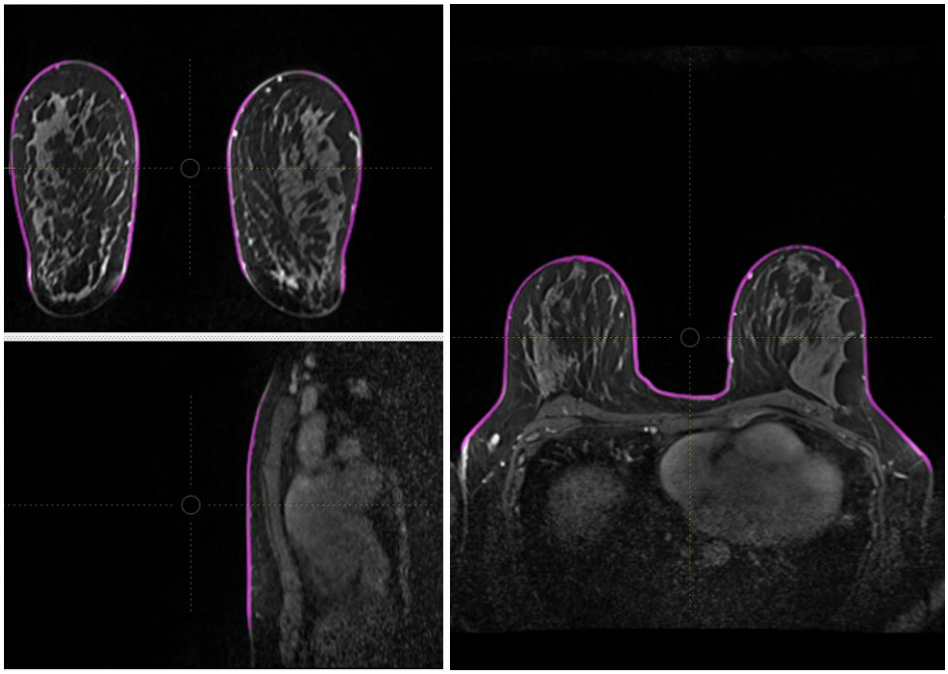


Figure 5.4: Segmentation (in pink) in the axial, sagittal and coronal planes of the subcutaneous fat layer of the breasts of a patient with invasive breast cancer.

5.2.3 Results

Figures 5.6 and 5.7 depict the raw images, the full mask segmented, the estimated bias field calculated overlaid on the raw images and the bias-corrected images for respectively T1-weighted DCE and fat-saturated T2 images of three different patients of the training set imaged with the three training devices. Figure 5.8 and Figure 5.9 show the effect of bias correction in the axial, coronal and sagittal planes for each modality.

Figure 5.10 and Figure 5.11 show four examples of bias fields estimated during the correction step for each of the three imaging devices of Institut Curie for both T2 and DCE modalities. It helps get a global visual pattern of the bias field obtained for each device, though as previously explained, bias field is also heavily influenced by the anatomy and position of the patient imaged.

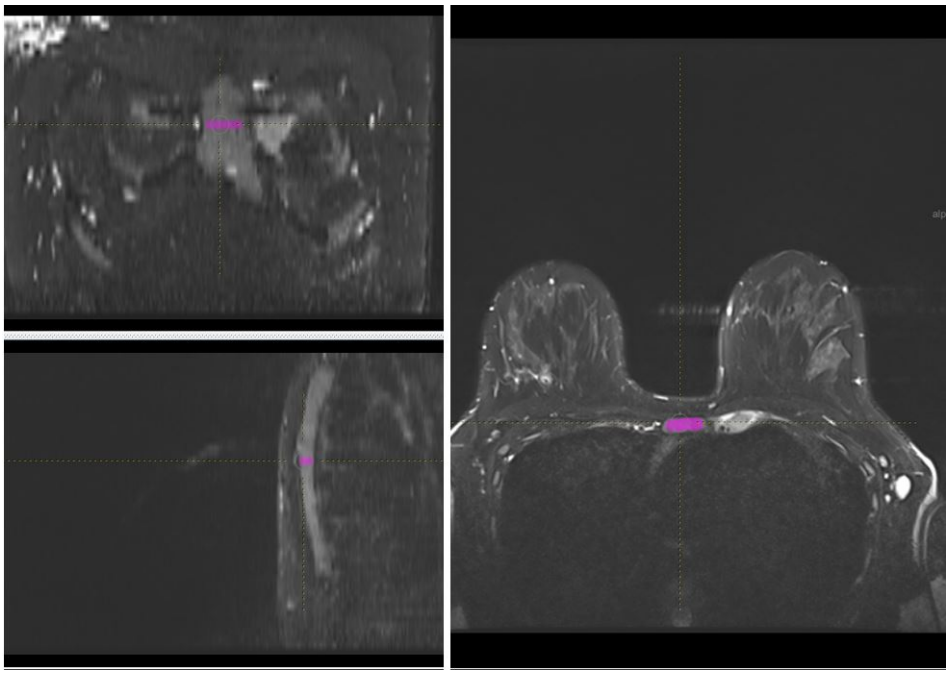
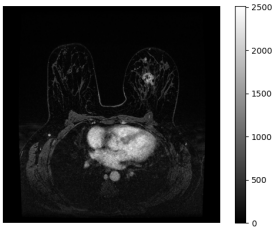


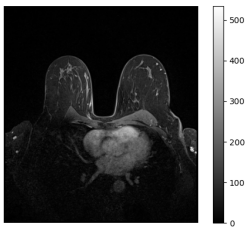
Figure 5.5: Segmentation (in pink) in the axial, sagittal and coronal planes of the several slices of the sternum of the same patient as Figure 5.4.

Figure 5.12 shows the correction pipeline in both modalities on a patient of the test set, imaged in a private imaging center on a DISCOVERY MR750 (GE), 3T with an HD breast coil.

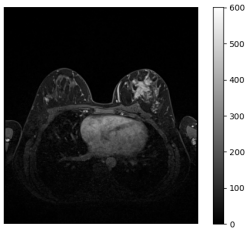
Figure 5.13 and Figure 5.14 show the significant decrease of the coefficient of variation in the fat and sternum of patients after bias correction. For the T1-weighted DCE images, the decrease is illustrated both in the training and test set. For T2-images, segmentations of the sternum were only available for 75 patients of the training set and changes of CV were thus compared on this subset of patients.



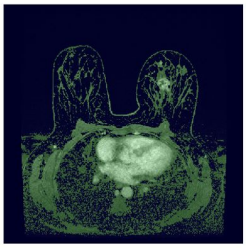
(a) Raw image



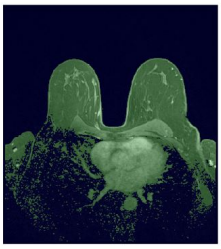
(b) Raw image



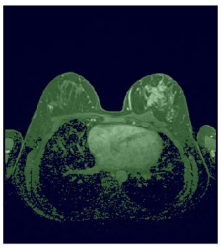
(c) Raw image



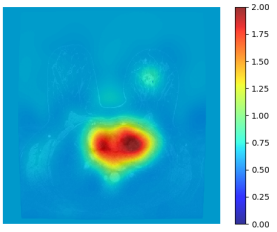
(d) Full mask (light green)



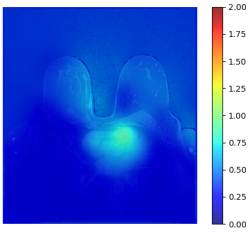
(e) Full mask (light green)



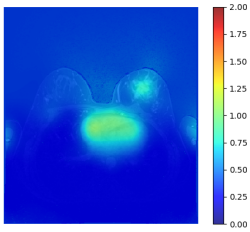
(f) Full mask (light green)



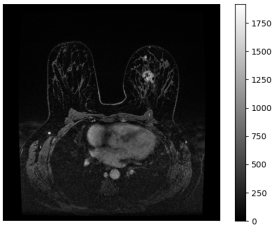
(g) Bias field overlaid on image



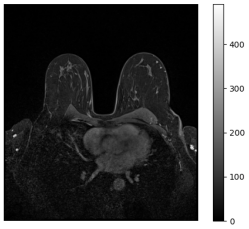
(h) Bias field overlaid on image



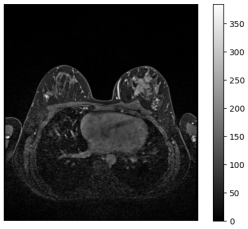
(i) Bias field overlaid on image



(j) Corrected image

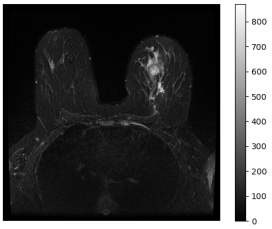


(k) Corrected image

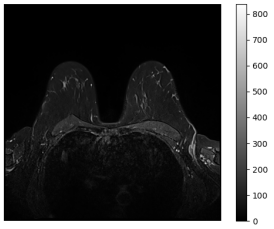


(l) Corrected image

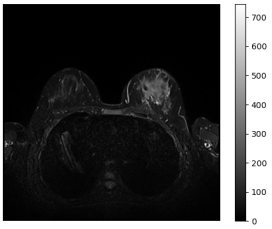
Figure 5.6: Column-wise illustrations in the axial plane of raw T1-weighted DCE images (1^{st} row), full masks (2^{nd} row), estimated bias field gain overlaid on raw images (3^{rd} row) and corrected images (4^{th} row) from patients scanned on the GE machine (Patient 12, 1^{st} column); the Siemens machine with the Sentinelle coil (Patient 1, 2^{nd} column) or with the 18-channel coil (Patient 4, 3^{rd} column).



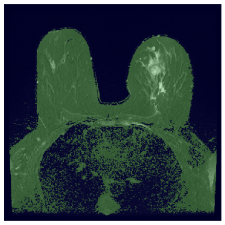
(a) Raw image



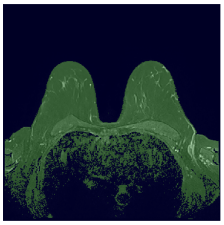
(b) Raw image



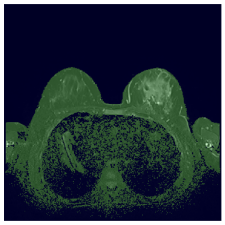
(c) Raw image



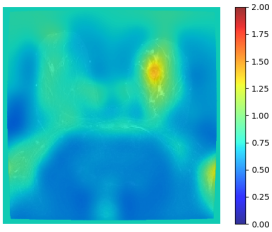
(d) Full mask (light green)



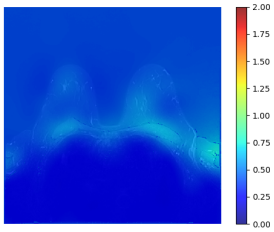
(e) Full mask (light green)



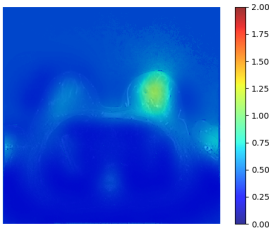
(f) Full mask (light green)



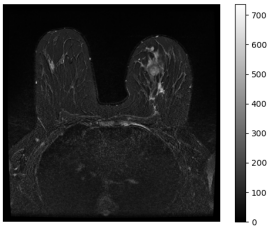
(g) Bias field overlaid on image



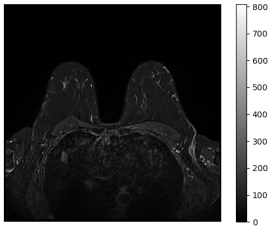
(h) Bias field overlaid on image



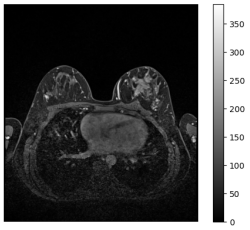
(i) Bias field overlaid on image



(j) Corrected image

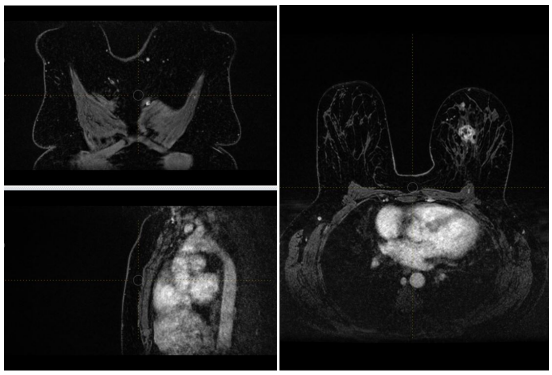


(k) Corrected image

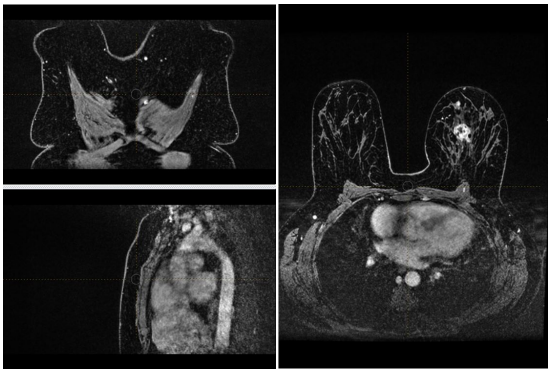


(l) Corrected image

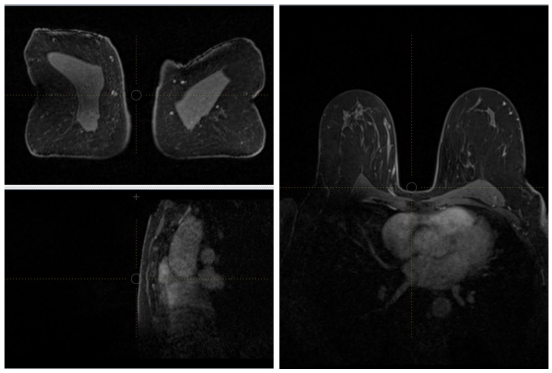
Figure 5.7: Column-wise illustrations in the axial plane of raw fat-saturated T2 images (1^{st} row), full masks (2^{nd} row), estimated bias field gain overlaid on raw images (3^{rd} row) and corrected images (4^{th} row) from patients scanned on the GE machine (Patient 12, 1^{st} column); the Siemens machine with the Sentinelle coil (Patient 1, 2^{nd} column) or with the 18-channel coil (Patient 4, 3^{rd} column).



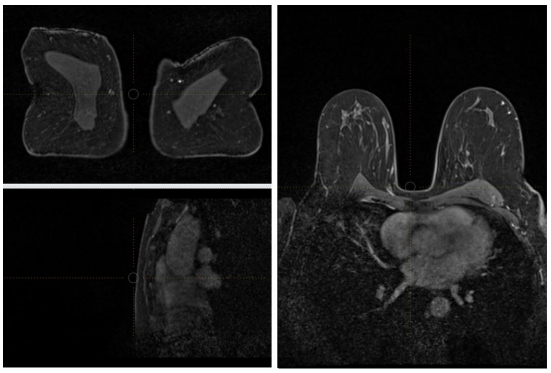
(a) Raw image (GE scanner)



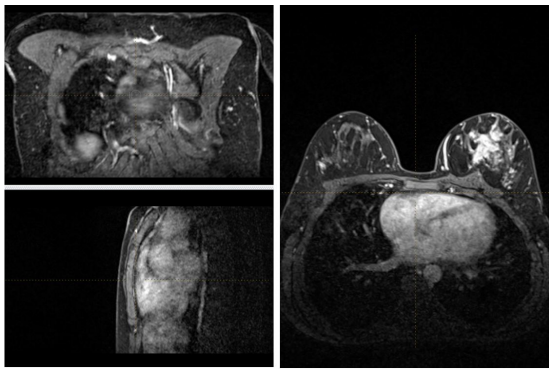
(b) Corrected image (GE scanner)



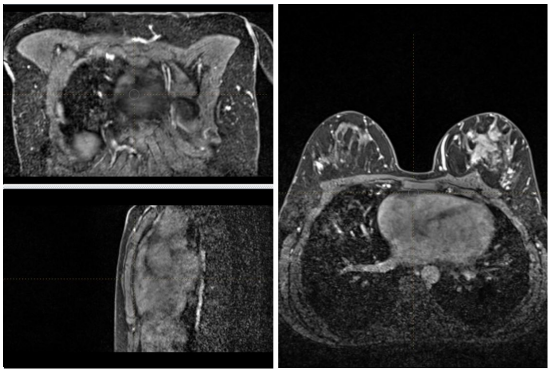
(c) Raw image (Siemens, Sentinelle coil)



(d) Corrected image (Siemens, Sentinelle coil)

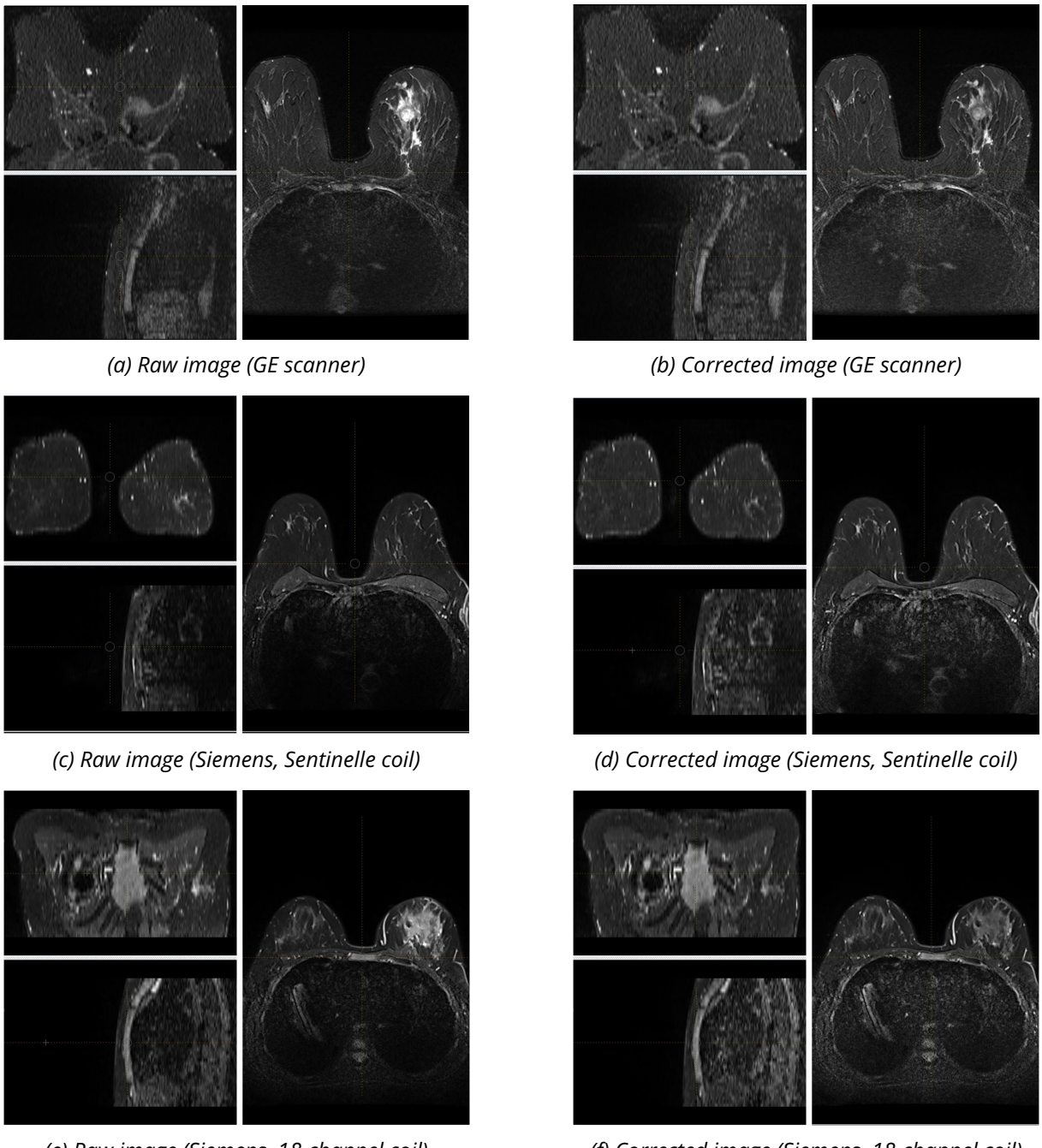


(e) Raw image (Siemens, 18-channel coil)



(f) Corrected image (Siemens, 18-channel coil)

Figure 5.8: Illustrations in the axial, sagittal and coronal planes of raw (1^{st} column) and corrected (2^{nd} column) T1-weighted DCE images from patients scanned on the GE machine (Patient 12, 1^{st} row); the Siemens machine with the Sentinelle coil (Patient 1, 2^{nd} row) or with the 18-channel coil (Patient 4, 3^{rd} row).



(a) Raw image (GE scanner)

(b) Corrected image (GE scanner)

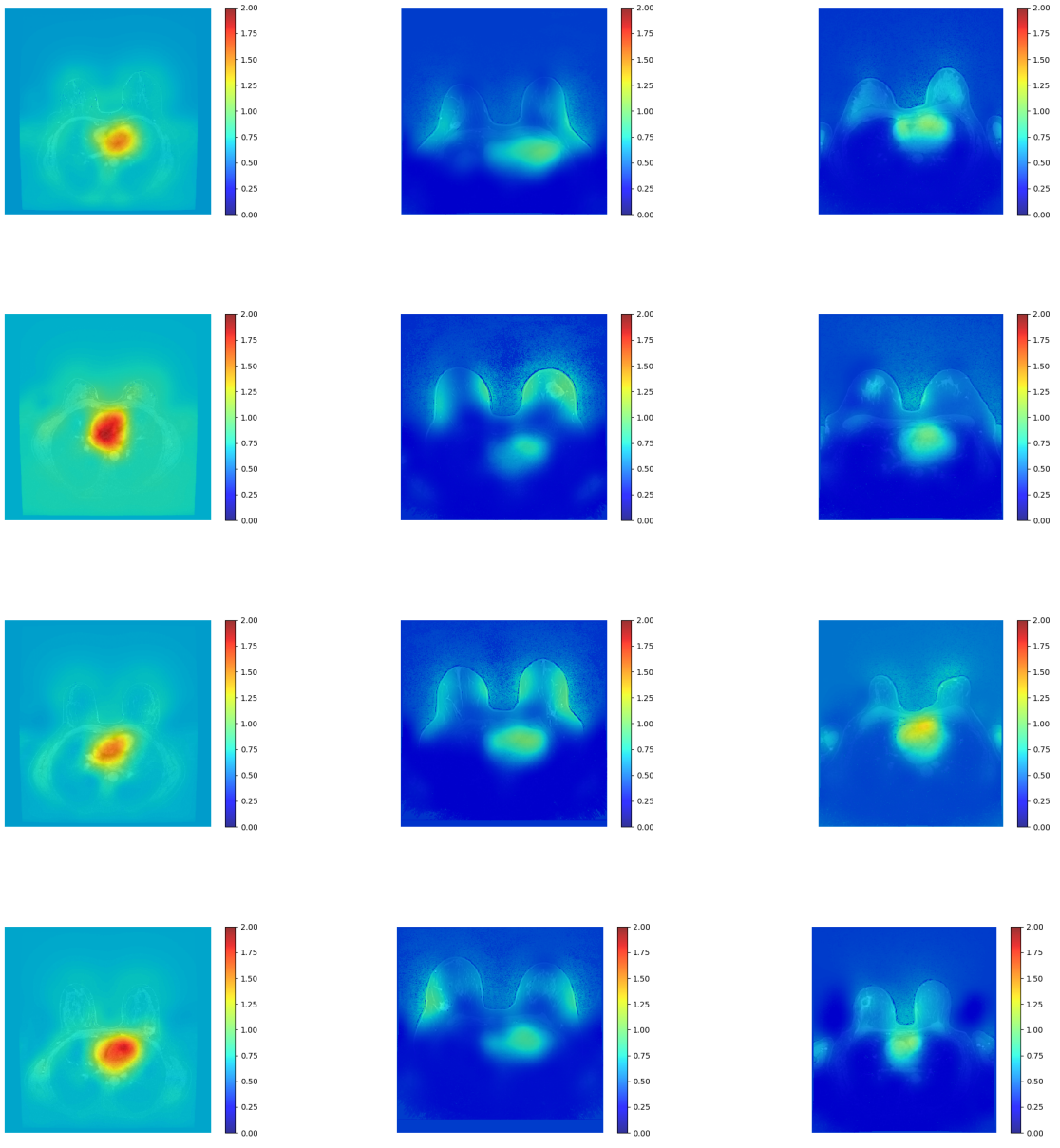
(c) Raw image (Siemens, Sentinelle coil)

(d) Corrected image (Siemens, Sentinelle coil)

(e) Raw image (Siemens, 18-channel coil)

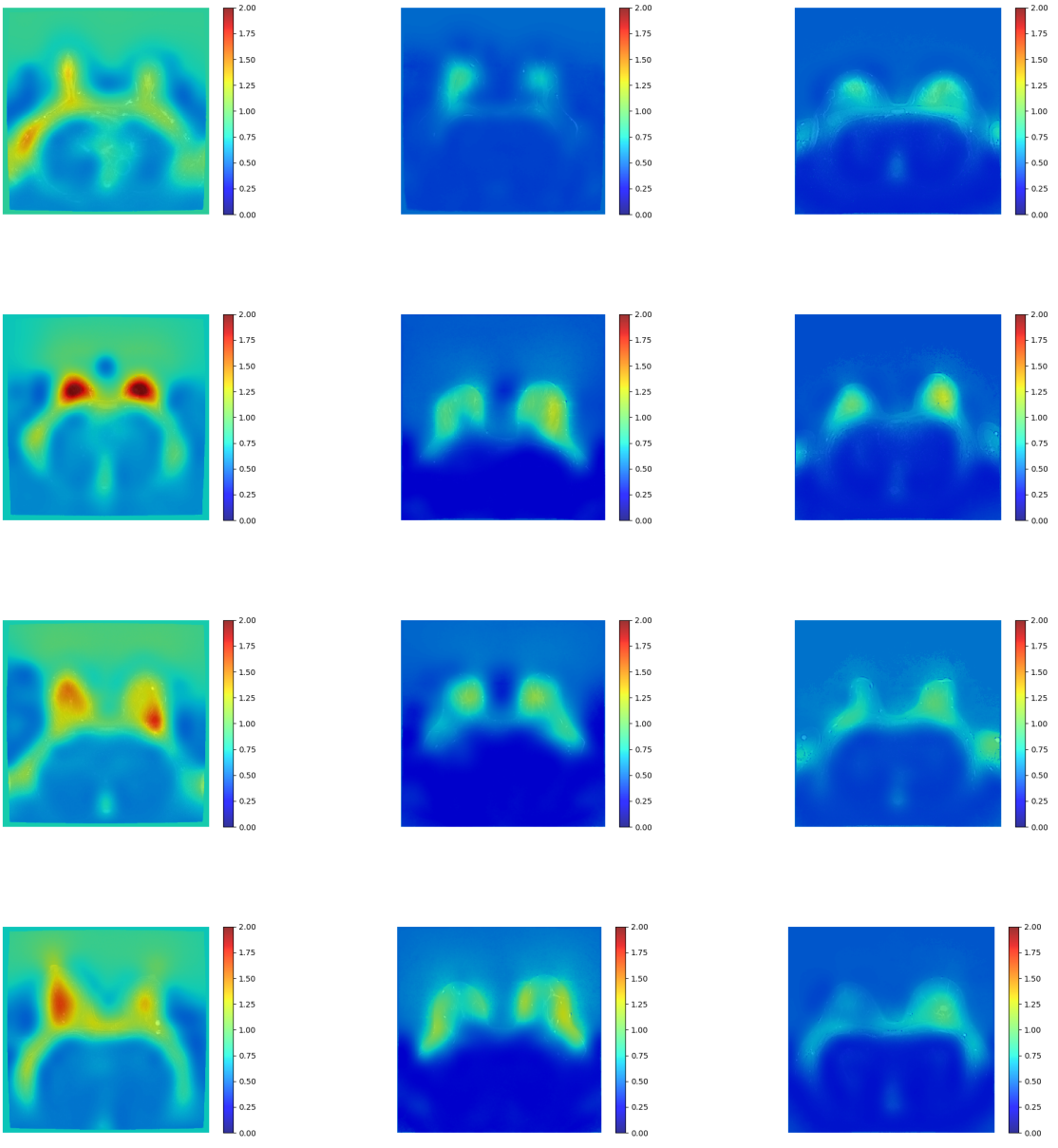
(f) Corrected image (Siemens, 18-channel coil)

Figure 5.9: Illustrations in the axial, sagittal and coronal planes of raw (1^{st} column) and corrected (2^{nd} column) fat-saturated T2 images from patients scanned on the GE machine (Patient 12, 1^{st} row); the Siemens machine with the Sentinelle coil (Patient 1, 2^{nd} row) or with the 18-channel coil (Patient 4, 3^{rd} row).



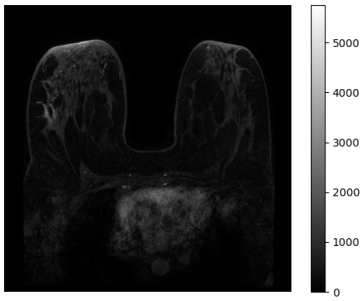
(a) GE scanner *(b) Siemens, Sentinelle coil* *(c) Siemens, 18-channel coil*

Figure 5.10: Column-wise illustrations in the axial plane of four examples of the estimated bias field calculated by the N4 algorithm on T1-weighted DCE images from four different patients scanned on the GE machine (1^{st} column), the Siemens machine with the Sentinelle coil (2^{nd} column) and the Siemens machine with the 18-channel coil (3^{rd} column).

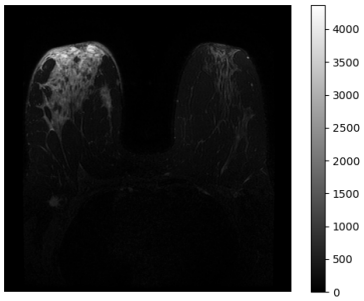


(a) GE scanner (b) Siemens, Sentinelle coil (c) Siemens, 18-channel coil

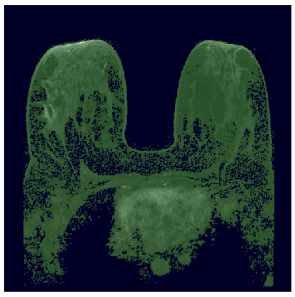
Figure 5.11: Column-wise illustrations in the axial plane of four examples of the estimated bias field calculated by the N4 algorithm on fat-saturated T2 images from four different patients scanned on the GE machine (1^{st} column), the Siemens machine with the Sentinelle coil (2^{nd} column) and the Siemens machine with the 18-channel coil (3^{rd} column).



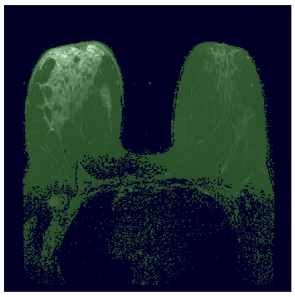
(a) T1-DCE raw image



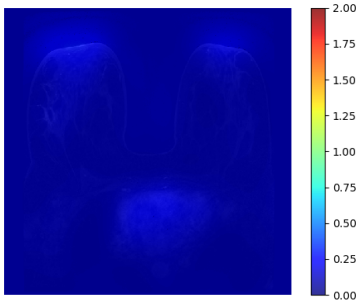
(b) T2 raw image



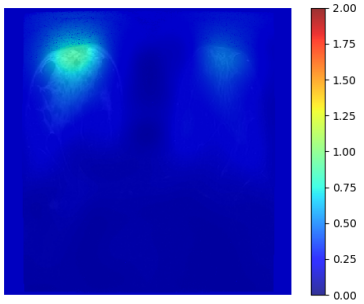
(c) Full mask (light green)



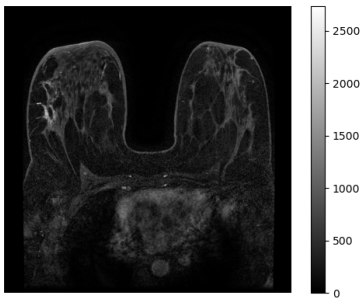
(d) Full mask (light green)



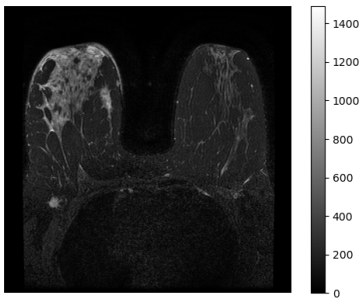
(e) Bias field overlaid on image



(f) Bias field overlaid on image

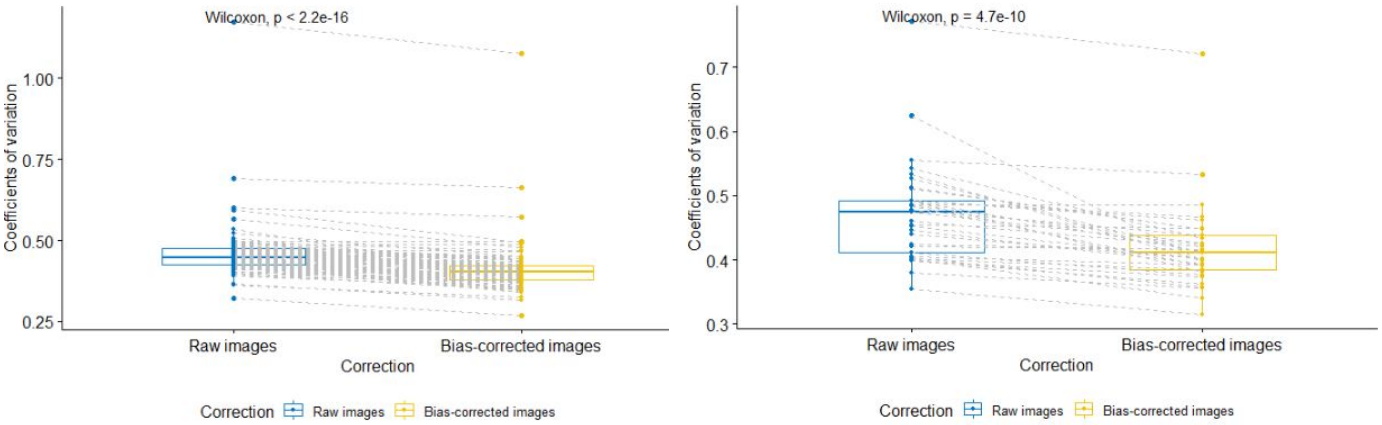


(g) Corrected image



(h) Bias field overlaid on image

Figure 5.12: Column-wise illustrations in the axial plane of raw T1-DCE and fat-saturated T2 images (1^{st} row), full masks (2^{nd} row), estimated bias field gain overlaid on raw images (3^{rd} row) and corrected images (4^{th} row) from test set patient 107, imaged on the DISCOVERY MR750 (GE, 3T) with an HD breast coil.



(a) Training test *(b) Test set*

*Figure 5.13: Boxplots of coefficients of variations calculated in the segmented subcutaneous fat layer of the breasts in raw and bias corrected images of **(a)** the training set; **(b)** the test set. Paired Wilcoxon test was performed. Dashed lines connects points corresponding to the same patients. Median of CV on raw images is 0.4480 and 0.4017 on corrected images of the training set. On the test set, the median of CV on raw images is 0.4744 and 0.4105 on corrected images*

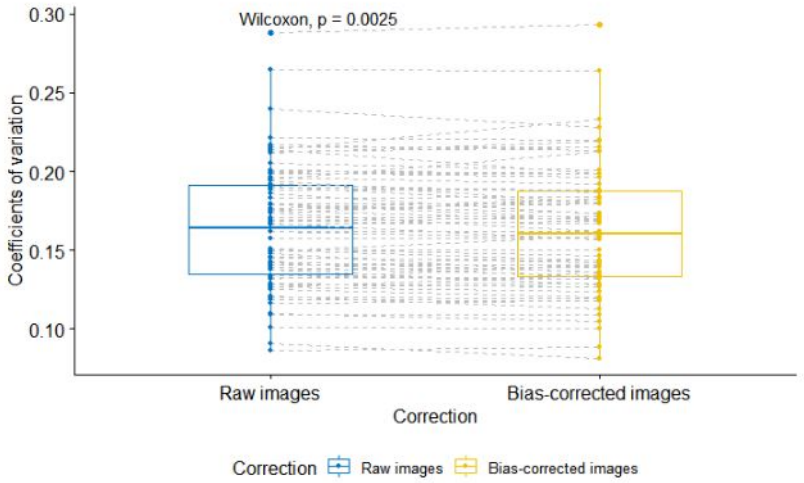


Figure 5.14: Boxplots of coefficients of variations calculated in the segmented sternum in raw and bias corrected images on the training set. Paired Wilcoxon test was performed. Dashed lines connects points corresponding to the same patients. Median of CV on raw images is 0.1644 and 0.1605 on corrected images.

5.2.4 Discussion & Conclusion

The first noteworthy difference observed between patient and phantom images concerns the GE scanner. The contrast observed on T1-weighted DCE images acquired on the GE machine is much stronger, making the parenchymal tissue appears almost as dark as the background. The automatic pipeline developed to segment the full mask of the breasts, consisting of a succession of thresholding and filling steps using morphological operations thus did not capture the parenchyma completely unlike in patients imaged in the other settings (Figure 5.6). No such effect was observed in T2 images (Figure 5.7). The segmented mask was still nevertheless much more complete than the default one of the N4 algorithm obtained by Otsu thresholding. After visual and quantitative comparisons between the results obtained with both masks, as no changes were observed, it was decided to apply the automatic segmentation pipeline. There was no particular problem when applying the segmentation pipeline on the test set despite the variety of scanners.

Figure 5.13 and Figure 5.14 shows the global decrease in CV, which is steep in the fat and more moderated in the sternum. The choice of a few slices of the sternum as a reference tissue to analyze the impact of the correction may not be the best as few voxels were considered and were nearly all at the same location. According to the Wilcoxon signed rank tests, distributions of CV before and after correction were significantly different ($p < 0.01$) in both modalities attesting that the correction carried on reduced local spatial inhomogeneities.

Corrected images of the training and test set were also checked one by one in the axial, sagittal and coronal planes and the estimated bias fields were visualized. As noticed in the phantoms, general patterns of the bias field were specific to the coils and scanners used and differences were observed between modalities (Figures 5.10 and 5.11). In T1-weighted DCE, a major difference with the phantom experiments is the presence in the all imaging settings of a physiological strong enhancement of the breast and the tumor region due to the injection of the contrast agent. This physiological enhancement must not be overcorrected in the clinically relevant zone of the MR images (breast). In the patient dataset, the N4 algorithm with the parameters selected on the phantoms was able to reduce local inhomogeneities in images while keeping a specific enhancement of the tumor region that is still highly noticeable after correction (Figure 5.8).

5.3 Image pre-processing: Normalization

5.3.1 Introduction

As pointed out in the phantom experiment, MR images must be normalized to ensure the reproducibility of radiomic features and it is especially important in MR imaging because of the absence of standard units.

Though an abundant literature exists on the normalization of MR brain images [228], there is however no clear consensus today about the best approach to standardise images in breast MRI. Among the 36 identified papers about handcrafted radiomics for neoadjuvant chemotherapy in breast cancer (Table 2.3), only five studies reported performing image normalization. Two studies performed histogram matching (HM) and three of them Z-score normalization. In a recent paper about the reproducibility of features extracted from brain MR images using both phantom and patient data, Li et al. [244] found that though bias field correction and intensity normalization did not remove the "scanner effect", image pre-processing did improve the robustness of radiomic features. Li et al. notably compared Z-score, White-Stripe (WS) normalizations and histogram matching and found no clear advantage of one method over the others. Reviewing studies about intensity standardisation of MR images before feature extraction in gliomas, Fatania et al. [245] underscored the rarity of studies comparing several types of normalization even in the global brain MR imaging field, which is much more developed than breast MR imaging. They nevertheless noticed a rise in deep learning-based normalization approaches. Finally, Destito et al. [246] compared radiomic features extracted from a sphere in healthy brain tissue in repeated acquisitions of the same group of patients on different scanners using several normalization techniques (no normalization, Z-score, HM, WS). They found that features extracted from Z-score normalized images displayed higher ICC and were thus deemed more reproducible.

5.3.2 Methods

Like in the phantom experiments, MR images were first corrected for bias field gain and spatially resampled before normalization. Two types of normalization were tested: Z-score and histogram matching.

Histogram matching

Histogram matching aligns intensity densities of raw images by matching predefined histogram landmarks of the images to the landmarks of a histogram template in a piecewise linear manner (Figure 5.15). As explained in Chapter 4, a histogram template can be provided, or landmarks can be learned through a training process on a set of images. On the phantom images, we used decile quantiles as landmarks, which resulted in a good alignment of intensity distributions between scanners. In patient images, the use of so many landmarks could prove too strong and reduce inter-subject variabilities and in the process remove useful information, which is a concern about the histogram matching approach in general [247]. Therefore, we decided to use a single landmark corresponding to the median intensity as defined by Nyul et al. [21]. As histogram matching was originally developed for a multiple sclerosis study, the influence of the presence of a large tumor in the image on the calculation of landmarks has

been discussed in [24]. Isaksson et al. [24] notably found, in a study on the impact of image normalization on prostate radiomics, that calculating landmarks for histogram matching in the healthy prostate tissue gave better results than taking the whole prostate tissue including the tumor area. Though the ratio of tumor volume to healthy tissue volume is generally smaller in breast, a single landmark corresponding to the median of healthy breast tissue intensity was selected for our study and learned on the training set.

The training process of histogram matching starts by segmenting the body of the patients using thresholding with the mean image intensity and morphological closing operations. The background is left aside during the whole procedure. A lower p_1 and upper p_2 percentiles in the body segmentation are selected to remove outliers. For every image i of the training set, lower p_{1i} and upper p_{2i} values are then mapped to s_1 and s_2, the minimum and maximum intensities of the standard scale. Intensities x of [p_{1i}, p_{2i}] are mapped linearly to x' of [s_1, s_2] following the below formula:

$$x' = s_1 + \frac{x - p_{1i}}{p_{2i} - p_{1i}}(s_2 - s_1) \tag{5.2}$$

Selected landmarks are then mapped to the standard scale. Final landmarks are obtained by averaging landmark values of all training images on the standard scale. In our case a single landmark μ_s corresponding to the median intensity of healthy breast tissue was mapped as illustrated in Figure 5.15.

Once the landmarks defined, the images are normalized by matching their intensities to the standard scale using several piecewise linear mappings. In the case of a single landmark μ_s, intensities are mapped according to two transformations. Let be x' the mapped intensity on the standard scale of intensity x of image i,

$$x' = \begin{cases} \mu_s + \frac{s_1 - \mu_s}{p_{1i} - \mu_i}(x - \mu_i), & \text{if } x \leqslant \mu_i \\ \mu_s + \frac{s_2 - \mu_s}{p_{2i} - \mu_i}(x - \mu_i), & \text{otherwise.} \end{cases} \tag{5.3}$$

Z-score normalization

Similarly, Z-score parameters (mean and standard deviation) were calculated in the healthy breast tissue, taking a full mask of the breast using mean thresholding, closing operations and cropping while removing the previously segmented tumors.

Quantitative & Qualitative assessment

There is no golden standard to evaluate the impact of normalization, which can be assessed in multiple manners. First, qualitative analysis can be carried on by visualizing the alignment of intensity densities of tissues according to the methods used. Quantitative and statistical analysis can also be performed. The effects of normalization could also be studied through the perspective of radiomics. Isaksson et al. [24] calculated the concordance correlation coefficient (CCC) between features extracted before and after normalization. Their assumption was that normalization techniques that would not overcorrect the images would preserve the

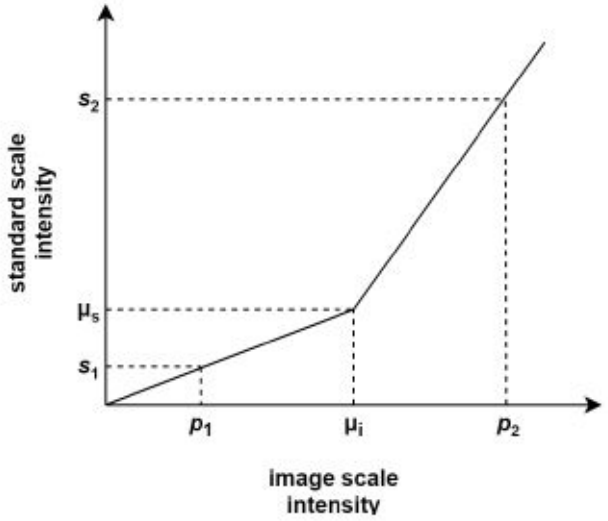


Figure 5.15: Mapping process of landmark μ_i of the image scale to μ_s on the standard scale. p_1 and p_2 are the minimum and maximum intensities on the image scale and s_1 and s_2 their equivalent on the standard scale. Source: Nyul et al. [21].

concordance between raw and normalized features. The CCC of two features x and y can be defined as

$$\rho_c = \frac{S_{xy}}{{\sigma_x}^2 + {\sigma_y}^2 + (\mu_x - \mu_y)^2} \tag{5.4}$$

where μ_x, μ_y and σ_x, σ_y are their respective mean and standard deviations and S_{xy} their covariance. In their study, Isaksson et al. [24] defined a cut-off of 0.8 to establish the concordance of features.

Finally, the impact of normalization can be assessed for our specific clinical task, i.e. predicting pCR to NAC, by comparing performances obtained in each case. This idea was carried out by counting, for each normalization approach, the number of features whose AUC lower bound were strictly superior to 0.5 when predicting the response to NAC. No correction for multiple testing was integrated in this experiment.

5.3.3 Results

Qualitative assessment

Figure 5.16 and Figure 5.17 show the alignment of distributions in T1-DCE and T2 modalities in the healthy breast tissue on one hand and tumor region on the other hand for the different normalizations.

Quantitative & Statistical assessment

Table 5.3 and Table 5.4 summarize the results obtained by calculating the concordance of radiomic features extracted from native and wavelet-filtered normalized images and their association with pCR to evaluate the different types of normalizations.

Table 5.3: Concordance correlation coefficients (CCC) of radiomic features depending on normalizations.

Normalization type	Number of features with CCC>0.8
Histogram matching	196/2553 (7.7%)
Z-score	248/2553 (9.7%)

CCC: concordance correlation coefficient.

Table 5.4: Association of radiomic features with pCR depending on normalizations.

Normalization type	Number of features with $AUC_{LowerBound}$>0.5
No normalization	84/2553 (3.2%)
Histogram matching	93/2553 (3.6%)
Z-score	125/2553 (4.9%)

$AUC_{LowerBound}$ *lower bound of the area under the receiver-operating characteristic curve.*

5.3.4 Discussion & Conclusion

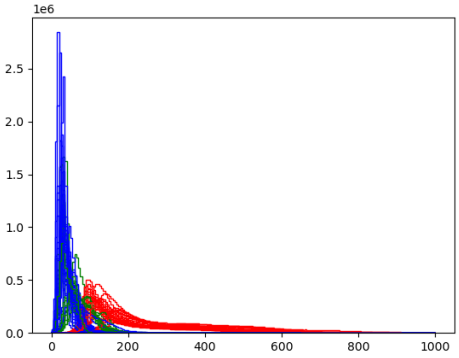
The figures showed in both modalities a much better alignment across scanners of the intensity densities in the healthy breast tissue than in the tumor area, in which results were very heterogeneous. Unlike in phantom experiments, the use of a single landmark for histogram-matching resulted in much noisier histograms. The alignment seemed slightly better with the Z-score normalization but there was no glaring difference between the two approaches.

Radiomic analyses highlighted that with Z-score normalization a 34% increase (Table 5.4) in the number of features with $AUC_{LowerBound}$>0.5 could be obtained compared to histogram matching. Regarding the concordance correlation calculation, several remarks must be made. First, features were globally heavily affected by the normalizations as an extremely small percentage of features (9.7%) was deemed robust which must lead to reassess the assumption of concordance between raw and normalized features. The differences between the two normalization types were also very small. Finally, the CCC could be affected by the low variance of some features and the parameters (bounds and bins) used to calculated features.

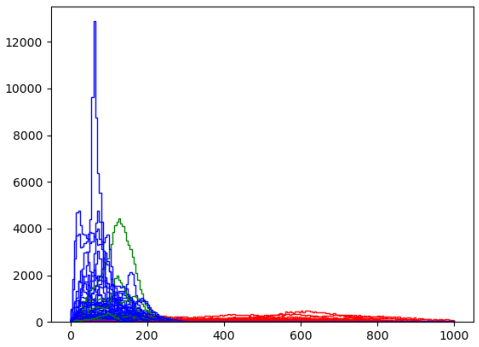
Therefore, taking into account the previous results and considering its ease of use, Z-score normalization in the healthy breast tissue was performed on our training and test sets.

Conclusion

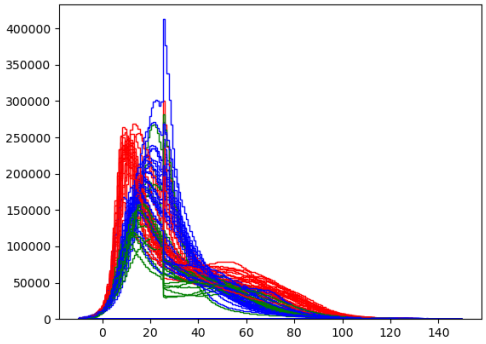
This chapter adapted the pipeline defined on the phantoms to patient images, highlighting the difficulties to go from the simple model of the phantoms to the much more complex patient images. Figure 5.18 summarizes all the parameters defined to build robust radiomic models in the context of a multicentric dataset. The remaining point that needs to be further investigated concerns the harmonization of radiomic features of the test set, where no conventional method can be used due to its small size. An original strategy will be developed in Chapter 6 to address this issue.



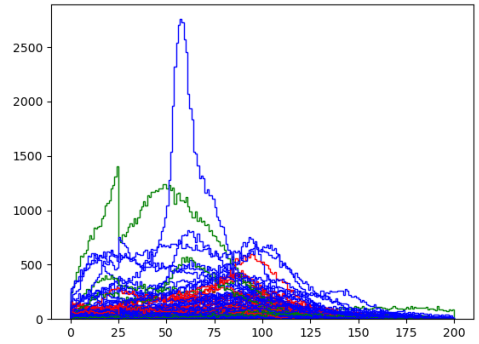
(a) Healthy breast tissue: BC image



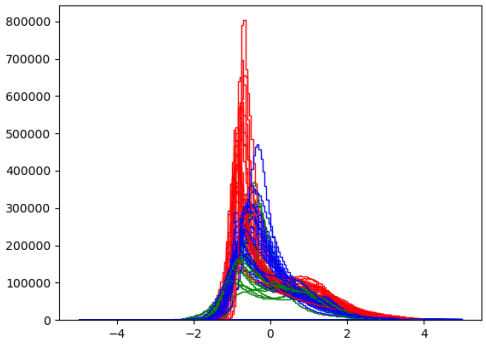
(b) Tumor: BC image



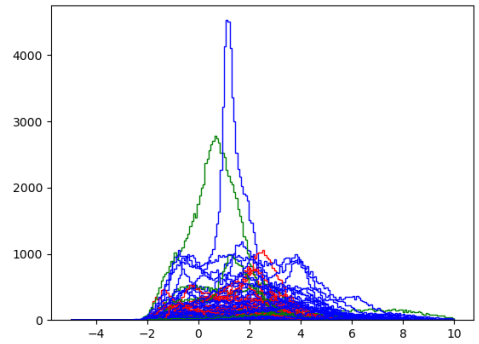
(c) Healthy breast tissue: HM-BC image



(d) Tumor: HM-BC image

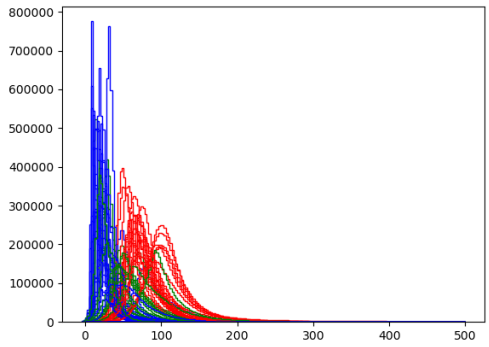


(e) Healthy breast tissue: Z-score BC image

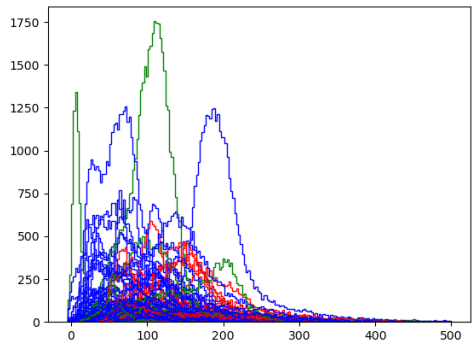


(f) Tumor: Z-score BC image

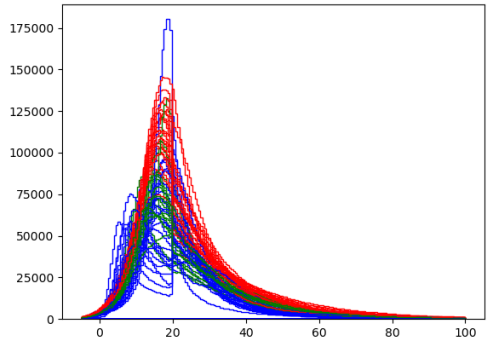
*Figure 5.16: Intensity histograms of healthy breast and tumor tissue from raw (1^{st} row), bias-corrected (BC) (2^{nd} row) and histogram-matched (HM) (3^{rd} row) or Z-score normalized BC (4^{th} row) **T1-DCE** images of the training set. **RED color** is used for patients scanned on the GE machine; **BLUE color** for those imaged on the Siemens machine with the Sentinelle coil and **GREEN color** when the 18-channel coil was used.*



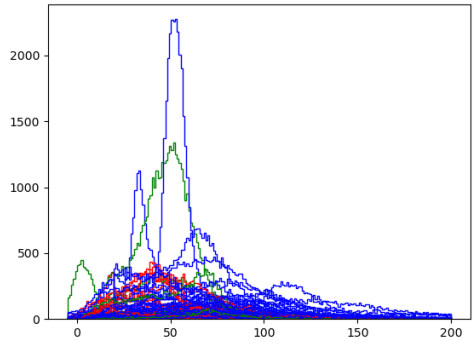
(a) Healthy breast tissue: BC image



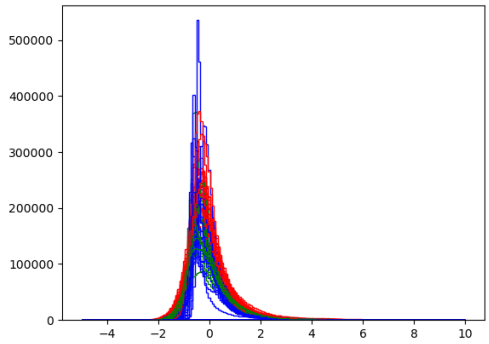
(b) Tumor: BC image



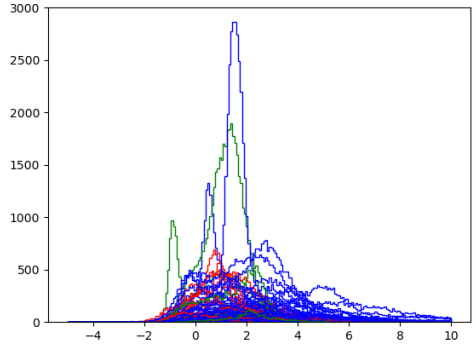
(c) Healthy breast tissue: HM-BC image



(d) Tumor: HM-BC image



(e) Healthy breast tissue: Z-score BC image



(f) Tumor: Z-score BC image

*Figure 5.17: Intensity histograms of healthy breast and tumor tissue from raw (1^{st} row), bias-corrected (BC) (2^{nd} row) and histogram-matched (HM) (3^{rd} row) or Z-score normalized BC (4^{th} row) **fat-saturated T2** images of the training set. **RED color** is used for patients scanned on the GE machine; **BLUE color** for those imaged on the Siemens machine with the Sentinelle coil and **GREEN color** when the 18-channel coil was used.*

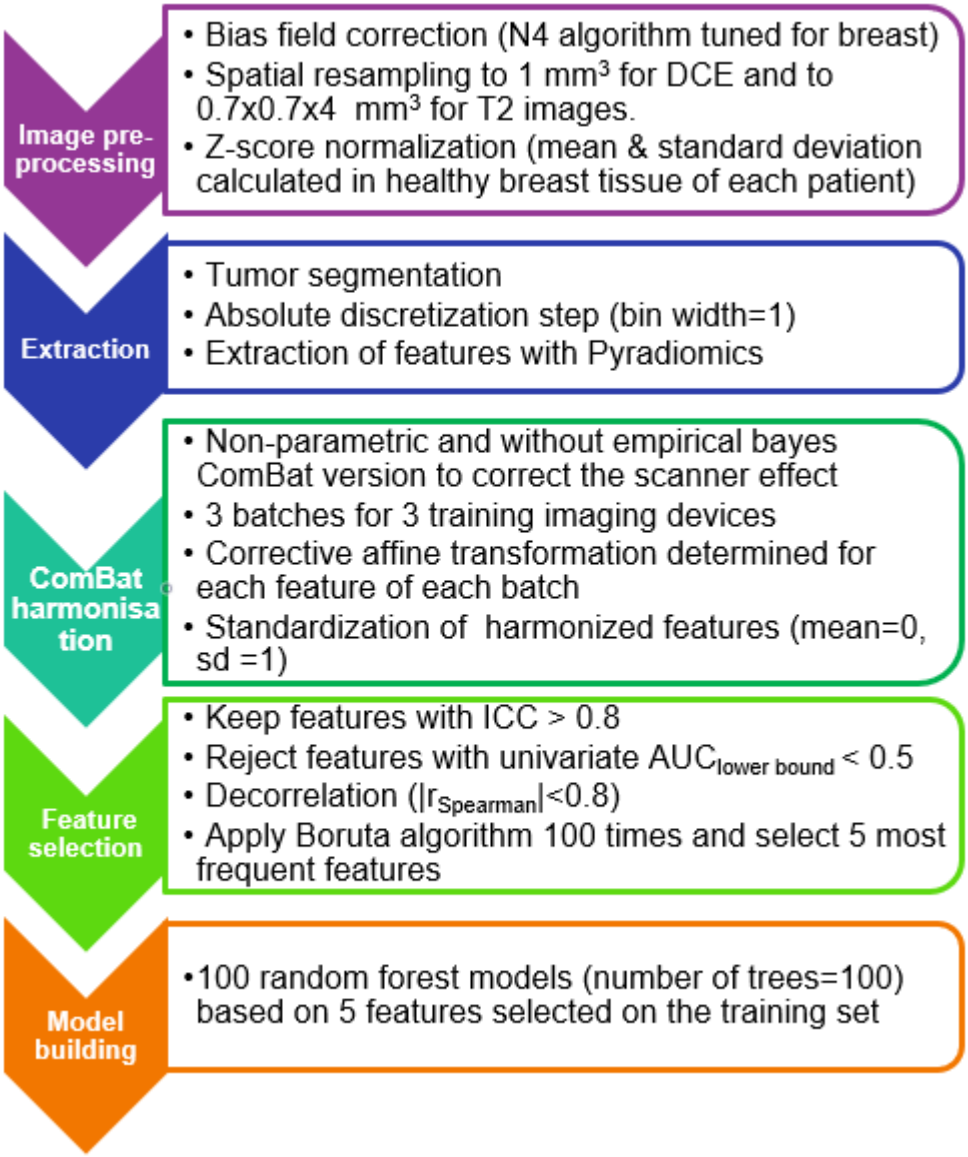


Figure 5.18: Summary of the different steps of the radiomic pipeline on the training set.

Chapter 6

Radiomic analyses to predict pCR to NAC

Preface

Following the pipeline defined in Chapter 5, this chapter will present first radiomic analyses based on pre-treatment images to predict pCR to NAC. The importance of delineating precisely the margins of the tumor to create the volume of interest from which to extract features, will be explored. An original harmonization method to tackle the "scanner effect" in small multicentric dataset when conventional harmonization methods like ComBat cannot be applied will be developed. Main results will be presented in the form of an article, in process of submission, which is an extension of the long abstract published in *The proceedings of IEEE EMBC* [25]. This chapter will use some tables and reintroduce the cohort previously defined in Chapter 3. The second part of the chapter present preliminary results on molecular subtype-specific models.

6.1 Introduction

The review of the literature in Chapter 2 showed a strong and growing interest in the radiomic field for the pre-treatment or early prediction of the response to NAC in breast cancer. Most predictive models used clinical and biological data, BI-RADS features, kinetic parameters, shape descriptors or first-order/texture features extracted from the tumor core and possibly their margins. Useful information in MR breast images to predict pCR to NAC must be better understood. Some studies looked into the potential use of peritumoral regions [131, 145, 187]. Other works have suggested that it is in a finer and multiresolution analysis of the tumor core heterogeneity, through the use of wavelet filtering for instance, that lies the information to predict pCR [146, 185]. Fractal-based texture analysis to explore structure irregularities in the tumor texture has also been investigated [185].

If many articles focused on the tumor core and the analysis of its peripheral regions, there has however never been an interest in establishing clearly the contribution of tumor heterogeneity and its shape and margins. The study of the tumor shape and margins often remains restricted to conventional shape parameters (Table 2.1) or qualitative BI-RADS descriptors. The chapter will thus delve into this issue. Contradictory findings like the interest of multiparametric signature over single sequence models, the relevance of molecular subtype-specific models and the exportability of radiomic models will also be investigated.

6.2 Article - Saint Martin et al., to be submitted

Multicentric export of MRI-based radiomic signatures to predict response to neoadjuvant chemotherapy in breast cancer.

IN PREPARATION FOR SUBMISSION

Marie-Judith Saint Martin[1], Frédérique Frouin[1], Irène Buvat[1], Pia Akl[2], Caroline Malhaire[1,3] and Fanny Orlhac[1]

[1] U1288-LITO, Inserm, Centre de Recherche de l'Institut Curie, Université Paris-Saclay, Orsay, France
[2] Department of Radiology, Hôpital Femme Mère Enfant, Hospices civils de Lyon, Lyon, France
[3] Department of Radiology, Ensemble Hospitalier de l'Institut Curie, Paris, France

Abstract

Background: MRI-based radiomic studies reported promising results to predict the response to neoadjuvant chemotherapy (NAC) in breast cancer but model evaluation on independent multicentric test sets is less documented. Besides, relevant information required for prediction (tumor shape, margins, intensity, or textural heterogeneity) is understudied.
Purpose: The aim of this study is to identify relevant tumor patterns to predict the response to NAC and to assess model transferability on an independent multicentric test set.
Materials and methods: In this retrospective study, fat-saturated T2 and T1-weighted contrast-enhanced pre-treatment MR images of 136 women treated with NAC between 2016 and 2020 at Institut Curie, were analyzed. Features extracted from four volumes of interest (VOIs) per patient, (tumors, bounding box surrounding tumors, bounding box on binarized tumor images, constant box inside tumors) were combined in fifteen experiments involving one to four VOIs and repeated using either T1 or T2 images only and then both modalities. Models were evaluated using leave-one-out cross-validation on a training set of 103 patients, acquired with three MR devices, and tested on an independent multicentric test set of 33 patients. An original feature harmonization strategy involving the projection of test patients to one of the three training imaging devices was developed.
Results: Among the 103 (33) patients (mean age 48 years $\pm$ 11 [SD]) in the training (test) sets, 49 (15) achieved pathological complete response. Models built with features extracted from binarized images of tumor lesions or from a combination of features from several VOIs, including precise delineation of tumors, yielded better performances than the model using only the standard tumor segmentation on the training and test set ($p < .001$). Harmonizing test feature values provide better or equivalent performances in most experiments (37/45).
Conclusion: Performances calculated with the Youden Index to predict the response to NAC

on a multicentric test set were improved by combining features from several VOIs including binarized images of tumor lesions and by using a harmonizing strategy.

Key words: Breast cancer, radiomics, MRI, NAC, pCR

Abbreviations:
NAC: Neoadjuvant Chemotherapy
pCR: Pathological Complete Response
VOI: volume of interest
T: Tumor
CB: Constant box
BB: Bounding box
bBB: Binary bounding box
LOOCV: Leave-one-out cross-validation
IQR: interquartile range

Introduction

Women with locally advanced breast cancer frequently receive neoadjuvant chemotherapy (NAC) before surgery. NAC aims to reduce the size of tumors to facilitate surgeries and enhances breast conservation rates [57]. Achieving pathological complete response (pCR) is also associated with better overall survival [60]. Predicting beforehand the response, which depends on multiple factors such as molecular subtype [248], would considerably improve patient care in offering women a tailored approach to their treatment. Numerous radiomic studies attempted to predict pCR using baseline MR images acquired before the beginning of NAC [222, 249]. These works built predictive models based on shape, intensity and textural features [149, 189] extracted from volumes of interest (VOIs) drawn in images, corresponding usually to the tumor core [196], sometimes extended to peritumoral regions [144, 187]. The specific influence of the extent, shape, and margins of the VOIs on the textural and intensity information captured by features, were however never addressed. It would therefore be of interest to unravel the different sources of information and evaluate their contribution in predictive models. Segmenting tumors is a time-consuming and radiologist-dependent task, often hampering the constitution of large radiomic cohorts needed to gather a substantial variability of patients. For instance, if the textural information captured by bounding boxes around tumors, like experimented in PET [250], was sufficient to predict pCR, it could considerably facilitate radiomic studies.

Radiomic studies also suffer from the impact of the "scanner effect" [251], i.e. the influence of image acquisition and reconstruction parameters on radiomic feature values. Thus, performances of radiomic models trained on data acquired with one imaging device are significantly degraded when applied to data acquired with another device [18]. Many radiomic studies therefore report results using cross-validation on the training set [13, 146, 149, 170, 184, 192, 196] or test sets gathering patients imaged on the same scanners as used during training [15, 144, 150, 185, 189, 190, 193, 252]. Though corrective methods, like the ComBat method [251], have proven their efficiency in reducing the scanner effect, they require a substantial number of patients ($\sim$20-30) from each scanner to be efficient [11]. This number proves to be difficult to be reached in multicentric datasets. The present study aims to identify the distinct

sources of relevant information to predict pCR using radiomic analyses based on four types of 3D VOIs, capturing different tumor characteristics quantifiable from medical imaging. A strategy to realign feature values extracted from multiple scanners is developed and transferability of the predictive models is assessed on a multicentric independent test set.

Methods

Study sample

This retrospective study, approved by our institutional review board (IRB number OBS180204), included all women with breast cancer consecutively treated with NAC in our institute between 2016 and 2020, who underwent initial breast MRI before the beginning of treatment. The requirement to obtain informed consent was waived. Based on 156 patients, discarding patients with missing modalities, 136 datasets were collected (Figure 6.1). In addition, clinical and biological information was reported (Table 6.1).

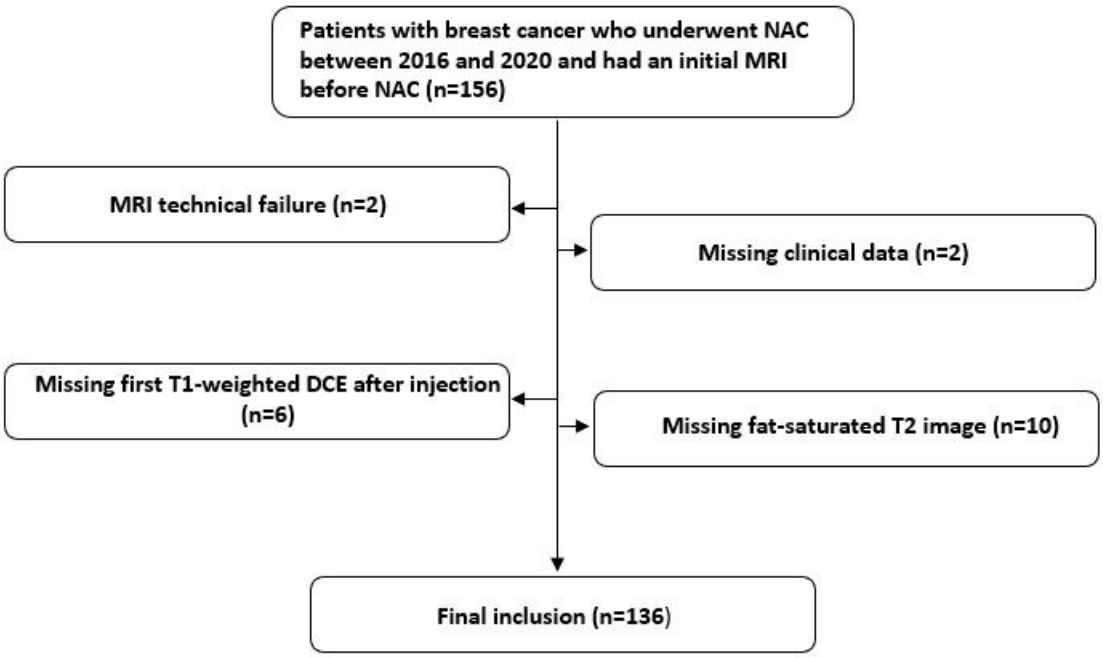


Figure 6.1: Flowchart of study inclusion. NAC= Neoadjuvant chemotherapy. DCE= Dynamic contrast-enhanced.

Imaging

Patients were imaged in one of the three imaging devices of the institute (n=110) or in other imaging centers with different scanners and coils (n=26) (Table 6.2). The training set gathered 103 women imaged at the institute while the test set was made of all patients imaged outside the institute (n=26) and 7 patients imaged at the institute and included later in the study. Detailed image parameter acquisitions are available in supplemental Table 6.5. Images were corrected for bias field gain using the N4 algorithm tuned specifically for the breast area [19].

Table 6.1: Clinical & biological data.

Label	Levels	Training (n=103)	Testing (n=33)	Total	p
Age (y)	Median (IQR)	48.0 (39.5 to 56.5)	46.0 (39.0 to 52.0)	47.5 (39.0 to 56.2)	0.594
BMI (kg.m^{-2})	Median (IQR)	23.4 (21.4 to 25.7)	23.4 (21.5 to 27.5)	23.4 (21.5 to 26.1)	0.359
Menopause	Postmenopausal	42 (40.8)	11 (33.3)	53 (39.0)	0.577
	Premenopausal	61 (59.2)	22 (66.7)	83 (61.0)	
T stage	0/I/II	91 (88.3)	24 (72.7)	115 (84.6)	0.059
	III/IV	12 (11.7)	9 (27.3)	21 (15.4)	
N stage	0	58 (56.3)	16 (48.5)	74 (54.4)	0.559
	I/II	45 (43.7)	17 (51.5)	62 (45.6)	
M stage	0	102 (99.0)	33 (100.0)	135 (99.3)	1.000
	I	1 (1.0)	0 (0.0)	1 (0.7)	
Histological Type	Ductal NOS	99 (96.1)	32 (97.0)	131 (96.3)	0.757
	Lobular	1 (1.0)	0 (0.0)	1 (0.7)	
	Mixt	2 (1.9)	0 (0.0)	2 (1.5)	
	Other	1 (1.0)	1 (3.0)	2 (1.5)	
Molecular subtype	HER2+	12 (11.7)	7 (21.2)	19 (14.0)	0.587
	Luminal B/HER2-	30 (29.1)	9 (27.3)	39 (28.7)	
	Luminal B/HER2+	13 (12.6)	3 (9.1)	16 (11.8)	
	TN	48 (46.6)	14 (42.4)	62 (45.6)	
Grade	2	34 (33.0)	8 (24.2)	42 (30.9)	0.464
	3	69 (67.0)	25 (75.8)	94 (69.1)	
Ki67 (%)	Median (IQR)	60.0 (32.5 to 75.0)	40.0 (30.0 to 60.0)	50.0 (30.0 to 75.0)	0.231
TILFactor	High	44 (42.7)	18 (54.5)	62 (45.6)	0.324
	Low	59 (57.3)	15 (45.5)	74 (54.4)	
Response to NAC	non pCR (npCR)	54 (52.4)	18 (54.5)	72 (52.9)	0.991
	pCR	49 (47.6)	18 (54.5)	64 (47.1)	

Patients and tumors characteristics in training and test sets. Continuous variables are represented by their median and interquartile range (IQR). Wilcoxon rank sum test and Pearson's Chi-square test were performed respectively for continuous and categorical variables between training and test sets. In circumstances where Chi-square test could not be used due to too few observations, Fisher's exact test was carried out.

Table 6.2: Imaging devices of training and test sets.

Imaging centers	Manufacturers	Devices	Magnetic field strength (T)	Coils	Training	Testing
Institut Curie	GE	Optima MR450w	1.5	8-channel coil	25	3
Institut Curie	Siemens	MAGNETOM Aera	1.5	18-channel coil	19	0
Institut Curie	Siemens	MAGNETOM Aera	1.5	Sentinelle (16-channel) coil	59	4
Other center	Siemens	MAGNETOM Aera	1.5	16-channel coil	0	4
Other center	Siemens	MAGNETOM Aera	1.5	18-channel coil	0	3
Other center	Siemens	MAGNETOM Aera	1.5	Spine 32-channel coil	0	1
Other center	Siemens	MAGNETOM Amira	1.5	18-channel coil	0	1
Other center	Siemens	MAGNETOM Avanto eco	1.5	Breast matrix coil	0	1
Other center	Siemens	MAGNETOM Avanto eco	1.5	16-Channel AI Breast coil	0	1
Other center	Siemens	MAGNETOM ESSENZA	1.5	Breast matrix coil	0	1
Other center	GE	Discovery MR 750	3	HD Breast coil	0	1
Other center	GE	Optima MR360	1.5	HD Breast coil	0	4
Other center	GE	Optima MR450w	1.5	HD Breast coil	0	2
Other center	GE	Signa Artist	1.5	HD Breast coil	0	3
Other center	GE	Signa HDxt	1.5	HD Breast coil	0	2
Other center	GE	Signa Voyager	1.5	HD Breast coil	0	2

***Training**: Number of patients included in training set; **Testing**: Number of patients included in test set; **Other centers**: imaging centers other than Institut Curie.*

Definition of the VOIs

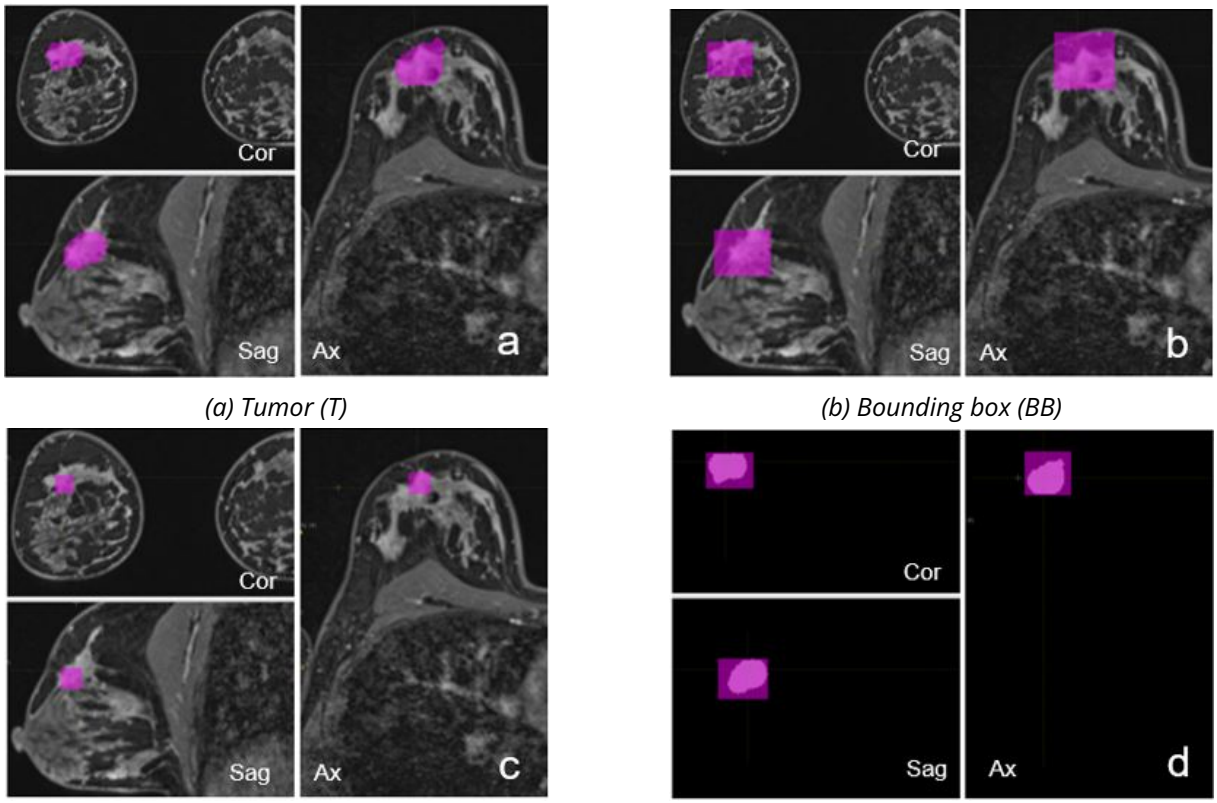
Two radiologists (with 14 and 3 years of experiences in breast MRI) equally and independently segmented the lesions in 3D on the first T1-weighted DCE image after gadolinium-based contrast media injection using the LIFEx software [235] (version 6.0, www.lifexsoft.org). Thirty lesions in the training set were segmented by both radiologists to assess segmentation variabilities. Segmentations were resampled to fit onto the fat-saturated T2 images. To get a more precise delineation of the tumor borders, segmentations were refined on the T1-weighted DCE image using a threshold equal to 40% of the maximum intensity inside tumor (tDCE). This segmentation with its 3 declinations: DCE, tDCE and T2 (Supplemental Figure 6.9) constitutes the first type of VOI. Denoted "T", it is the standard VOI used in the majority of radiomic studies (Figure 6.2a). Features from this VOI capture tumor heterogeneity and intensity variations but are influenced by shape and border outlines of the lesions. To dissociate the different sources of information involved in the prediction, three other VOIs were automatically defined:

- The minimal 3D bounding box around the tumor, on which was added a one-voxel

border, formed a second VOI ("BB" for bounding box). Its features transcribed the heterogeneity of the tumoral and peritumoral regions and can be influenced to a lesser extent by the volumes of the boxes (Figure 6.2b).

- The third VOI ("CB" for constant box) is a 12-pixel wide cube of fixed size across the database. It was chosen so that the cubes could fit in the lesions and positioned at their center of mass. It focuses on the tumor heterogeneity and intensity and brings no information about shape or volume (Figure 6.2c).

- The last VOI ("bBB" for binary bounding box) has the same shape as the bounding box but is applied on binarized DCE and T2 images where tumor voxels value is 2 while the rest of the image is set to 1. It only apprehends shape and margins of the lesion (Figure 6.2d).



(a) Tumor (T)

(b) Bounding box (BB)

(c) Constant box (CB)

(d) Binary bounding box (bBB)

Figure 6.2: Coronal (Cor), sagittal (Sag) and axial (Ax) views of the 4 VOIs on the first dynamic contrast-enhanced (DCE) image after gadolinium-based contrast media injection of a 44-year-old woman with breast cancer. ***(a)*** *Full tumor lesion VOI (abbreviated T);* ***(b)*** *Bounding box VOI surrounding the tumor (abbreviated BB);* ***(c)*** *Constant Box VOI of fixed size centered inside the tumor (abbreviated CB);* ***(d)*** *Binary bounding box (abbreviated bBB).*

Prediction Pipeline

To investigate the opportunity of combining features from different VOIs, 15 experiments were defined, carrying out all possible combinations using features extracted from either one (T, BB, CB or bBB), two, three or all the four VOIs. Though multiparametric signatures based on several MR sequences have achieved good results [15, 134], using only one sequence sometimes proved to yield better results [13]. The 15 experiments were therefore repeated three times: the first round of experiments used only T1-weighted DCE images (DCE and tDCE segmentations), the second one only fat-saturated T2 images (T2 segmentations) and the last one both modalities (DCE, tDCE, T2). For every experiment, the feature selection process and model building followed the pipeline described in Figure 6.3.

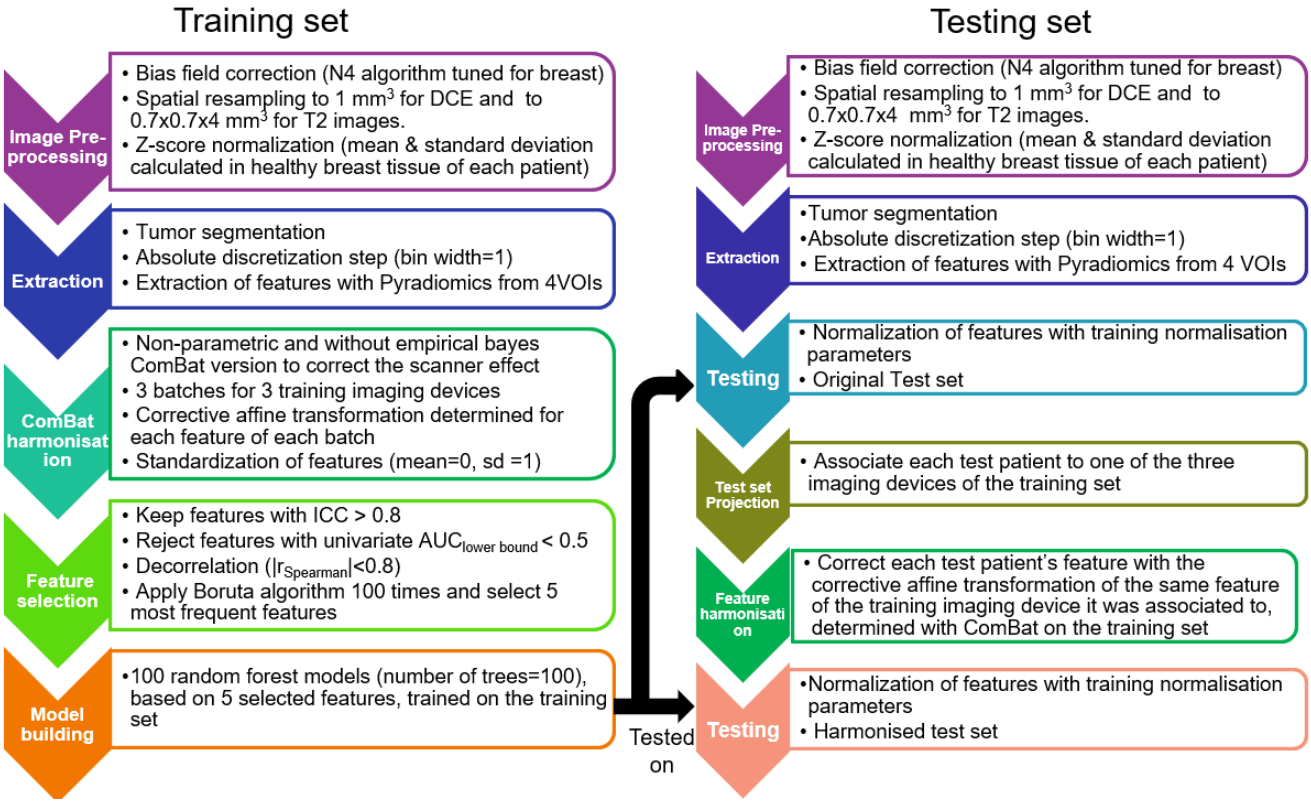


Figure 6.3: Diagram summarizing the pre-processing and selection process on the training set to build predictive models, tested after a preprocessing step on the original and harmonized test set. ICC= Intraclass correlation coefficient; $r_{Spearman}$= Spearman's rank correlation coefficient; $AUC_{lowerBound}$ = lower bound of the area under the receiver operating characteristic curve. VOIs= volumes of interest.

Radiomic features were computed with the IBSI-compliant [143] Pyradiomics software (v 3.0.1) [152] on the native and wavelet-filtered images. The same set of 107 features (93 texture features and 14 shape features) was extracted from each declination of each VOI for each image. In order to differentiate features during the selection process, a suffix was added to each feature to identify the VOI it was extracted from (T, BB, CB, bBB) and the modality of the images (T2, tDCE and DCE forms).

Harmonization of the test set

Due to the many imaging devices used in the test set (Table 6.2) and the low number of patients imaged on each of them, the ComBat harmonization method could not be directly used. Therefore, an original harmonization strategy of the feature values of the test set, involving the projection of test patients onto one of the three training imaging devices was developed.

In the tumor area, the "scanner effect" is mixed with the disruptions due to the tumor signal. In healthy tissue, the specific impact of the "scanner effect" can be more clearly assessed. A box of constant dimensions (12-pixel wide cube) was therefore placed in the healthy breast tissue of normalized bias-corrected T2 and T1-DCE images of every patient from training and test sets. Radiomic features were extracted from this box using absolute bounds with fixed bin size. As parameters of the T1-DCE and T2 sequences of imaging devices of the test sets may be closer to different training imaging devices, the harmonization strategy was carried out separately for the two sequences.

Using training set patients only, features that were heavily impacted by the "scanner effect" were selected (Kruskal-Wallis $p < 10^{-14}$ for T1-DCE features, and $p < 10^{-8}$ for T2 features). 17 T1-DCE and 10 T2 features were selected and normalized. Using the selected features, the centroids of the three clusters formed by patients of the three training imaging devices were calculated. Test patients were then assigned to the cluster whose centroid they were the closest to using Euclidean distance and the previously selected features. Every feature of a test patient extracted from a tumor VOI could then be corrected using the ComBat affine transformation determined during the training phase for this feature for its assigned scanner.

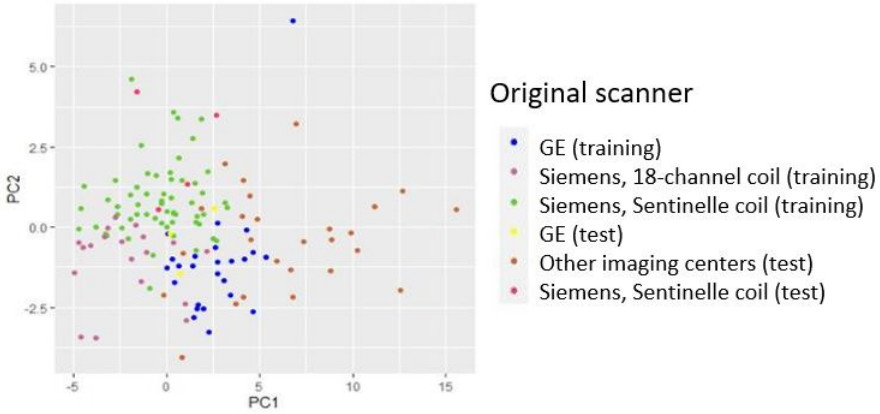
Figure 6.4 illustrates in 2D with principal component analysis (PCA) the projection process for the T2 sequence. PCA visualization shows the clustering of patients according to their scanners. Table 6.3 and Table 6.4 reports the projection results for T1-DCE and T2 sequences. As 7 out of the 33 test patients were imaged on one of the training imaging devices (4 on the Siemens with Sentinelle coil, 3 with the GE machine), it was possible to assess the projection method based on their respective assigned scanners. For both modalities, 5 out of 7 cases were projected accurately while the two other patients were projected on the Siemens device (with 18-channel coil).

Table 6.3: Projection of test set patients on closest scanners for T1-DCE sequence.

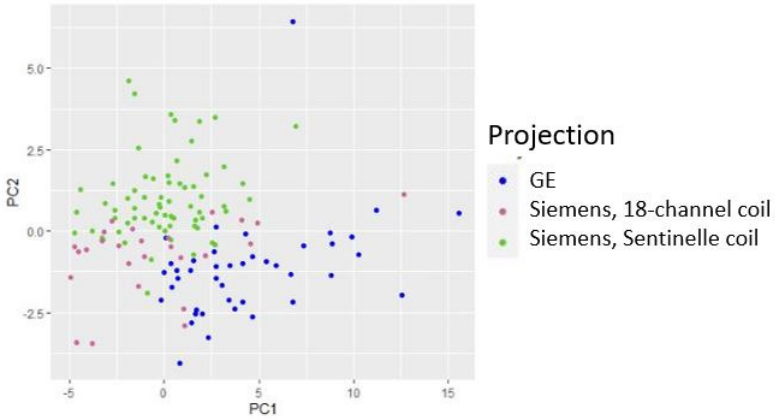
$Projection$ / $OriginalScanner$	Siemens (Sentinelle coil)	Siemens (18-channel coil)	GE
Siemens (Sentinelle coil)	3	1	0
GE	0	1	2
Other centers	13	8	5

Table 6.4: Projection of test set patients on closest scanners for T2 sequence.

$Projection$ / $OriginalScanner$	Siemens (Sentinelle coil)	Siemens (18-channel coil)	GE
Siemens (Sentinelle coil)	4	0	0
GE	0	2	1
Other centers	5	5	16



(a) Patients of training and test sets colored according to their original scanner.



(b) Patients of training and test sets colored according to their projected scanner for the test set and original scanner for the training set.

Figure 6.4: Principal component analysis (PCA) representation of patients from training and test sets using T2 features impacted by the "scanner effect". Each dot represents a patient and is colored: ***(a)*** *by the original scanner on which patients were imaged;* ***(b)*** *by their projected training scanner for test patients and by their original scanner for training patients.*

Statistical analysis

Statistical analyses were performed in R (version 4.1). Performances of random forest models on the training set were evaluated with the median and interquartile range (IQR) of the Youden index (Y=sensitivity+specificity-1) using 100 repetitions of leave-one-out cross-validation (LOOCV). Performances on the test set were obtained using 100 random forest models made of the five features selected during the training process (Figure 6.3) and tested on the original and harmonized feature values issued from the test set. Kruskal-Wallis tests were carried out to compare the performances of the 15 experiments of each round globally then Dunn's tests were performed to make 2-by-2 comparisons.

Results

Patient characteristic

A total of 136 patients with mean age 48 $\pm$ 11 years were included (Figure 6.1). The tumor response was assessed on surgical specimens after NAC using the Residual Cancer Burden (RCB) score. The training and test set presented both a relatively balanced proportion of responders and non-responders to the treatment (Table 6.1) and included patients diagnosed with tumors from different molecular subtypes.

Training results

When using T1-weighted DCE images only (Figure 6.5a), best results were obtained with the "bBB & T" experiment (median Youden index, 0.52, [IQR, 0.50, 0.54]) while the "T & CB & bBB" experiment achieved best results when using only T2 images (0.47, [0.43, 0.49], Figure 6.5b) or both modalities (0.50, [0.46, 0.52], Figure 6.5c). Supplemental Figure 6.10 ranks all the 45 training experiments according to the median of their Youden index distributions. Features selected in the best experiments are reported in Supplemental Table 6.5.

As there was no constraint on the selection algorithm to select features from each VOI or modality the experiment uses, several experiments can select the same set of features like the experiments "All (T2)" and "T & CB & bBB (T2)" or the experiments "All (DCE & T2)" and "T & CB & bBB (DCE & T2)".

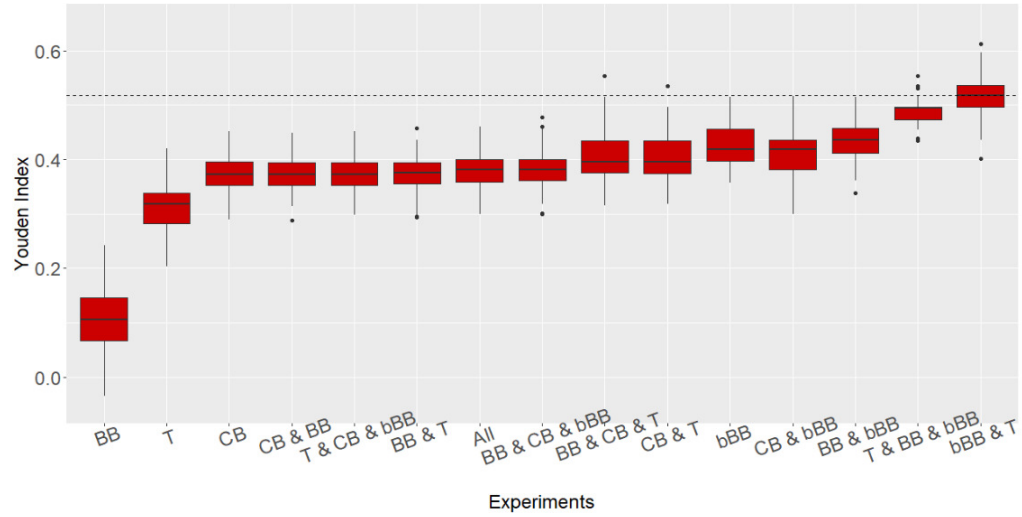
Kruskal-Wallis test revealed that the medians of Youden index distributions of the experiments were not equal ($p < 10^{-12}$). Subsequent analysis of Dunn's test (Figure 6.7) showed that there was no significant difference between the best experiments of each round ("bBB & T (DCE)","T & CB & bBB (DCE & T2)", "T & CB & bBB (T2)") but that they all significantly outperformed experiments based on the tumor segmentation alone ($p < 0.05$, Figure 6.7d). Statistical results in Figure 6.7d were calculated using Dunn's test on the 45 experiments together but for clarity's sake, only partial results are depicted in the figure.

Testing results

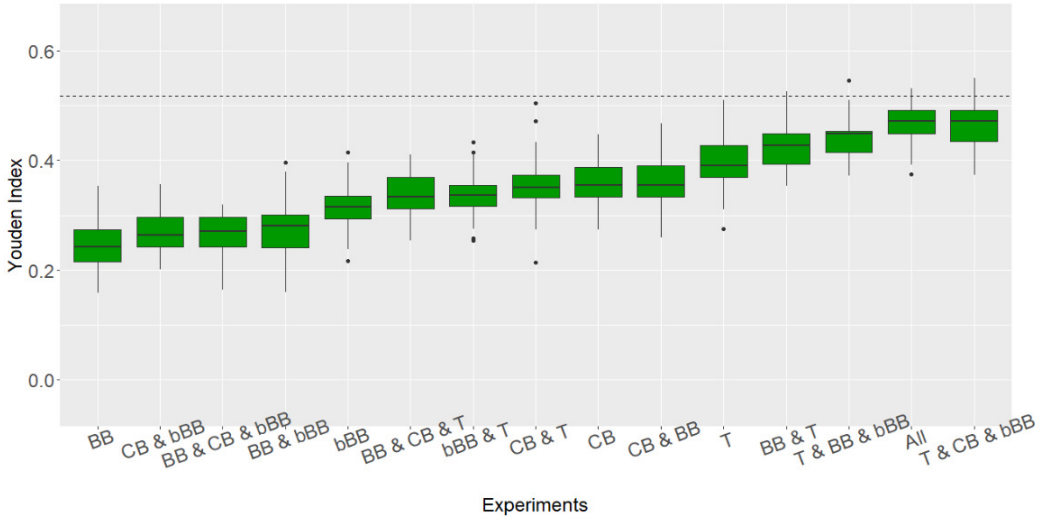
Harmonization of the test set following the new projection approach led to better or equivalent results (Wilcoxon signed rank test between results before and after harmonisation $p < 0.05$) in 13/15 experiments when using only T1-DCE images (Figure 6.6a), in 12/15 with both modalities (Figure 6.6c) and in 12/15 with T2 images (Figure 6.6b).

Performances on the test set were disparate amongst experiments, going from negative Youden index values to a median of 0.44, IQR [0.39, 0.50] for the best experiment "bBB" in T1-DCE after harmonization.

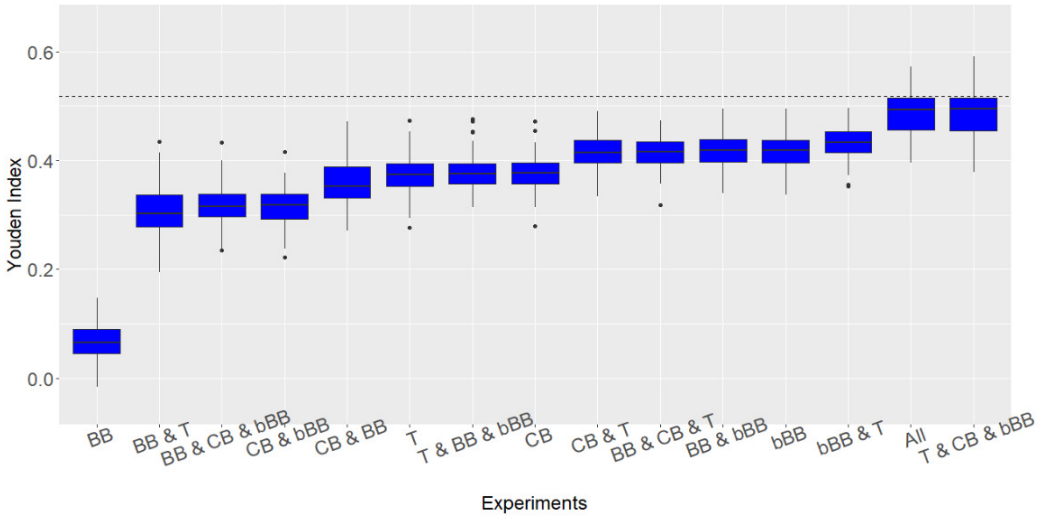
Similarly to the training experiments, results of testing experiments on the harmonized test set were significantly different in each round (Kruskal-Wallis ($p < 10^{-12}$). Dunn's tests comparing the best results across modalities with the tumor experiments (T (DCE), T (T2), T (DCE & T2)) show that they achieved significantly better performances ($p < 0.05$, Figure 6.8).



(a) Using T1-weighted DCE images

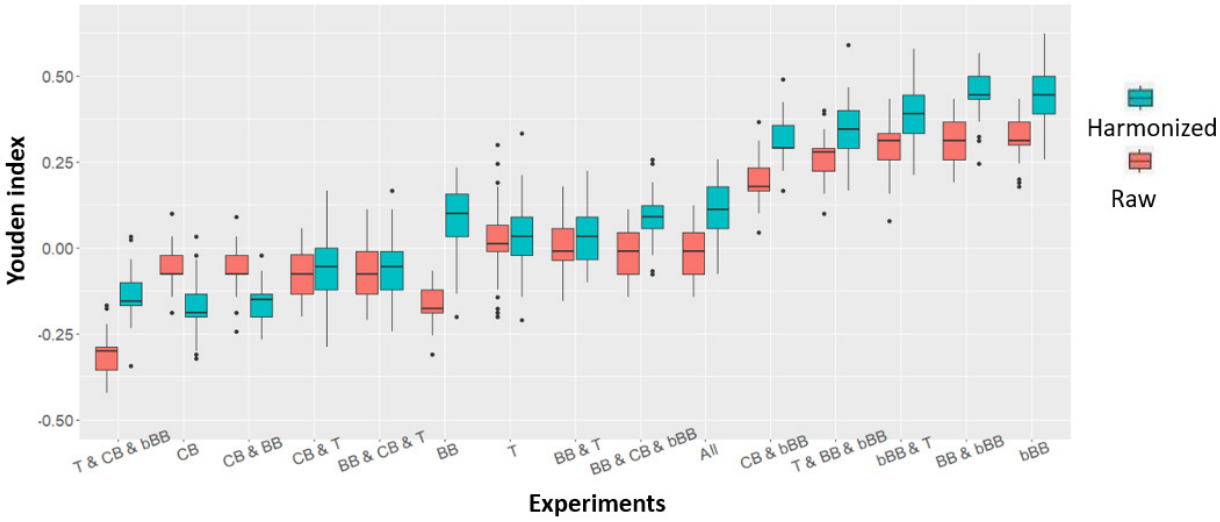


(b) Using fat-saturated T2-weighted images

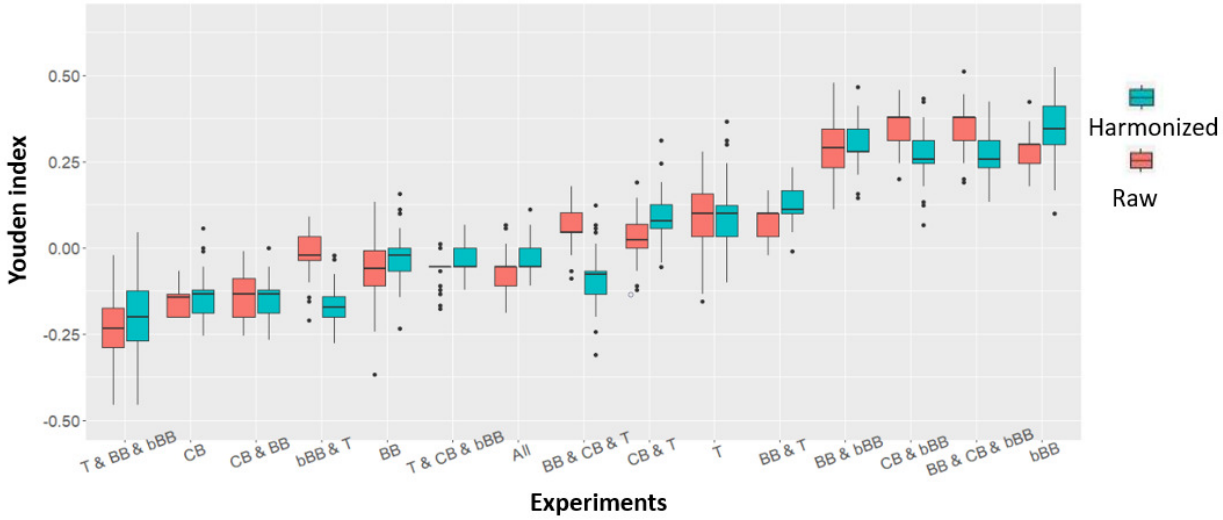


(c) Using both fat-saturated T2 & T1-weighted DCE images

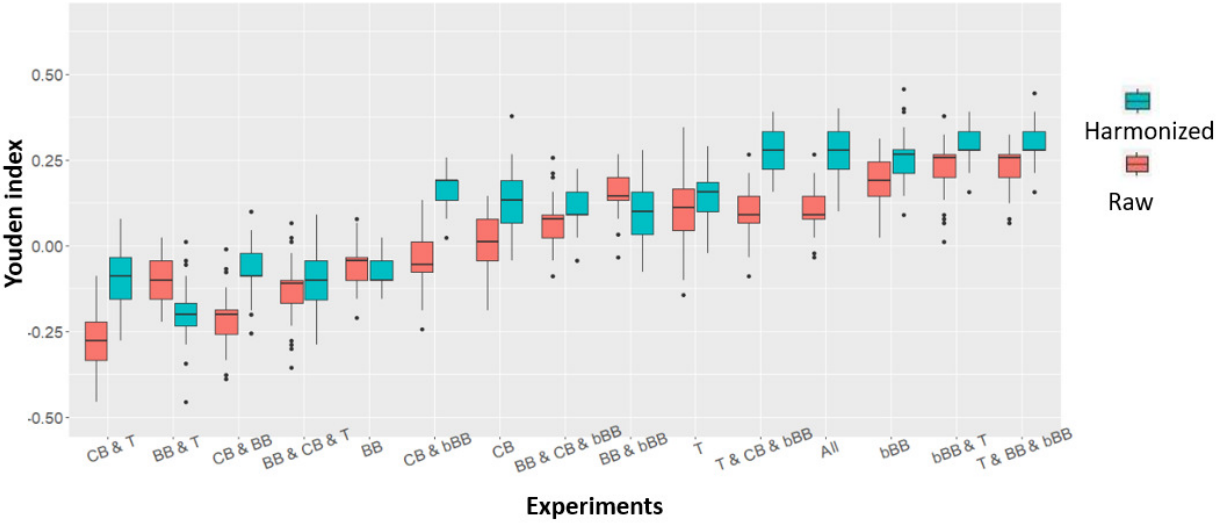
*Figure 6.5: Training results ranked according to the median value of the Youden index using LOOCV with segmentations based on **(a)** T1-DCE images; **(b)** fat-saturated T2 images; **(c)** images from both modalities. Dashed lines represent the median Youden index value of the best experiment: "bBB & T" on T1-DCE images.*



(a) Using T1-weighted DCE images

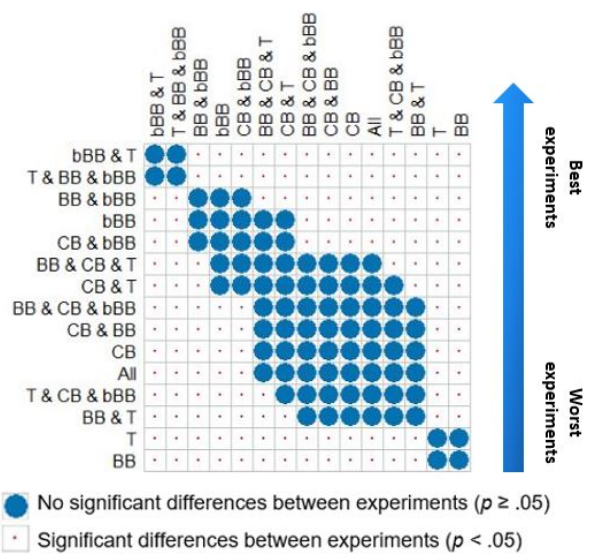


(b) Using fat-saturated T2-weighted images

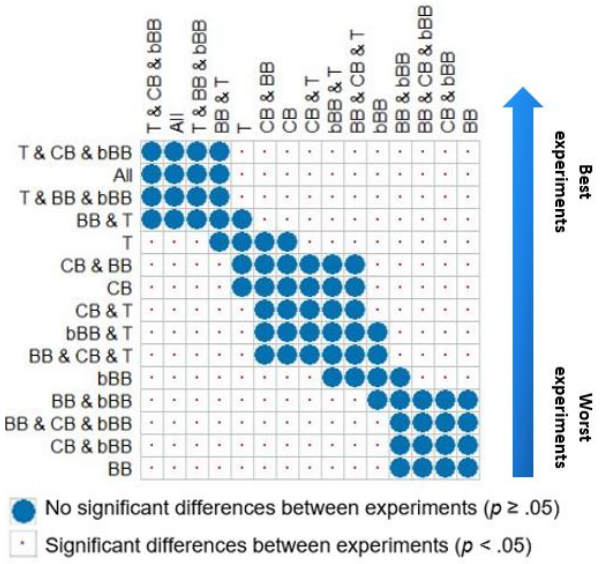


(c) Using both fat-saturated T2 & T1-weighted DCE images

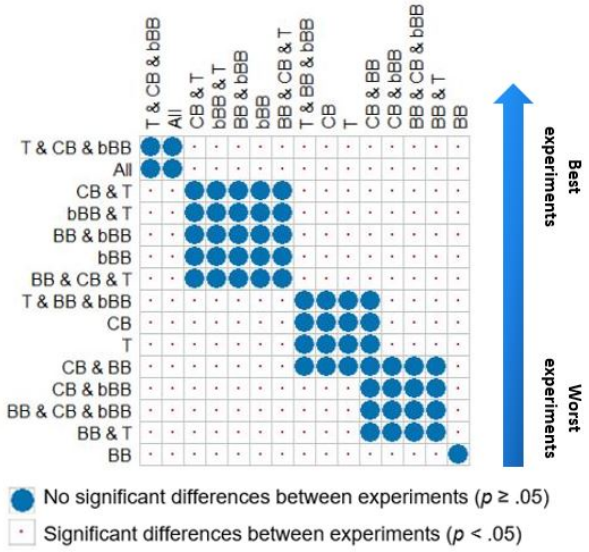
*Figure 6.6: Test results ranked according to the median value of the Youden index obtained using 100 random forest models with segmentations based on **(a)** T1-DCE images; **(b)** fat-saturated T2 images; **(c)** images from both modalities.*



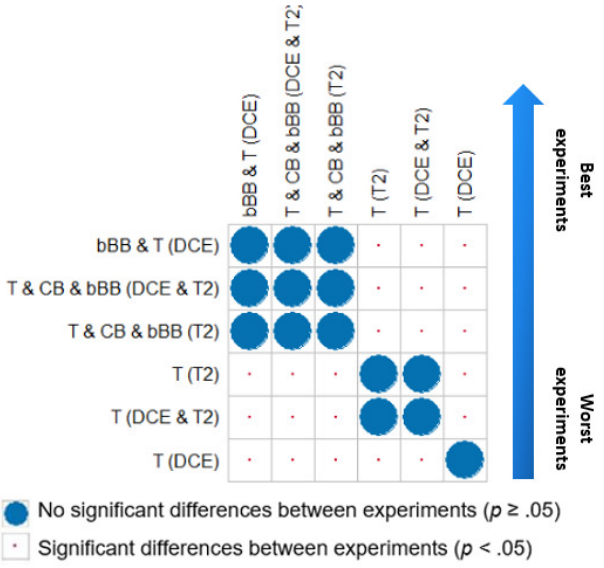
(a) Results using T1-weighted DCE images



(b) Results using fat-saturated T2-weighted images



(c) Results using both T2 & T1-weighted DCE images



(d) Comparing best results with Tumor experiments

Figure 6.7: Statistical analysis of training results using Dunn's test. ***(a)*** *comparisons of results using only T1-DCE images;* ***(b)*** *comparisons of results using only T2 images;* ***(c)*** *comparisons of results using T2 & T1-DCE images,* ***(d)*** *comparisons of best results across modalities and tumor experiments (T DCE, T T2, T DCE & T2).*

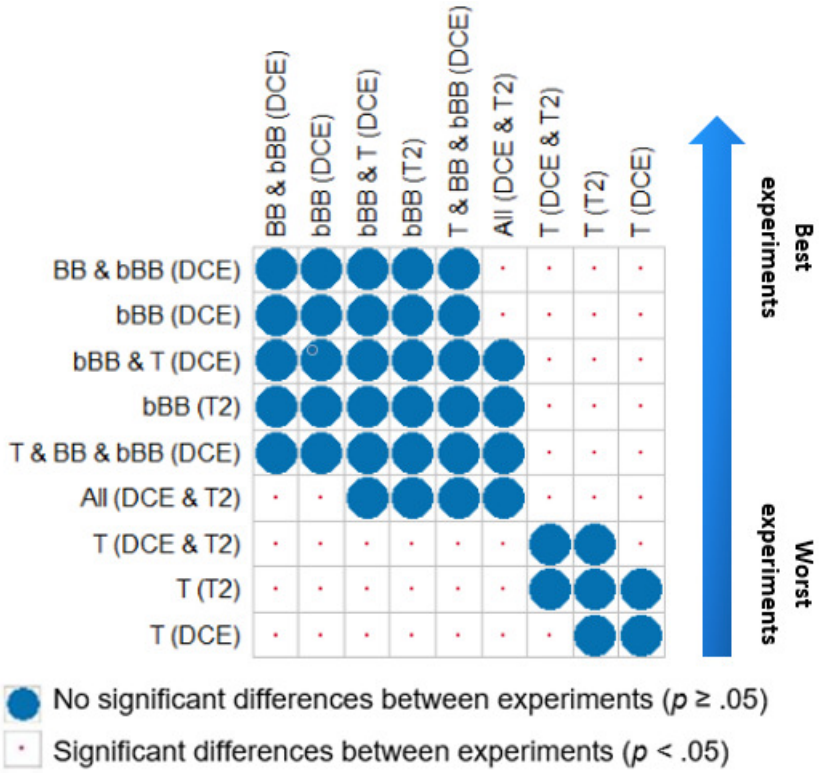


Figure 6.8: Statistical analysis of test results using Dunn's test. The six best results overall on the test sets covering the three cases (DCE, T2 and DCE & T2) were compared with the tumor experiments (T (DCE), T (T2), T (DCE & T2)). Like in Figure 6.7d, Dunn's tests were performed on the 45 experiments but for clarity's sake, only partial results are depicted.

Discussion

This study investigated which tumor patterns quantified from MR images contribute the most to radiomic model prediction of pathological complete response (pCR) to neoadjuvant chemotherapy in breast cancer. It also aimed to assess model transferability on an independent and multicentric test set with and without a harmonization strategy of radiomic feature values.

Based on our experiments involving four types of volumes of interest (VOIs) apprehending tumors in diverse ways, we found that significantly better performances than those of the models using the standard tumor delineation alone ("T"), could be achieved. The constant box ("CB") experiment that only assesses tumor heterogeneity outperformed the tumor models in the DCE and DCE & T2 training results. Similarly, the binary bounding box ("bBB") that relies on shape and borders information, also surpassed in these conditions the tumor experiments (Figure 6.5a,6.5c). On the other hand, the bounding box ("BB") experiments always trailed behind. However, results from the training set suggest that the combination of features from different VOIs achieved the best results. In the round of experiments based on T2 or DCE &T2 images, the combination of features from three VOIs (T, bBB and CB) topped results. These findings suggest that complex shape or border information and textural heterogeneity could

both present interest to predict pCR, relegating the idea of using only constant or bounding boxes, and that information found in the different VOIs is complementary and not redundant.

Results of the test set showed a marked drop in performances in models built on raw features. The designed harmonization strategy improved predictions in 73% (33/45) and gave equivalent performances in 9% (4/45) of the cases. Best results of the test set were obtained after harmonization and were globally consistent with the training results in the DCE and DCE & T2 rounds of experiments. In the T2 round however, best training experiments achieved poor performances and the harmonization strategy did not work as well. In all rounds, the performance of the "bBB" was particularly good and consistent with training Youden indexes. The "bBB" captures the information found in the shape and margins of tumor lesions beyond what traditional shape parameters calculated on the standard tumor VOI in Pyradiomics, such as volume, elongation, or sphericity, could apprehend. Unlike features extracted from the T VOI, features from the bBB, achieving already relatively good performances on the test set before harmonization, seemed also more robust to the scanner effect. This is quite logical since the binarization process reduced its dependence on the intensity values affected by the scanner effect. The analysis of the test set showed that the best performances were obtained by combining features from several VOIs or using the "bBB" experiment and that they always surpassed tumor model results (T (DCE), T (T2), T (DCE & T2)). No definitive conclusion on the benefit of multiparametric signature over single-sequence models could be reached but T1-DCE images used alone bring the best performances for the test set.

Drops in model performances on the multicentric test set are in line with recent studies [18, 251], highlighting the core problem of radiomic model transferability in the context of the scanner effect. Da-ano et al. [253] proposed in the similar circumstances of an heterogeneous and independent test set, to gather test patients into groups using hierarchical clustering and to consider these groups as new imaging devices to apply the ComBat algorithm. However, due to the much smaller dataset available in our study (33 patients versus 98 patients in Da-ano et al.), separating the test set in several groups would have created too small batches to properly apply ComBat. We opted instead for an original approach that consists in identifying the device of the training set closest to each patient in the test set to apply the most suitable ComBat transformation. Unlike Da-ano et al., the projection of patients was carried out using radiomic features calculated in a VOI located in the healthy contralateral breast of patients instead of using features from tumor lesions. Healthy breast tissue should indeed give a clearer outlook of the scanner effect, as it is not disturbed by tumor signals.

It is difficult to compare performances between already published studies due the variability of databases, methods and the number of selected features for instance, but our results were comparatively superior to those obtained by Granzier et al. [18] when using a multi-scanner test set (slightly lower median sensitivity 67% (10/15) versus 73% (36/49), and improved specificity 83% (15/18) versus 36% (43/119) when comparing the "bBB (DCE)" experiment to results on the ZMC cohort in [18]). Though the test set used by Granzier et al. was larger than ours, it was less diverse (three imaging devices from a single medical center versus more than 15 devices from multiple medical centers).

Our study nevertheless suffers from the small size of our dataset and especially our test set. The distribution of molecular subtypes in the cohort also made it difficult to build molecular subtype-specific models as in [15, 144, 167, 168] for every subtype while preserving an independent test set. Our dataset is however much closer to what could be observed in routine clinical practice, where the subtype is not always known at the time of baseline MRI.

Conclusion

In conclusion, models built with features extracted from binarized images of tumor lesions or from a combination of features from several VOIs automatically derived from the tumor segmentation outperformed models relying solely on the standard tumor segmentation to predict response to neoadjuvant chemotherapy. Our new harmonization strategy of the test set feature values improved in most experiments the models transferability on an independent multicentric test set.

Supplemental data

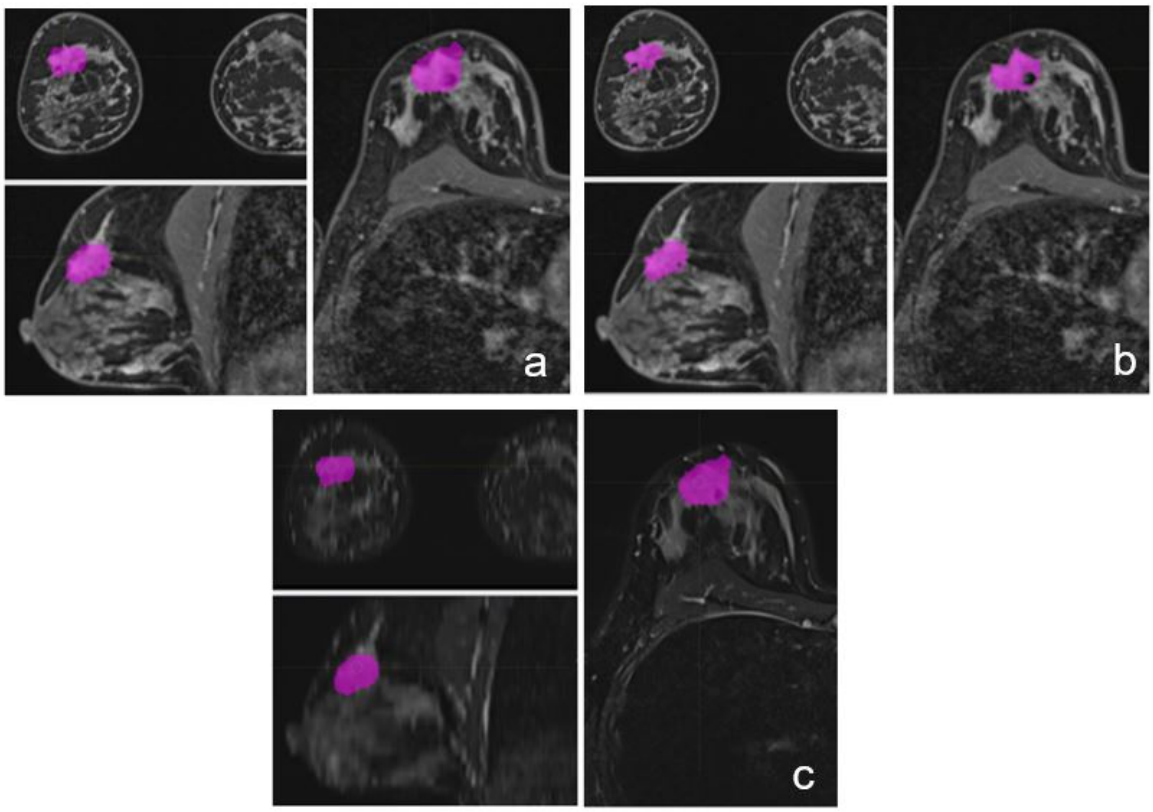


*Figure 6.9: Illustrations of **(a)** the full segmentation delineated by radiologists on first T1-weighted DCE image after contrast injection; **(b)** the thresholded segmentation on DCE image; **(c)** the full segmentation on a fat-saturated T2 image.*

Table 6.5: Features selected in models of interest and associated Youden index on test set.

Experiments	Features	Youden index (media, [IQR]) on test set
BB & bBB (DCE) *	original glszm SizeZoneNonUniformityNormalized (tDCE bBB) original glszm SmallAreaLowGrayLevelEmphasis (tDCE bBB) waveletHHH firstorder Maximum (tDCE bBB) waveletHHL firstorder Minimum (DCE bBB) waveletHHL firstorder Range (DCE bBB)	0.44, [0.43, 0.50]
bBB & T (DCE)	original glszm SizeZoneNonUniformityNormalized (tDCE bBB) original glszm SmallAreaLowGrayLevelEmphasis (tDCE bBB) waveletHHH firstorder Maximum (tDCE bBB) waveletHLL firstorder Range (DCE bBB) waveletLLH glszm ZoneEntropy (tDCE T)	0.39, [0.33, 0.44]
bBB (T2)	waveletHHH firstorder Maximum (T2 bBB) waveletHLH firstorder Minimum (T2 bBB) waveletHLH firstorder Range (T2 bBB) waveletLLH glszm SmallAreaEmphasis (T2 bBB) original glszm HighGrayLevelZoneEmphasis (T2 bBB)	0.34, [0.30, 0.41]
T & BB & bBB (DCE)	original glszm SizeZoneNonUniformityNormalized (tDCE bBB) original glszm SmallAreaLowGrayLevelEmphasis (tDCE bBB) waveletHHH firstorder Maximum (tDCE bBB) waveletHHL firstorder Range (DCE bBB) waveletLLH glszm ZoneEntropy (tDCE T)	0.34, [0.29, 0.40]
T & CB & bBB (DCE & T2)**	original glszm SizeZoneNonUniformityNormalized (tDCE bBB) original glszm SmallAreaLowGrayLevelEmphasis (tDCE bBB) waveletHHH firstorder Maximum (tDCE bBB) waveletHHH firstorder Mean (T2 T) waveletHLL glszm LargeAreaHighGrayLevelEmphasis (T2 CB)	0.28, [0.22, 0.33]
T (DCE & T2)	waveletLHH glcm ClusterShade (tDCE T) waveletLLL firstorder 10Percentile (T2 T) waveletLLH glszm ZoneEntropy (tDCE T) waveletLHH glcm MaximumProbability (tDCE T) waveletHHH glrlm LongRunLowGrayLevelEmphasis (tDCE T)	0.16, [0.10, 0.18]
T (T2)	original gldm DependenceEntropy (T) waveletLLL firstorder 10Percentile (T) waveletLLL gldm LargeDependenceHighGrayLevelEmphasis (T) waveletLLH glcm Imc1 (T) waveletLLH firstorder Median (T)	0.10, [0.03, 0.12]
T (DCE)	waveletLHH glcm ClusterShade (tDCE T) waveletLLH glszm ZoneEntropy (tDCE T) waveletLHH glcm MaximumProbability (tDCE T) waveletLLL glszm ZoneEntropy (T) waveletHLH firstorder Minimum (tDCE T)	0.03, [-0.02, 0.09]

** "BB & bBB (DCE)" and "bBB (DCE)" selected the same set of features. ** "T & CB & bBB (DCE & T2)" and "all (DCE & T2)" selected the same set of features.*

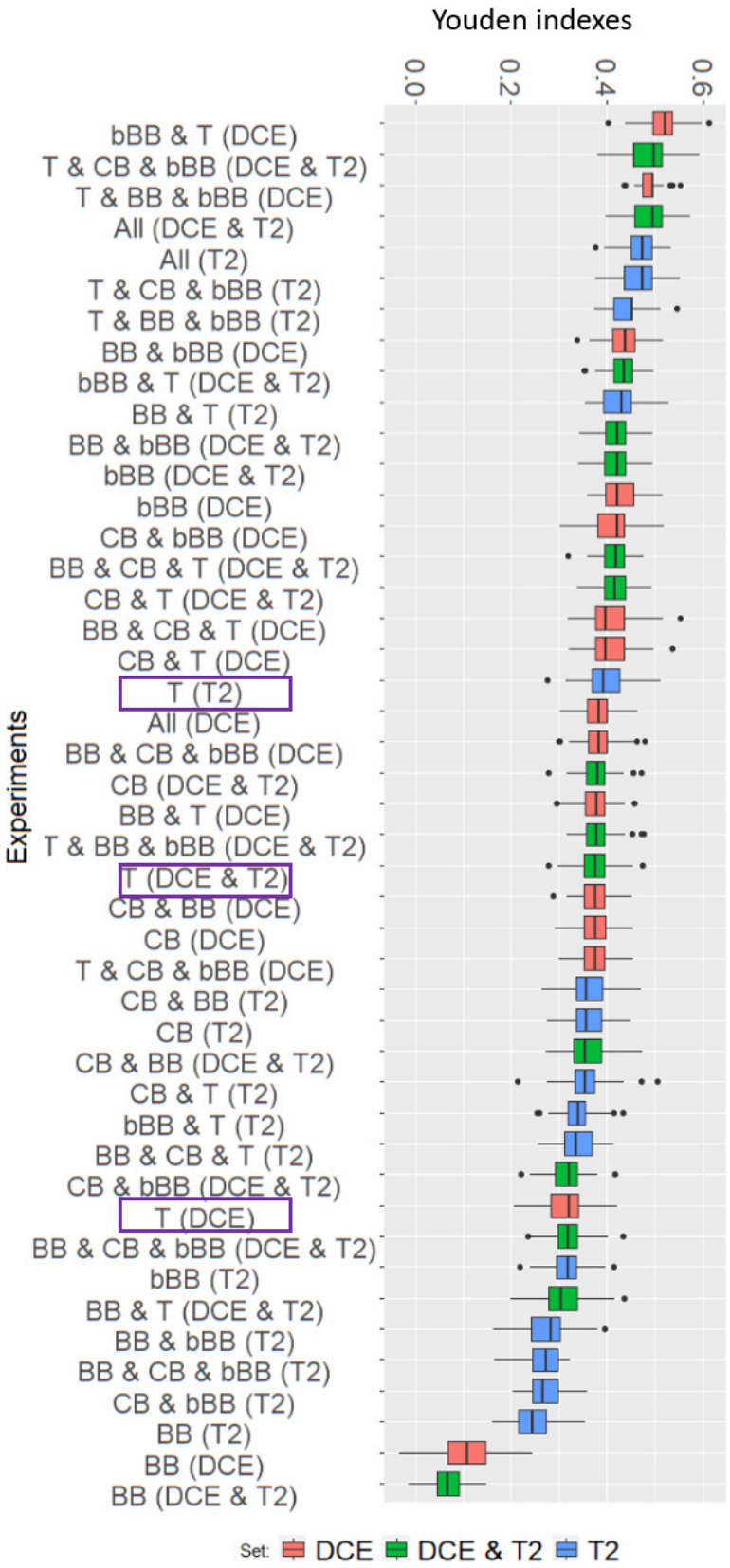


Figure 6.10: All 45 training experiments ranked according to the median of their Youden indexes. Tumor experiments (T (DCE), T (T2), T (DCE & T2)) are in boxes to spot them easily.

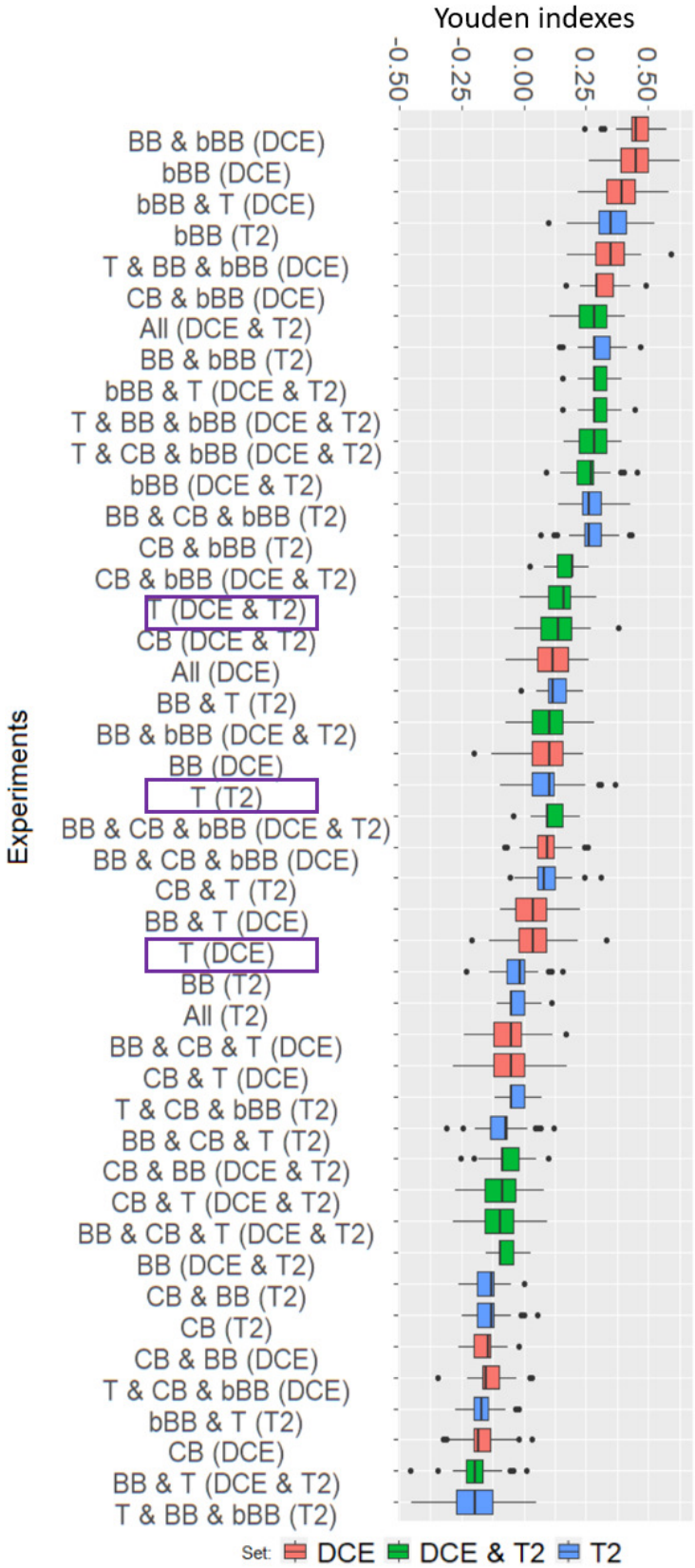


Figure 6.11: All 45 testing experiments ranked according to the median of their Youden indexes. Tumor experiments (T (DCE), T (T2), T (DCE & T2)) are in boxes to spot them easily.

6.3 Molecular subtype-specific models

6.3.1 Introduction

As described in Chapter 2, the different molecular subtypes of breast cancer have varied characteristics, prognoses and responses to treatments. NAC has become the standard of care for Luminal B, HER2-enriched and triple-negative cancers. Though some radiomic studies [14, 16] included Luminal A tumors, it is a rarer occurrence as they have very low chances of achieving pCR ($< 7.5\%$) [6].

Some studies have designed molecular subtype-specific models to create more homogeneous databases with the hope of revealing effects that would exist more strongly in a certain subtype and that would be swamped when mixing all tumors together. Improved performances were reported in [134]. Most of the time, studies focused on HER2-enriched specific models [187, 188, 206] or models designed for TN tumors [15, 167, 198]. HER2-enriched tumors follow a particular course of treatment including Herceptin on top of the standard Anthracycle/Taxane regimen, which explains the constitution of a distinct database for this subtype. Similarly, TN tumors are extremely aggressive with the worst prognosis and often considered separately from other subtypes.

Nevertheless, to refine the dataset while keeping as many patients as possible to preserve statistical power, common division of TN/HER2+ (TN, HER2+, Luminal B/HER2+) [144, 150] versus HR+/HER2- (Luminal A, Luminal B/HER2-) [134, 144] tumors has been experimented on. Compared with HR+/HER2- tumors, TN/HER2+ patients have higher chances of achieving pCR and a better overall survival associated with it [150].

6.3.2 Methods

Considering the molecular subtype composition of our database, a single model designed for TN/HER2+ tumors was built (Table 6.6). The training set collected 73 patients (42 pCR/31 npCR) and the test set 24 patients (14 pCR/10 npCR). To keep up with our rule of thumb of one feature for every 10 pCR patients in the training set, previously detailed in Chapter 3, the number of features was set to four.

The same VOIs and prediction pipeline including the harmonization strategy for the test set defined in Figure 6.3 and previous sections were used. Random forest models were however tuned to take into account the slight class imbalance and they were evaluated with AUC and 95% CI. AUCs were used to facilitate comparisons with performances from the literature reported in Table 2.3. Three rounds of experiments (DCE, T2, DCE & T2) were carried out like in the previous section.

Molecular subtype	Training set (n=103)	Test set (n=33)
TN	48	14
HER2+-enriched	12	7
Luminal B/HER2+	13	3
Luminal B/HER2-	30	9

Table 6.6: Distribution of molecular subtypes between the training and test sets including all subtypes.

6.3.3 Results

Among the three rounds of experiments (T2, DCE, DCE & T2), best results on the training set were obtained when using both T1-DCE and T2 sequences. AUCs of DCE & T2 experiments are reported in Figure 6.12. Best models performances are summarized in Table 6.7 to which is added the performances of the tumor (T) experiment and the best model for all subtypes obtained in Section 6.2.

Table 6.7: Training and test performances of radiomic models of interest.

TN/HER2+ Models	AUC, [95%CI] (training)	AUC, [95%CI] (testing)
CB & BB	0.87, [0.79, 0.95]	0.66, [0.43, 0.90]
CB	0.86, [0.78, 0.94]	0.68, [0.45, 0.91]
T & CB & bBB	0.86, [0.78, 0.94]	0.76, [0.55, 0.96]
BB & CB & bBB	0.86, [0.78, 0.94]	0.79, [0.60, 0.98]
all	0.85, [0.76, 0.94]	0.77, [0.57,0.96]
CB & bBB	0.85, [0.76, 0.94]	0.76, [0.55, 0.96]
T	0.73, [0.62, 0.84]	0.51, [0.26, 0.70]
Best global model	AUC, [95%CI] (all subtypes training)	AUC, [95%CI] (all subtypes testing)
bBB & T (DCE)	0.80, [0.71, 0.89]	0.72, [0.53, 91]

6.3.4 Discussion

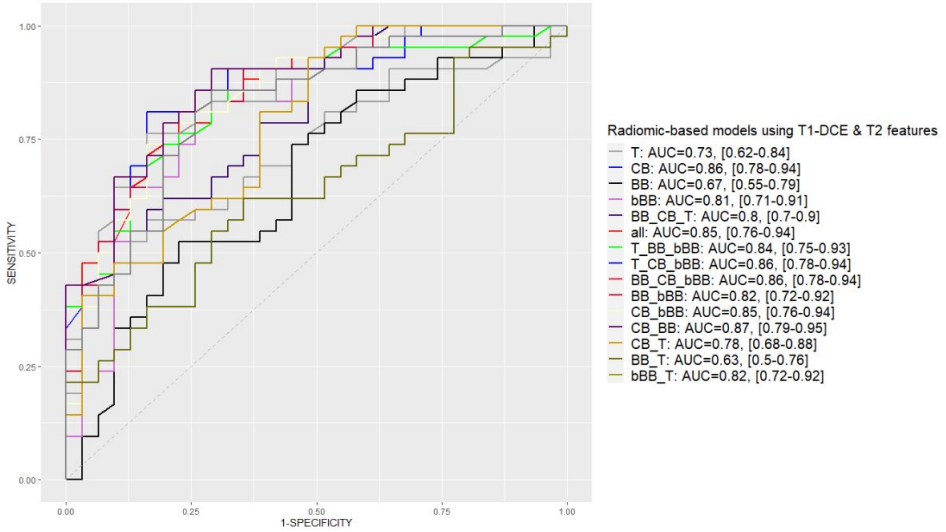
On the cohort restricted to the TN/HER2+ patients, several models ("CB & BB", "CB", "T & CB & bBB", "all", "BB & CB & bBB", "CB & bBB") reported AUC values around 0.86 [0.78, 0.94] with no significant differences between them. Their respective results on the test were spread inside the range [0.66, 0.79]. As observed in Section 6.2, better performances on the training set were achieved by using a combination of features from different VOIs or from another type of VOI (in this case, the constant box) than by using the standard tumor segmentation (training AUC = 0.73, [0.62, 0.84]). On the test set, the best performances were achieved in the model "BB & CB & bBB" (0.79, [0.60, 0.98]).

On the cohort gathering all molecular subtypes, presented in Section 6.2, the "bBB & T" experiment achieved the best results on the training set. Using AUC metrics, "bBB & T" reached on the training set an AUC = 0.80, [0.71, 0.89] and on the test set an AUC of 0.72 [0.53, 0.91]. Thus, there seems to be a slight trend in achieving higher performances with subtype-specific models than with global signatures, as reported in [16, 134, 144], though no significant statistical results could be obtained.

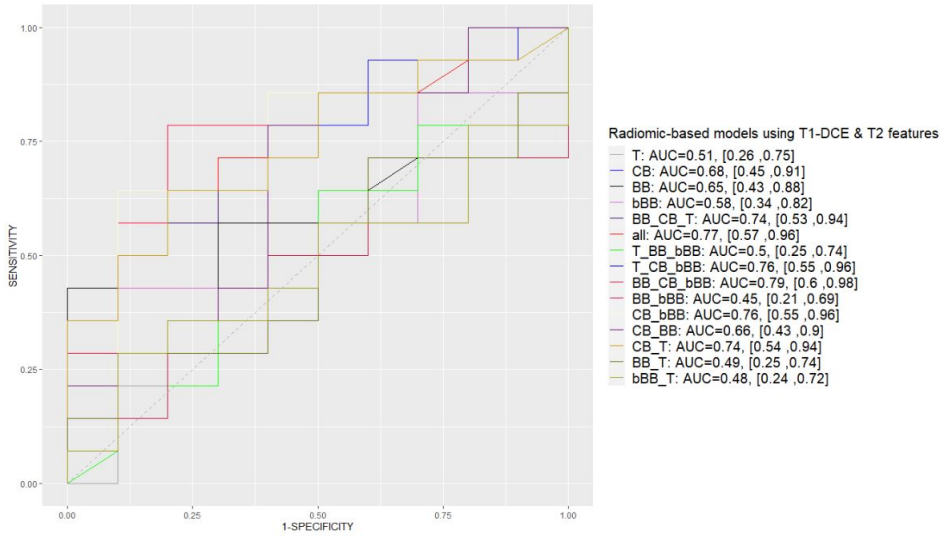
Comparing our performances with the literature remains tricky as when subsetting their dataset, many studies drops the evaluation of the model on an independent test set in favor of cross-validation on a new subtype-specific training set. The impact of the "scanner effect" in our multicentric dataset must also not be forgotten. On a multi-scanner independent test set, Cain et al. [150] reported an AUC=0.70 $\pm$ 0.06 for their TN/HER2+ model while

Braman et al. [144] achieved an AUC=0.83 $\pm$ 0.03 with CV on a 47-patient training set. The performances obtained in our experiments are thus in par with the literature.

Our experiments are however limited by the small size of our dataset and especially of the test set (n=24). HR+/HER2- specific models should also be tested while other divisions of the subtypes could be investigated.



(a) Training



(b) Testing

*Figure 6.12: Evaluation with AUC and 95% CI of radiomic-based models using T1-DCE & T2 features on **(a)** training set; **(b)** test set.*

Conclusion

In this chapter, a large number of experiments were conducted to investigate which information found in MR images could be relevant to predict pCR to NAC. Based on several types of VOIs (tumor, bounding box surrounding tumors, bounding box on binarized tumor images, constant box inside tumors), experiments highlighted that the information contained by the shape and margins of the lesions, conveyed by performing advanced texture analysis on a binarized version of the images, could be of great interest for the prediction. Results also suggested that combining features from different VOIs could improve performances and outperform models based on the standard tumor delineation.

The experiments also underscored the difficulty of exporting radiomic models to multi-centric test sets in the context of the "scanner effect". The original harmonization strategy developed offered an encouraging boost to test performances but still needs to be tested in other situations.

The experiments did not allow us to reach a definite opinion on the added benefit of multiparametric signatures. A slight trend of achieving better results in subtype-specific models than in models based on all subtypes has been reported but further investigation should be carried out.

Chapter 7

Automatic segmentation

Preface

This chapter introduces a deep learning-based ensemble approach to segment breast tumors on T1-weighted DCE images. Methods and results were reported in an article that was accepted for publication in *European Radiology* [26]. Some descriptions of the image acquisition process and other information about parameters and radiologist inputs defined in Chapter 3 are repeated in the article.

This work was conducted in collaboration with Michel Koole and Masoomeh Rahimpour from the Nuclear Medicine & Molecular Imaging team of KU Leuven university (Belgium). My contribution to this work involved the design of the study, data management and preprocessing, conducting the evaluation process and statistical analyses, researching the literature and writing the manuscript.

7.1 Introduction

In handcrafted radiomic analyses, the segmentation of lesions is a crucial step. Segmenting lesions in 3D is a time-consuming and tedious task for radiologists. As it requires the input of specialists, it can constitute a bottleneck from a time, practical and sometimes even financial point of view. Being able to segment automatically or semi-automatically the tumors would considerably alleviate radiologists' workload and facilitate the development of radiomic studies. Besides, depending on their relative experience, training or simply their preferences, radiologists will segment lesions differently. However, inter-radiologist variabilities in segmenting lesions affect radiomic feature values.

The cohort introduced in this chapter does not completely match the 136-patient dataset used to build the predictive models in Chapters 3 and 5. The modality requirements were indeed different as the segmentation approach requires both the first post-contrast T1-DCE image and its subtraction image but do not need T2 images. A few scans identified at a later stage as mid-course MRI were also kept in the cohort for the segmentation task while they were excluded from radiomic analyses. Furthermore, the split between training and test sets was made differently than in Chapter 3, to keep the 30 tumors segmented by both radiologists in the test set in order to compare results with inter-operator reproducibility.

7.2 Article - Rahimpour, Saint Martin et al., Eur Rad, 2022.

Visual ensemble selection of deep convolutional neural networks for 3D segmentation of breast tumors on dynamic contrast enhanced MRI.

ACCEPTED in European Radiology on August 14^{th}, 2022.

Masoomeh Rahimpour[1]*, Marie-Judith Saint Martin[2]*, Frédérique Frouin[2], Pia Akl[3], Fanny Orlhac[2], Michel Koole[1]*, Caroline Malhaire[2,4]*

[1] Department of imaging and pathology, KU Leuven, Belgium
[2] U1288-LITO, Inserm, Centre de Recherche de l'Institut Curie, Université Paris-Saclay, Orsay, France
[3] Department of Radiology, Hôpital Femme Mère Enfant, Hospices civils de Lyon, Lyon, France
[4] Department of Radiology, Ensemble Hospitalier de l'Institut Curie, Paris, France

* Masoomeh Rahimpour and Marie-Judith Saint Martin equally contributed to the paper
* Michel Koole and Caroline Malhaire equally contributed to the paper

Abstract

Objectives: To develop a visual ensemble selection of deep convolutional neural networks (CNN) for 3D segmentation of breast tumors using T1-weighted dynamic contrast enhanced (T1-DCE) MRI.
Methods: Multi-center 3D T1-DCE MRI (n=141) were acquired for a cohort of patients diagnosed with locally advanced or aggressive breast cancer. Tumor lesions of 111 scans were equally divided between two radiologists and segmented for training. The additional 30 scans were segmented independently by both radiologists for testing. Three 3D U-Net models were trained using either post-contrast images, or a combination of post-contrast and subtraction images fused either at the image or feature level. Segmentation accuracy was evaluated quantitatively using the Dice Similarity Coefficient (DSC) and the Hausdorff distance (HD95) and scored qualitatively by a radiologist as excellent, useful, helpful, or unacceptable. Based on this score, a visual ensemble approach selecting the best segmentation among these three models was proposed.
Results: Mean and standard deviation of DSC and HD95 between the two radiologists were equal to 77.8±10.0% and 5.2±5.9 mm. Using the visual ensemble selection, a DSC and HD95 equal to 78.1±16.2% and 14.1±40.8 mm was reached. The qualitative assessment was excellent (resp. excellent or useful) in 50% (resp. 77%).
Conclusion: Using subtraction images in addition to post-contrast images provided complementary information for 3D segmentation of breast lesions by CNN. A visual ensemble selection

allowing the radiologist to select the optimal segmentation obtained by the three 3D U-Net models achieved comparable results to inter-radiologist agreement, yielding 77% segmented volumes considered excellent or useful.

Key words: Breast Neoplasms; Magnetic Resonance Imaging; Neural Networks; Computer; Image Processing; Computer-Assisted.

Key points:

- Deep convolutional neural networks were developed using T1-weighted post-contrast and subtraction MRI to perform automated 3D segmentation of breast tumors.
- A visual ensemble selection allowing the radiologist to choose the best segmentation among the three 3D U-Net models outperformed each of the three models.
- The visual ensemble selection provided clinically useful segmentations in 77% of cases, potentially allowing for a valuable reduction of the manual 3D segmentation workload for the radiologist and greatly facilitating quantitative studies on non-invasive biomarker in breast MRI.

Abbreviations:
CNN: Convolutional Neural Network
DCE: Dynamic Contrast Enhanced
DSC : Dice Similarity Coefficient
HD95 : 95th percentile of Hausdorff Distance
ReLU: Rectified Linear Unit
SubT1: Subtraction image (first post-contrast DCE-MRI minus pre-contrast DCE-MRI)
T1c: first post-contrast DCE-MRI

Introduction

MR imaging, alongside mammography, is one of the standard imaging modalities for the detection, diagnosis, and treatment follow-up of breast cancer [254]. Dynamic contrast-enhanced MRI (DCE-MRI) is commonly used in quantitative analysis such as radiomic studies [8] to assess the malignancy of breast lesions, tumor extension, or predict their response to neoadjuvant therapy [222]. The analysis requires a precise segmentation of the breast tumor, but a manual delineation of lesion is time-consuming, often tedious and prone to inter-and intra-radiologist variability [255]. It frequently constitutes a bottleneck for the quantitative analysis of larger imaging studies using breast MRI. By providing an easy access to robust 3D quantitative features extracted from tumoral lesions, an automated 3D tumor segmentation would considerably improve the identification of non-invasive biomarkers in breast MR imaging.

The recent rise of deep learning methods has brought a renewed interest to tackle organ and lesion segmentation [255]. Deep convolutional neural networks (CNNs) have established themselves as state-of-the-art methods to segment medical images in 2D [256, 257] and in 3D [258, 259]. Many public databases and segmentation challenges are available online to train and test CNN models. Although the Medical Segmentation Decathlon [260] intends to build models that could segment multiple organs using different imaging modalities, most challenges focus on specific lesions such as brain tumors with The Brain Tumor Segmentation (BraTS) Challenge [261] or liver with the Liver Tumor Segmentation (LiTS) Challenge [262] benchmarks. To the best of our knowledge, no challenge for breast tumor segmentation using DCE-MRI has been reported. There are fewer studies using deep learning methods to segment breast tumors using DCE-MRI than using mammograms, partly due to the availability of very large mammography datasets [263]. Studies based on DCE-MRI used well established CNN segmentation models [264–267] based on U-Net [256], DeepMedic [268] or SegNet [269] architectures or less common models [270, 271]. Several studies [265, 267, 270, 272] took advantage of all the information given by the DCE-MRI by using the different post-contrast or subtraction (post-contrast minus the pre-contrast acquisition) images. For instance, Piantadosi et al. [272] used images from three different time points (pre-contrast, first and last post-contrast images). In the same way, Hirsch et al. [267] built several models taking different post-contrast images as input while Zhang et al. [270] fed both post-contrast and subtraction images as input to a hierarchical CNN

Though all these studies aimed to integrate segmentation results into a clinical workflow, the practical evaluation was only based on quantitative criteria. However, a visual assessment is still necessary to detect outliers, and should be integrated in the evaluation process. The key objective of this study was therefore to define a clinically useful tool to assist radiologists in breast lesion segmentation on DCE-MRI. Three different 3D U-Nets models were considered using either the first post-contrast T1 DCE-MRI (denoted T1c) or a fusion of T1c and subtraction images (denoted SubT1), with SubT1 images defined as the difference between the first post-contrast image and the pre-contrast image. Fusion of T1c and SubT1 images was implemented at both the image and features level, resulting in three 3D U-Net models. These models were trained and the visual ensemble selection was considered where the most optimal segmentation was selected visually by a radiologist to take advantage of the complementarity of the different U-Net models and to select the best segmentation for each patient.

Materials & Methods

Image database and ground truth definition

Breast MR images (n=141) were collected from a cohort of women diagnosed with locally advanced or aggressive breast cancer (see Table 7.1 for clinical characteristics) and undergoing neoadjuvant chemotherapy in Institut Curie between 2016 and 2020. This retrospective study was approved by our institutional review board (IRB number OBS180204) and written informed consent was waived for it. The 3D T1 fat suppressed DCE-MR images were acquired in a multi-center setting, with the majority of scans (77%) coming from Institut Curie with three acquisition devices (see Table 7.2). A dedicated breast coil was used in all cases. For DCE-MRI, gadolinium-based contrast material was injected using a power injector, followed by a saline solution flush. Representative acquisition parameters for T1 fat-suppressed DCE

sequences are given in Supplemental Table 7.6. On the whole database, in-plane voxel size varied between 0.62x0.62 and 1.0x1.0 mm, while voxel thickness ranged from 0.7 to 2.2 mm. The MRI performed outside Institut Curie were reviewed to control the quality of the images and the compliance with the recommendations of the American College of Radiology for the performance of contrast-enhanced MRI of the breast [273].

A set of 111 tumoral lesions was evenly segmented in 3D by two radiologists (see Supplemental Figure 7.5). Radiologist R1 had 15 years of experience in breast imaging while radiologist R2 had 3 years of experience. Tumors were manually segmented using the LIFEx software (v6.0, www.lifexsoft.org) [235] and were used as ground truth labels for training and validating the CNN models. The remaining 30 lesions were segmented by both radiologists and defined as the test dataset.

Table 7.1: Clinical information related to the 141 breast scans involved in the study. Quantitative features are given by mean values ± standard deviation, qualitative features are given by the number of cases (percentage).

Age of patients	48±11 years
Largest diameter of tumor	29±13 mm
Primary Tumor : T Stage I / II / III / IV	34 (24%) / 83 (59%) / 19 (13%) / 5 (4%)
Regional Lymph Node: N Stage 0 / I / II	77 (55%) / 62 (44%)/ 2 (1%)
Distant Metastasis: M Stage 0 / I	139 (99%) / 2 (1%)
Tumor type	Ductal NOS 137 (97%) / Others 4 (3%)
Breast Cancer Subtype Luminal / HER2+ / TN	41 (29%) / 37 (26%) / 63 (45%)

NOS: Not Otherwise Specified; HER2+: Human Epidermal growth factor Receptor 2 positive; TN: Triple-negative.

Table 7.2: MRI scanners and breast coils used to acquire the training and test databases.

MRI settings	**Database**	**Cases**
Institut Curie - GE Healthcare - 1.5T 8 channel breast coil	Training	13
Institut Curie - Siemens Healthineers - 1.5T Sentinelle breast coil	Training	50
Institut Curie - Siemens Healthineers - 1.5T 18 channel breast coil	Training	16
External Centres (n=10) - GE Healthcare - breast coil1.5T breast coil	Training	21
External Centres (n=6) - Siemens Healthineers - breast coil1.5T breast coil	Training	11
Total	**Training**	**111**
Institut Curie - GE Healthcare - 1.5T 8 channel breast coil	Test	13
Institut Curie - Siemens Healthineers - 1.5T Sentinelle breast coil	Test	13
Institut Curie - Siemens Healthineers - 1.5T 18 channel breast coil	Test	4
Total	**Test**	**30**

Image preprocessing

All MR images were corrected for bias field gain using the N4 algorithm as described in [19], resampled to get isotropic 1x1x1 mm^3 voxels across the whole database then cropped in a fixed size bounding box (300x160x200 mm^3) ensuring that the whole breast area and armpit were included in the images. Next, images were resampled to the voxel size of 2 mm to reduce memory requirements for the segmentation model. In addition, images were normalized by dividing the intensity values of each image volume by the 95^{th} percentile of its intensity values to avoid a normalization based on intensity outliers.

Segmentation models

The basic architecture of the models was a 3D U-Net similar to the implementation in No New-Net [258]. The U-Net contained 4 pathways, each consisting of 2 convolutional layers with kernel size of (3,3,3) and (3, 3, 1) (see Supplemental Figure 7.6 and Supplemental Table 7.7). All convolutional layers were followed by an instance normalization and a leaky rectified linear unit (Leaky ReLU) activation function. Two fully-connected layers followed by a softmax layer were added as final layers to classify the image voxels into healthy or tumoral tissue. Three different configurations of the U-Net model were elaborated. The first model (referred as "U-Net (T1c)") was trained by the T1c image while the other two models were trained by a combination of the first post-contrast and the first subtraction images using an image or feature level fusion strategy to combine images. For the image-level fusion approach (denoted "U-Net ILF (T1c+SubT1)"), the two both MR images were concatenated to form defined a single dual-channel, before being used as input for the CNN model. For the feature-level fusion approach (abbreviated "U-Net FLF (T1c+SubT1)"), a U-Net architecture was used in which the encoder part consisted of two independent channels fed by the post-contrast and subtraction images, respectively. In the bottleneck of the U-Net, feature maps were concatenated and provided as the input to the decoder part, as illustrated in Figure 7.1. The models were implemented using DeepVoxNet [274], a high-level framework based on Tensorflow/Keras but specifically designed and optimized for 3D medical image data. All models were trained using a combined loss function L (defined by Equation 7.1) defined as a weighted combination of cross-entropy (L_{CE}) and soft Dice (L_{SD}) losses [275]:

$$L = \alpha L_{CE} + (1 - \alpha) L_{SD} \tag{7.1}$$

α being the weighting factor of the two loss terms. For training and validation, the Adam optimizer with default Keras settings (v 2.2.4 with Tensorflow backend) was used with the initial learning rate set at 0.001. When the validation Dice Similarity Coefficient (DSC) reached a plateau, the learning rate was reduced by a factor of 5, and training was stopped when the DSC on the validation dataset did not improve during the last 500 epochs. For this implementation, a single epoch consisted of feeding 12 entire image volumes to the model with a batch size of 2. All computations were performed on the Flemish supercomputer (CentOS Linux 7) using 2 NVIDIA P100 GPUs (CUDA v11.0, GPU driver v450.57) and 1 Intel Skylake CPU (18 cores). During training, a five-fold cross-validation was performed to determine the optimal number of epochs and a grid search was performed within a range of [0.1, 0.9] and a step size of 0.1 to find the optimal value for the hyperparameter α. The

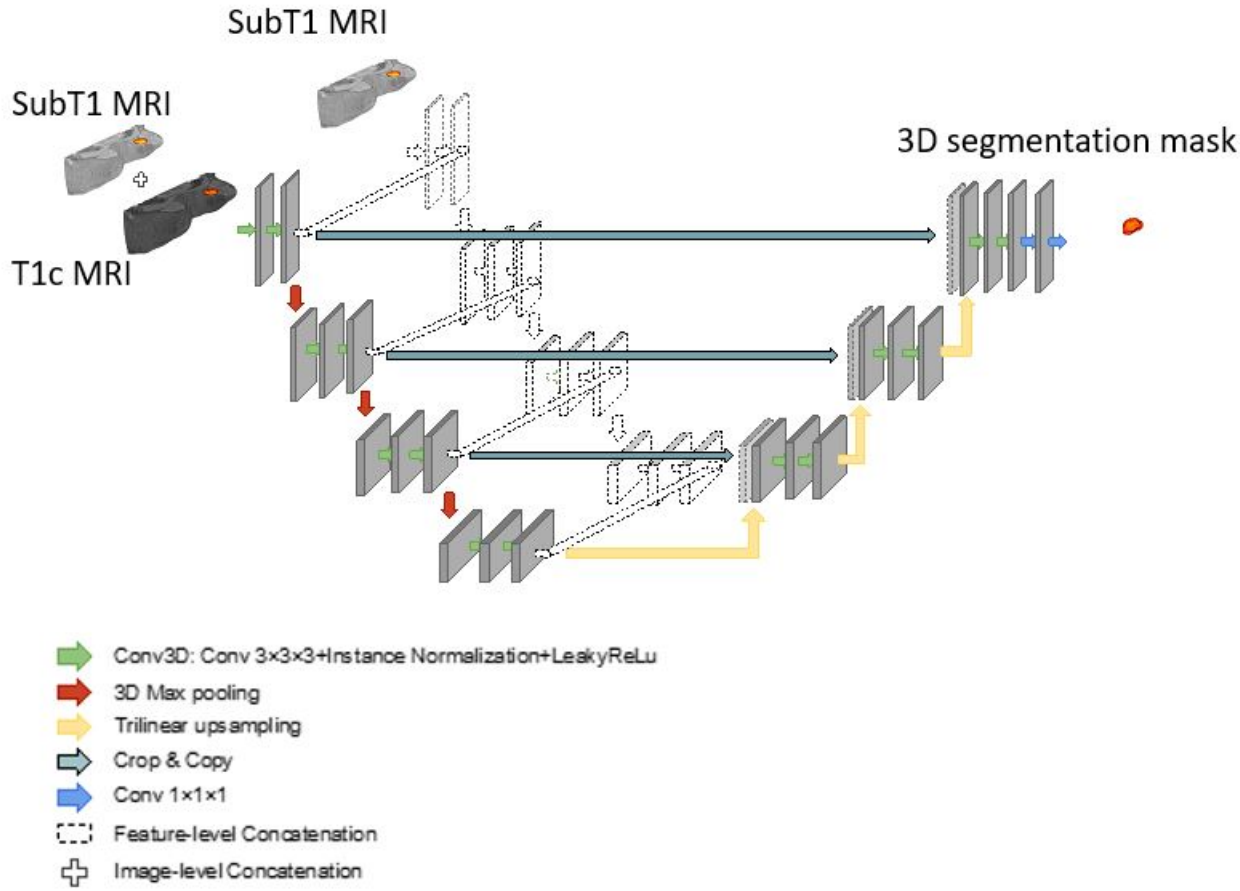


Figure 7.1: Schematic description of U-Net architecture used for Image-level fusion (ILF) and Feature-level fusion (FLF). The colored part represents the ILF where the T1c and SubT1 images are concatenated before being used as input for the CNN model. The dotted part is added to implement the FLF where T1c and SubT1 images are used as the input to two separate encoding parts and the extracted features from each level are concatenated.

highest DSC was achieved when α was set to 0.5 to appropriately weight soft Dice and cross-entropy loss functions. At the end of the training/validation, five models were saved and then used to generate the predictions on the test dataset. For final performance comparisons, the segmentation masks were averaged over the five models of the five-fold cross-validation, and then were up-sampled to a 1 mm^3 voxel size for comparison with the ground truth labels.

Visual ensemble selection

Radiologist R1 visually assessed the quality of the automated segmentations obtained by the three models: U-Net (T1c), U-Net ILF (T1c+SubT1) and U-Net-FLF (T1c+SubT1) and scored the segmentation quality as 'excellent', 'useful', 'helpful', or 'unacceptable'. Score 4 was given to excellent segmentations that could be used clinically without further modification. Score 3 was given to useful segmentations for which modifications (less than 25% of the total number of slices) could be achieved in a reasonable time (less than 50% of the time required for segmentation from scratch). Score 2 was given to helpful results that require

substantial modifications on a larger number of slices (between 25% and 66% of the total number of slices). Score 1 was given to unacceptable results, corresponding to very large errors in the tumor delineation or cases for which tumor was not detected. A novel patient-centric approach denoted visual ensemble selection was thus defined where, for each patient, the best segmentation was selected by Radiologist R1, when the visual scores were identical.

Quantitative analysis

To compare segmentations, the volumes of the lesions, DSC measuring the percentage of overlap ranging from 0% (no overlap) to 100% (perfect overlap) and the 95th percentile of the symmetric Hausdorff distance (denoted HD95) measuring how far the two segmentations are distant from each other were calculated for each case of the test database. Inter-radiologist agreement was estimated by comparing the segmentations from R1 and R2. The 3D segmentations produced by the three U-Net models and the visual ensemble selection were compared to the ground truth labels defined by R1 and R2.

Statistical analysis

Statistical analysis was performed using R software (version 4.1), with a significance level equal to 0.05. The distribution of DSC and HD95 values of segmentations obtained by the visual ensemble selection versus R1 and R2 were compared to the inter-radiologist DSC and HD95 using a Friedman test. The distribution of DSC and HD95 issued from the three 3D U-Net models were globally compared using the Kruskal-Wallis test according to the four qualitative scores and then compared using the Dunn's test and Bonferroni correction.

Results

Quantitative analysis

Table 7.3 shows the volumes of the lesions as assessed by the two radiologists, the three 3D U-Net models, and the visual ensemble selection on the test database. Table 7.4 provides the mean and standard deviation of DSC and HD95 for the comparison of R1 and R2 segmentations (inter-radiologist criteria) and the comparison of the three 3D U-Net models and the visual ensemble selection with the segmentations provided by either R1 or R2. Figure 7.2 illustrates these results, providing box plots for each configuration.

Qualitative analysis

Figure 7.3 and Supplemental Figure 7.7 show the boxplots obtained for DSC and HD95 according to the four quality scores of visual assessment. The three 3D U-Net models achieved comparable results with 27% to 33% of cases scored as excellent, 20% to 30% as useful, 20% to 27% as helpful, and 17% to 27% as unacceptable. When using the visual ensemble selection, 50% of cases were scored as excellent, 27% as useful, while only 23% of cases were scored as helpful or unacceptable. The global performance of the three 3D U-Net models was reduced by some outliers, while the visual ensemble selection reduced the number of outliers, which highlights the complementary role of the three 3D U-Net models.

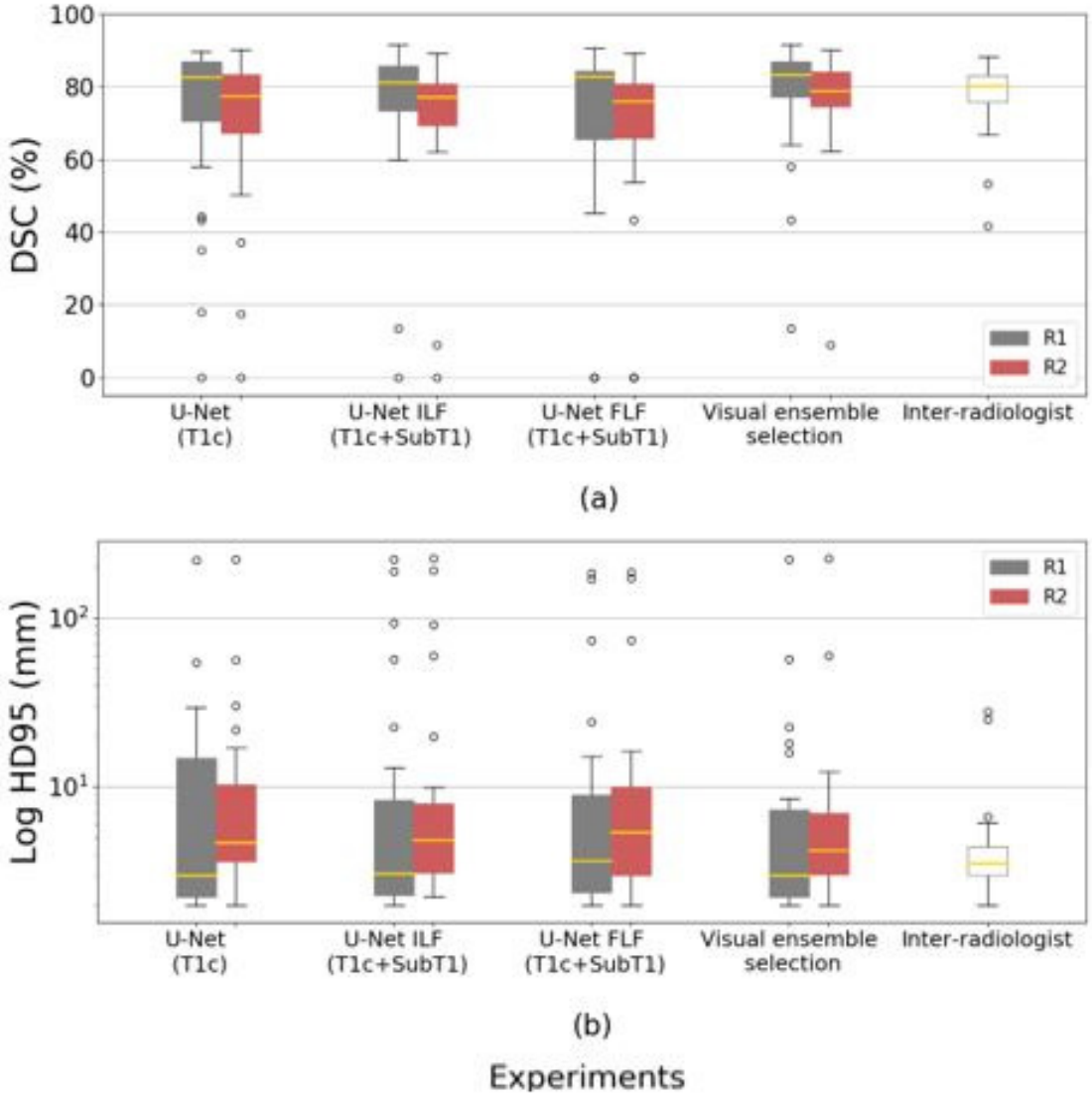


*Figure 7.2: Boxplot presenting **(a)** the DSC (%); **(b)** HD95 (mm) obtained by the different segmentation models on the test database: a U-Net using only T1c images, a U-Net trained by a combination using an image-level fusion of T1c and SubT1 images, a U-Net trained by a combination using a feature-level fusion of T1c and SubT1 images, and the visual ensemble selection. DSC and HD95 were determined using the manual delineation of the two independent radiologists (R1 and R2) as the ground truth labels. Inter-radiologist DSC and HD95 were added for comparison.*

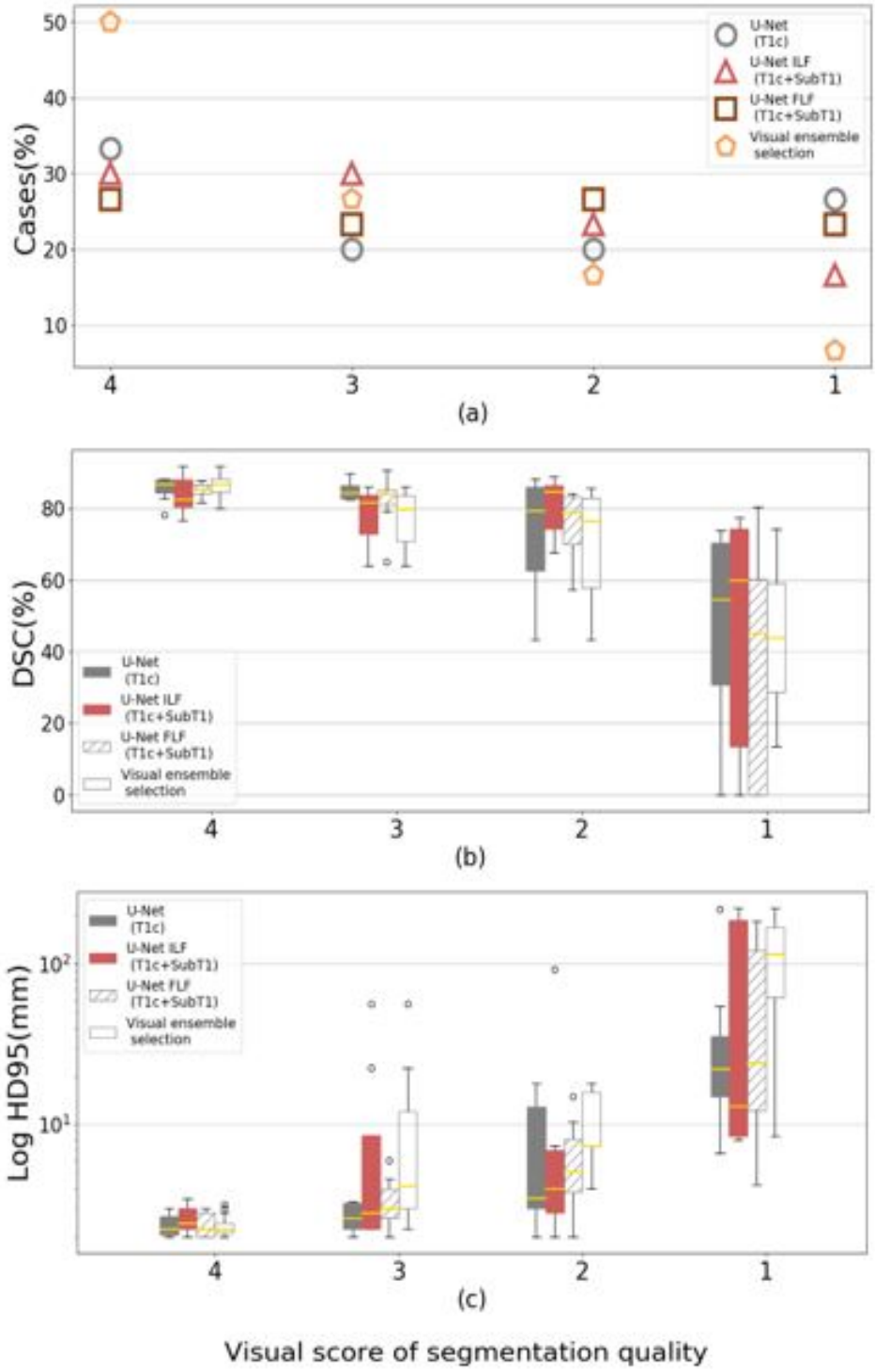


*Figure 7.3: Distribution of **(a)** automated segmentations according to the visual score 4 (Excellent), 3 (Useful), 2 (Helpful) and 1 (Unacceptable); together with the boxplots presenting **(b)** the DSC (%); **(c)** HD95 (mm). Results are shown for the different segmentation models: a U-Net using only T1c images, a U-Net using an image-level fusion of T1c and SubT1 images, a U-Net using a feature-level fusion of T1c and SubT1 images, and the visual ensemble selection.*

Table 7.3: Volumes of lesions (mean values ± standard deviation) as estimated by the two radiologists (R1 and R2), the three CNN models, and the visual ensemble selection on the test database.

Readers or Models	Volumes (cm3)
Radiologist R1	12.9±14.9
Radiologist R2	14.8±17.2
U-Net (T1c)	9.8±6.3
U-Net ILF (T1c +SubT1)	14.4±16.0
U-Net FLF (T1c+SubT1)	11.5±9.8
Visual ensemble	12.6±13.5

Table 7.4: Mean values ± standard deviation of quantitative criteria (DSC and HD95) to assess the segmentation provided by three CNN models and the visual ensemble selection, using either R1 or R2 as the ground truth on the test database.

	DSC (%)		HD95 (mm)	
Radiologist or Model	Radiologist R1	Radiologist R2	Radiologist R1	Radiologist R2
Radiologist R2	77.8±10.0		5.2±5.9	
U-Net (T1c)	72.7±22.8	70.6±20.8	15.6±40.3	15.9±40.6
U-Net ILF (T1c +SubT1)	74.9±20.3	71.9±19.7	22.9±53.2	23.6±53.6
U-Net FLF (T1c+SubT1)	70.2±26.1	67.3±25.0	19.3±45.1	19.8±45.4
Visual ensemble selection	78.1±16.2	76.5±14.5	14.1±40.8	14.1±41.2

Statistical analysis

For the segmentations obtained by three 3D U-Net models, the mean values of quantitative criteria (DSC and HD95) on the test database were significantly different ($p<0.0001$) for the unaceptable score (excellent, useful, helpful versus unacceptable) from the visual assessment provided by R1 (Supplemental Figure 7.7). Using R1 as the ground truth and based on paired rank analysis (Friedman tests), the DSC values between the segmentations provided by the visual ensemble selection and R1 were slightly better (p-value<0.03) than the DSC values computed from the segmentations provided by R1 and R2. There was no statistically significant difference for the HD95 results. Using R2 as the ground truth, the DSC values between the segmentation provided by the visual ensemble selection and by R2 were not significantly different from the DSC values computed using the segmentations provided by R1 and R2 (p-value=0.27).

Quantitative analysis according to visual assessment

Table 7.5 displays the mean and standard deviation of DSC and HD95 of the visual ensemble selection compared to the segmentations provided by either R1 or R2, according to the visual assessment. For the test cases scored as excellent, the mean DSC was higher than 81%, and with the standard deviation less than 6%, showing better results compared to the inter-radiologist DSC for the whole test database. Additionally, for these cases the mean HD95 was

less than 4 mm with the standard deviation less than 2 mm.

Table 7.5: Mean values ± standard deviation of quantitative criteria (DSC and HD95) to compare the segmentation provided by the visual ensemble selection, using either R1 or R2 as the ground truth, according to the four scores of qualitative assessment on the test database.

	DSC (%)		HD95 (mm)	
Visual ensemble selection	Radiologist R1	Radiologist R2	Radiologist R1	Radiologist R2
Score 4 - Excellent (n=15)	86.3±3.3	81.2±6.4	2.4±0.4	3.9±1.5
Score 3 - Useful (n=8)	76.9±8.7	76.7±8.1	13.1±18.9	11.4±19.7
Score 2 - Helpful (n=5)	69.3±18.1	77.4±5.8	10.6±6.1	8.1±3.1
Score 1 - Unacceptable (n=2)	43.9±43.0	39.3±42.9	116±151	117±154

Illustrative cases

Representative segmentation results of the test dataset are illustrated in Figure 7.4. These exemplified cases demonstrate the interest of the visual ensemble selection while highlighting the complementary role of the three 3D U-Net models. For instance, U-Net trained with T1c images could provide excellent results (case #1) and largely underestimated volumes (cases #2 and #3). For cases #2 and #3, the 3D U-Net using image-level fusion of T1c and SubT1 images as input, provided the best segmentation, scored as helpful (13 slices out of 30 need some correction) for case #2 and as useful (7 slices out of 37 need some correction) for case #3.

Discussion

We proposed a new CNN-based approach for breast tumor segmentation in a clinical setting. In our implementation, three 3D U-Net models were trained using different strategies: using only the post-contrast image or a combination of post-contrast and subtraction images using fusion at either the image or feature level. These three models were tested on 30 independent cases and none of them outperformed the other two. Following a subsidiarity principle, the best segmentation among the three was ultimately selected for each patient by the radiologist, defining a visual ensemble selection. Using appropriate display tools available in LIFEx [235], the additional workload required for the visual selection is low, compared to the time that is required to check one single segmentation carefully. Furthermore, the visual ensemble selection proves to provide acceptable segmentation results in 77% of the test cases and results are globally within inter-radiologist reproducibility.

Our approach provides a 3D segmentation of breast lesions, while some of the most recent studies still segment in 2D [264, 267, 272], despite tumor volume measured by MR imaging is a strong predictor of patient survival [276]. For advanced radiomic studies or follow-up studies, 3D segmentation is also an important task to achieve [222]. The CNN models were trained using multi-centric MRI, a prerequisite for a higher generalization of these models, and they were also evaluated using a multi-scanner test dataset. Compared to many studies, for which DSC was the only evaluation criteria [266, 267], HD95 was added as a criterion for the maximal distance between two segmentations. Contrary to DSC, this criterion was not included in the loss function for the training of the models and was therefore more independent. The

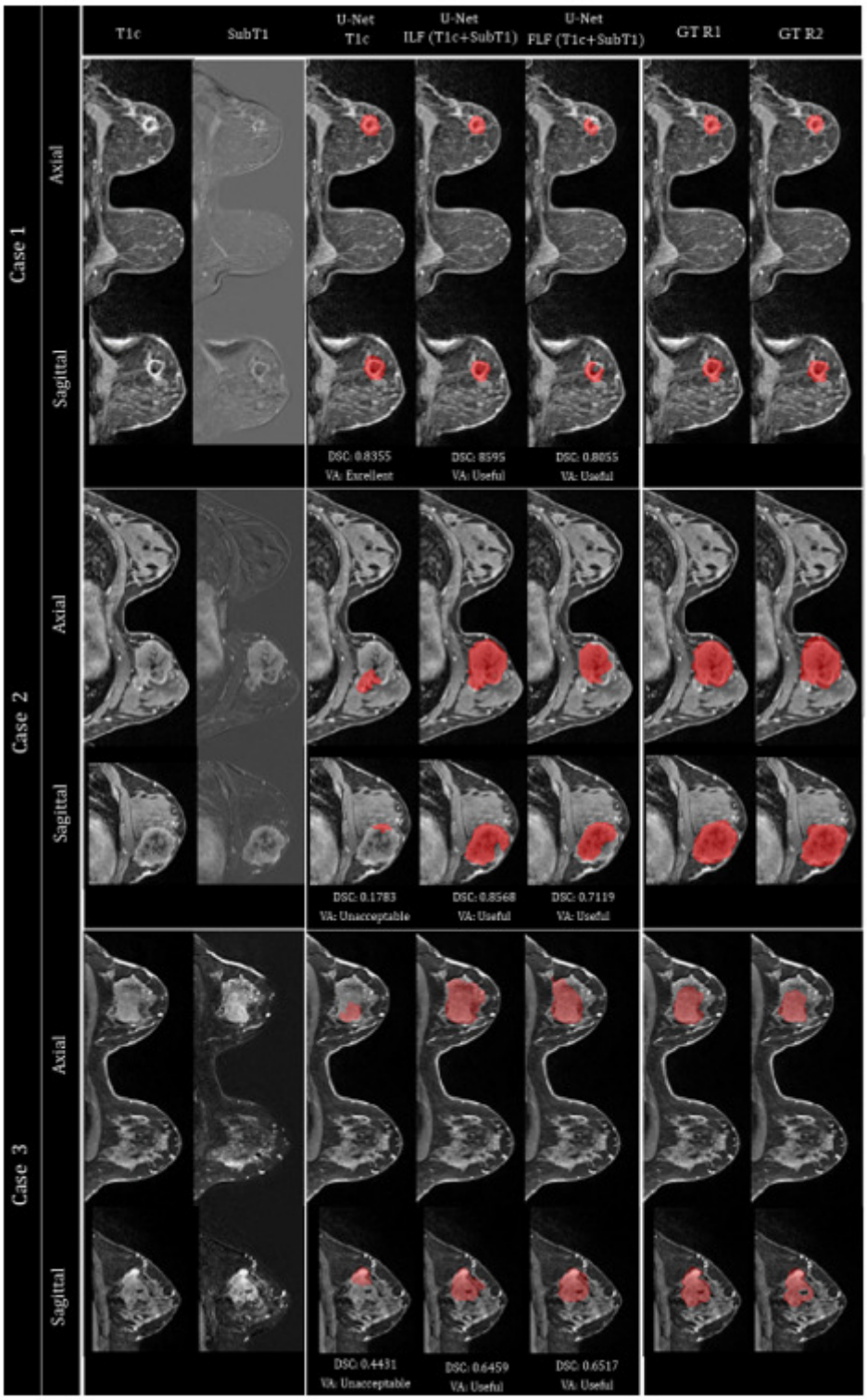


Figure 7.4: Illustration of representative segmentation results (axial and coronal views) on three cases of the test database. From left to right: T1c volume, SubT1 volume, segmentation provided by U-Net (T1c), U-Net ILF (T1c+SubT1), U-Net FLF (T1c+SubT1), ground truth (GT) provided by R1 and R2. DSC (%) and visual scoring (VS) given by R1 are included below each case. The visual ensemble selection corresponds to segmentation provided by U-Net (T1c) for case 1, and U-Net ILF (T1c+SubT1) for cases 2 and 3.

models designed in this study were based on the state-of-the-art U-Net architecture similar to the model proposed in [271] but without residual blocks. While Khaled et al. [271] generated a breast ROI mask during the pre-processing step and used it as the input to the segmentation model along with the 3D DCE-MRI, we did not provide the U-Net models with this prior information. The prior knowledge on the breast ROI mask was also used in [267] to train the CNN segmentation models with the U-Net architecture. We only used the T1c or/and SubT1 as the input to train the U-Net models, and not a full series of DCE-MRI for training as in [271], nor T1 and T2-weighed MRI sequences as in [267]. The Deepmedic architecture with a patch-based training method was evaluated in [267] demonstrating lower performance compared to the U-Net model. This evaluation confirms our choice to use the U-Net architecture. Furthermore, performance of our U-Net models was in the same DSC range [65%-80%] as reported in literature [204, 266, 267, 270–272], though it is difficult to compare methods evaluated on different datasets with DSC computed in 2D or in 3D. The mean 3D DSC between R1 and R2 was similar to the mean 3D DSC [78%-83%] for different observer combinations studied in [211]. Our database included locally advanced tumors or aggressive tumors, for which the irregular shape is difficult to segment.

The principle of an ensemble approach that combines the output of independently trained CNN models was also proposed in [271]. The authors compared a strategy of majority-voting and union operation to integrate the results of several CNN models trained with different post-contrast and subtraction images. We tested the automated ensemble approaches, but they did not improve final results (see Supplemental Table 7.8).

Despite the improved segmentation performance, our study had some limitations. The database used for training and testing was limited in terms of datasets but adding progressively new cases could gradually improve the performance of the different CNN models, even if the ideal number of cases is unknown. Further use of other post-contrast images needs to be investigated as well as the potential value of adding other modalities such as diffusion weighted images and apparent diffusion coefficient maps. Finally, the strategy we proposed is not fully automated and requires an additional visual assessment, but to the best of our knowledge, no current automated segmentation method included a self-assessment, even if a recent study [266] proposes solutions to address this issue.

Conclusion

This study proposes a visual ensemble selection as a new pragmatic segmentation method where the radiologist is asked to select the best segmentation among the results obtained by three different 3D U-Net models. This visual ensemble selection provided results comparable to inter-radiologist agreement with excellent or useful segmentations in 77% of the cases versus 60% of the cases for the 3D U-net model using image-level fusion of post-contrast and subtraction images, while it required little additional workload when compared to the visual evaluation of one single segmentation.

Supplemental data

Table 7.6: Representative values of MR acquisition parameters for the three settings at the Institut Curie.

	GE Healthcare (8-channel breast coil)	Siemens Healthineers (18-channel breast coil)	Siemens Healthineers (Sentinelle breast coil)
TR (ms)	6.81	5.2	5.2
TE (ms)	3.3	2.4	2.4
Slice thickness (mm)	1.0	0.9	0.9
Spacing between slices (mm)	1.0	0.9	0.9
Pixel spacing (mm)	0.82 x 0.82	0.91 x 0.91	0.91 x 0.91
Bandwidth	111	350	350
Pixel bandwidth	434	355	355
Flip angle	15	10	10
Fat saturation	DIXON	SPAIR	SPAIR
Parallel Imaging	ARC	GRAPPA	GRAPPA

Table 7.7: Parameters of U-Net model

Pathway	Block	Number of kernels	Kernel size
1 - EC	Conv3D	20	(3, 3, 1)
1 - EC	Conv3D	40	(3, 3, 3)
2 - EC	Conv3D	40	(3, 3, 1)
2 - EC	Conv3D	80	(3, 3, 3)
3 - EC	Conv3D	80	(3, 3, 1)
3 - EC	Conv3D	160	(3, 3, 3)
4 - EC	Conv3D	160	(3, 3, 3)
4 - EC	Conv3D	160	(3, 3, 3)
3 - DC	Conv3D	160	(3, 3, 3)
3 - DC	Conv3D	80	(3, 3, 1)
2 - DC	Conv3D	80	(3, 3, 3)
2 - DC	Conv3D	40	(3, 3, 1)
1 - DC	Conv3D	40	(3, 3, 3)
1 - DC	Conv3D	20	(3, 3, 1)
1 - FC	Conv3D	20	(1, 1, 1)
2- FC	Conv3D	1	(1, 1, 1)

EC, DC, and FC stand for encoder part, decoder part and fully-connected layer.

Table 7.8: Mean values ± standard deviation of quantitative criteria (DSC and HD95) to assess the performance of the automated ensemble approaches: majority voting and averaging. For the averaging, the segmentation predictions obtained by the three CNN models were averaged and thresholded by 0.5.

	DSC (%)		HD95 (mm)	
Automated ensemble approach	Radiologist R1	Radiologist R2	Radiologist R1	Radiologist R2
Majority Voting	74.5± 21.6	71.6±20.5	20.6±50.2	20.6±50.5
Averaging	74.9±22.0	72.3±20.8	14.0±25.13	14.24±25.1

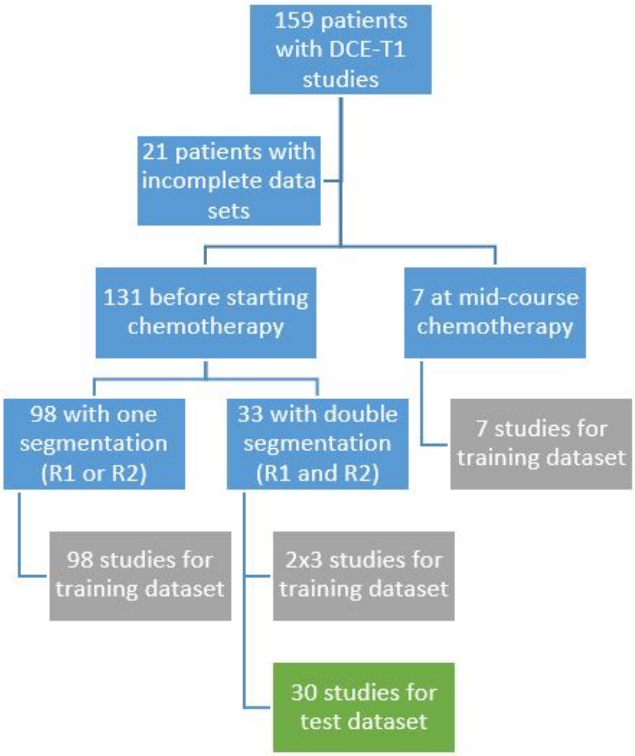


Figure 7.5: Flowchart for the definition of training and test datasets.

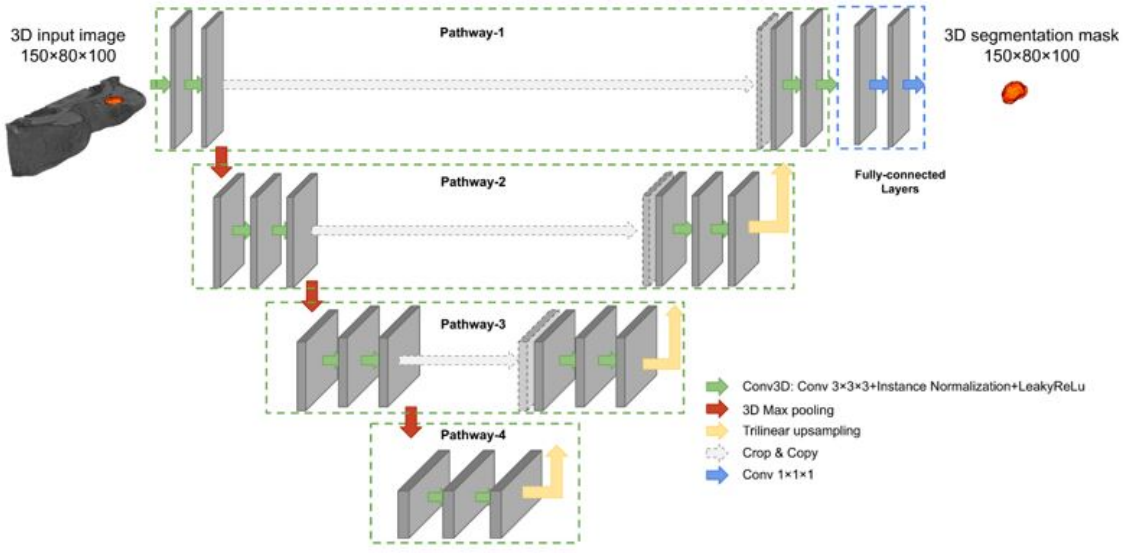


Figure 7.6: Overview of the 3D U-Net model.

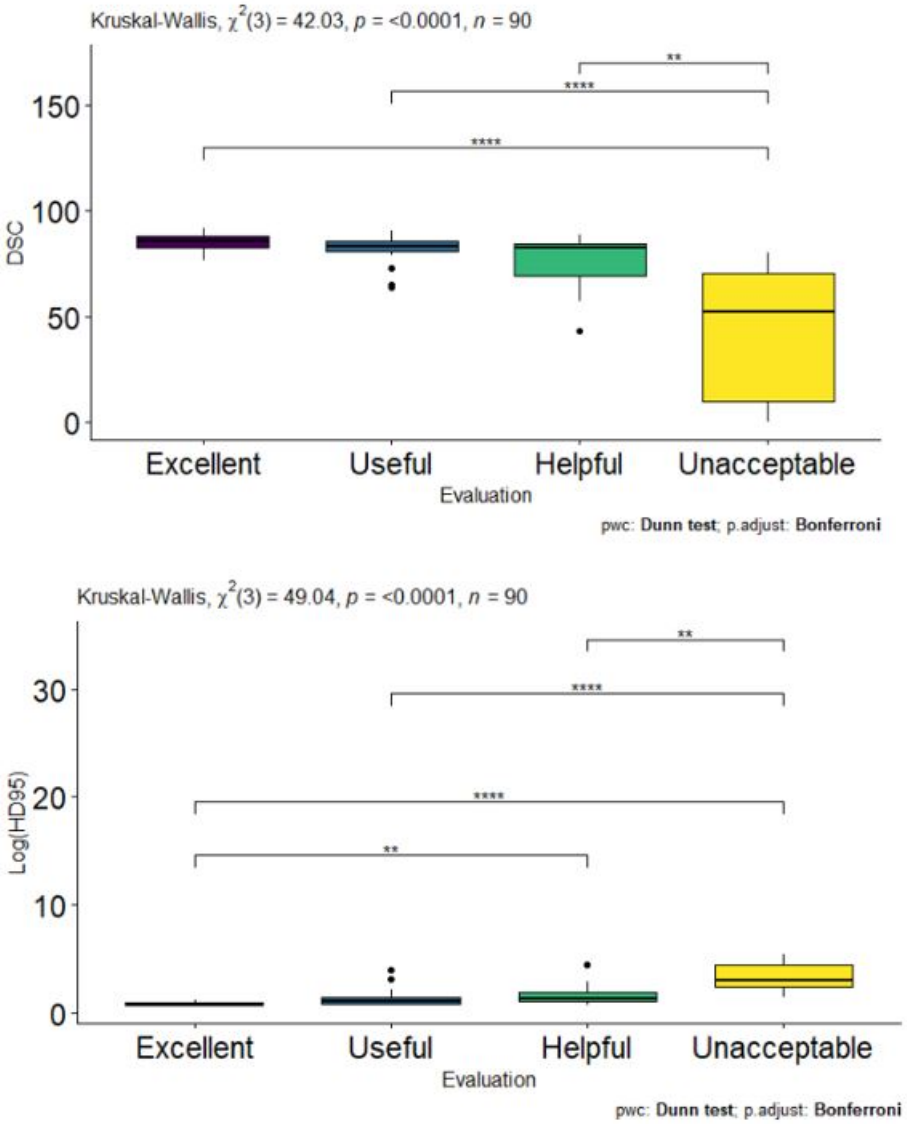


*Figure 7.7: Statistical analysis performed on the quantitative criteria (DSC and HD95) obtained by the three 3D U-Net models and using R1 as the ground truth according to the four visual scores (excellent, useful, helpful, and unacceptable), showing some significant differences of the mean values of DSC and HD95 according to the visual scores (****p-value<0.0001, ** p-value<0.01).*

7.3 Discussion

In the study, the different segmentation models and the ensemble approach were assessed with DSC or HD95 combined with a radiologist visual score. Using the 30 lesions segmented by both radiologists, agreement between each radiologist and each model, measured with the two metrics, was compared with the inter-radiologist agreement. As the goal of this model is to segment breast lesions for radiomic analyses, a radiomic-based evaluation of the methods could however be considered. Based on the 30 lesions segmented twice, the number of robust features (ICC $>$ 0.8 as defined in Chapter 5) obtained when using the two radiologists' segmentations could be compared with the number of robust features calculated when using automatic and radiologist segmentations.

This model was developed throughout this thesis as the database progressively increased in size and was therefore not used in the first radiomic analyses. However, inter-radiologist variabilities in segmenting lesions affect radiomic feature values which hinders the exportability of radiomic signatures. These considerations led to define procedures based on the ICC to select robust features (Chapter 5) and incorporate them in the global prediction pipeline. Taking into account the robustness of features to segmentation caused to discard more than 27% of the radiomic features calculated on native or wavelet-filtered images. Amongst the discarded features, 26 were associated with pCR ($p < 0.05$). Therefore, attempting to counter inter-radiologist variabilities resulted in rejecting potentially useful information for the prediction. Developing a segmentation approach and making it available to other researchers alongside our radiomic models would increase their exportability and allow us to use otherwise rejected features to build predictive models.

Conclusion

This chapter thus presents a patient-centric deep learning-based ensemble approach to segment breast lesions on T1-DCE MR images with performances on par with inter-radiologist agreement. This segmentation approach could facilitate radiomic studies and improve radiomic models' exportability. In future work, this approach will be used to segment mid-treatment images to study the changes in radiomic features between pre and mid-treatment.

Conclusion & future work

The prediction of the response to neoadjuvant chemotherapy in breast cancer has rose to prominence in recent years with the increasing development of precision medicine. Early identification of non-responders would constitute a major step forward in personalized treatment. An abundant literature on MRI-based radiomic signatures to predict pCR to NAC exists. Contradictory findings nevertheless emerged from the review of the literature. Pesapane et al. [14] found, for instance, no benefit in adding clinical and biological data to radiomic models while Peng et al. [200] and Hussain et al. [12] reported increased performances when combining clinical, biological and radiomic features. Whether to use pre-treatment or a mix of pre and mid-treatments images was questioned, as well as the benefit of multiparametric over single-sequence signatures, the necessity to build molecular subtype-specific models or the potential of peritumoral regions. The prediction of the response is thus a complex question that remains the subject of intense debates.

Beyond these considerations specific to the prediction of pCR to NAC, MRI-based radiomic analyses in general are faced with several challenges: local inhomogeneities in MR images, lack of standardized units in which to express MR signal, problems of reproducibility and exportability of features due to little robustness to segmentation and the impact of the "scanner effect". In [18], Granzier et al. concluded that the potential of radiomic features to predict pCR could not be properly investigated because of their lack of reproducibility. While the question of the feature robustness to segmentation is often addressed, little interest is shown towards normalizing MR images or harmonizing features to reduce the "scanner effect". Besides, testing of radiomic models on independent multicentric test sets remains limited in the literature.

Based on a cohort of 136 patients treated for NAC at Institut Curie, this thesis proposed methodological solutions to develop robust and exportable MRI-based radiomic models trained on a multi-scanner training set and tested on an independent multicentric test set, while investigating the potential of radiomic features extracted from pre-treatment T1-weighted DCE and fat-saturated T2 images and clinical and biological data in predicting pCR to NAC.

Taking advantages of their lack of biological effects and artefacts, experiments based on breast phantom images, acquired using the three different devices of Institut Curie, were conducted to build an MRI correction pipeline. Analyses concluded that the N4 algorithm, originally designed to correct bias field gain in brain MRI, needed to be tuned specifically for the breast area. It was also established that both bias field correction and normalization of MR images were needed to correct variations of image intensities (intra and inter scanner) but that further harmonization of the radiomic features using the ComBat method was required to reduce the "scanner effect".

Results obtained on phantoms were then adapted to patient images and a specific pipeline

to carry out radiomic analyses was outlined. Once the images acquired, the training set should be processed according to the following 7 steps: image pre-processing, lesion segmentation, feature extraction, feature harmonization, feature selection including robustness assessment, model building with a limited number of features and finally model evaluation. For the test set, the first three steps should be repeated similarly. As the test set gathered patients from many different imaging centers, there were not enough patients per center to apply the conventional ComBat method to reduce the "scanner effect". An original harmonization approach consisting in assigning patient to one of the three training imaging devices using information from the healthy breast tissue was designed. Based on the final goal of predicting patients with pCR, results after harmonization were at least equal or better in 82% of the experiments and performances were improved in 73% of the cases.

Finally, a deep learning-based approach to segment breast tumors on T1-weighted DCE images (including first image after injection and subtracted image) was developed to alleviate radiologists' workload and improve feature reproducibility and model exportability. This approach, requiring the final input of a user, is patient-centric as it allows to choose for each patient the best fit amongst three models using either post-contrast images or an early or late fusion of post-contrast and subtraction T1-DCE images.

Thus, this work proposed a robust pipeline to carry out MRI-based radiomic analyses, including pre-processing techniques specifically tuned for breast MR images, a new patient-centric segmentation approach to improve feature reproducibility and an original harmonization strategy that handles small multicentric test sets where conventional harmonization methods fail. This pipeline or some steps such as the harmonization strategy, could be used in all kinds of radiomic analyses beyond the prediction of the response to NAC.

Using the defined pipeline, investigations were conducted to determine what information found in MR images could be relevant to predict pCR. The contribution of shape and margins of the lesions to the prediction model compared with tumor heterogeneity was assessed using an original multi-VOI approach. Experiments concluded that models built with features extracted from binarized images of tumor lesions that apprehend the outline of the tumor beyond what traditional shape parameters can convey or from a combination of features from several volumes of interest (the standard tumor segmentation, a bounding box on the binarized tumor images and a constant box placed inside the tumor) achieved better performances than models using the standard tumor segmentation alone (median Youden index on the test set: 0.44, IQR [0.43, 0.50] vs 0.16, [0.10, 0.18], $p < 0.05$). Using the AUC to measure the performances, the best experiment achieved an AUC of 0.80, [0.71, 0.89] on the training set and 0.72, [0.53, 0.91] on the test set. Performances were thus comparable to the performances of the models built using only clinical and biological data (test AUCs in range [0.66, 0.76]). Defining models specific to the different molecular subtype led to better performances (best AUC=0.79, [0.60, 0.98] on the test set), but further analyses (including training and testing) with larger datasets are required.

Testing models published in the literature on our dataset is challenging as on top of the "scanner effect", feature extraction and normalization parameters and weights of the models (when applicable) are rarely disclosed. Comparing performances of the models developed with results from the literature is also a tricky task. Differences in composition of the datasets including distribution of molecular subtypes may impact results. In this study, performances

were obtained on an independent multicentric test set, which remains rare in the literature. The test set was also highly unusual compared with the literature due to its extreme variability with images collected from more than 15 imaging centers though it accurately represents databases collected in clinical routine. Performances of the models developed were nevertheless in par with those reported in the literature.

Several leads could be followed in future work. First, to confirm the trends observed in our experiments, a larger dataset and especially a larger test set should be acquired. Then, as new patients are recruited into the study, the deep learning-based segmentation model should be included in the treatment pipeline to ease the segmentation step but also to let features discarded as not robust enough to segmentation be used in models. They could bring new and relevant information.

This thesis focused on a handcrafted radiomic approach. This choice was made due to the limited size of our dataset at the beginning of our study, that increased throughout the thesis to reach 136 patients. It was also interesting to study what kind of performances could be obtained with low computational models and delve into the issue of their exportability. However, with an increased dataset, deep learning-based approaches to predict pCR could also be investigated. Though for the moment they are limited in number compared with handcrafted approaches, there is a rise in the use of deep learning that affects all areas. Deep radiomic models built with biological and clinical data and features extracted from the final layer of a CNN could also be tested. Peng et al. [200] reported increased performances with deep learning-based models compared with handcrafted radiomic models (AUC 0.83 vs 0.78, $p < 0.001$). Deep learning methods nevertheless require a larger number of images and greater computational power.

Voxel-based analysis like performed in [277] could also be investigated to determine if there exist patterns in sub-regions of the tumor associated with the prediction of pCR.

To improve prediction, new sources of information could be integrated into the study such as DW images or ADC maps. Kinetic parameters were not the major focus of this thesis, but they have been used in many models with high performances reported. Texture analysis of kinetic parametric maps has even been explored and could thus be included in our models.

The database at Institut Curie also includes mid-treatment images, whose acquisitions were recently completed. These images open a new path for research to better interpret the radiomic signatures and test the reproducibility of the whole pipeline. Though not as groundbreaking as prediction before treatment, being able to predict the response after a few cycles of chemotherapy would still represent a major step forward for patient care. From a radiomic point of view, features extracted from mid-treatment images could feed a variety of models. They could be used alone, or in combination with features from pre-treatment images. Changes and variations of texture features, sometimes called "delta-radiomics", between the pre-treatment and mid-treatment phases could also be studied as they express changes in the biological properties of tissues. These changes often appear before any modification of the general shape of the tumors, that could still be observed at mid-treatment.

Going beyond the prediction of the response, radiomic features could be further analyzed to determine if they convey information about the risk of breast cancer recurrence. As, though patients achieving pCR have better outcomes, tumors can still reoccur.

Glossary

ACR American College of Radiology.

ADC Apparent Diffusion Coefficient.

AJCC American Joint Comittee on Cancer.

AUC Area Under the Curve.

BB Bounding Box.

bBB binary Bounding box.

BC Bias Corrected.

BES Breast Edema Score.

BH Benjamini and Hochberg multiple comparisons correction method.

BI-RADS Breast Imaging Reporting & Data System.

BMI Body Mass Index.

CB Constant Box.

CCC Concordance Correlation Coefficient.

CNN Convolutional Neural Network.

CV Coefficient of variation.

DCE Dynamic Contrast-Enhanced.

DCIS Ductal Carcinoma In Situ.

DSC Dice Similarity Coefficient.

DWI Diffusion-weighted imaging.

EFS Event-Free Survival.

ER Estrogen Receptor.

FDA Food and Drug Administration.

FISH Fluorescence In Situ Hybridization.

GLCM Gray Level Co-occurrence Matrix.

GLDM Gray Level Dependence Matrix.

GLRLM Gray Level Run Length Matrix.

GLSZM Gray Level Size Zone.

HD95 95th percentile of Hausdorff Distance.

HDI Human Development Index.

HER2 Human Epidermal growth factor Receptor 2.

HM Histogram Matching.

HTR High Temporal Resolution.

IBSI Image Biomarker Standardization Initiative.

ICC Intraclass Correlation Coefficient.

IHC Immunohistochemistry.

IQR Interquartile Range.

LCIS Lobular Carcinoma In Situ.

LOOCV Leave-One-Out Cross-Validation.

MIP Maximum Intensity Projection.

mRMR minimum Redundancy Maximum Relevance.

NAC Neoadjuvant chemotherapy.

NGTDM Neighboring Gray Tone Difference Matrix.

npCR non Pathological Complete Response.

OS Overall survival.

PCA Principal Component analysis.

pCR Pathological Complete Response.

PR Progesterone Receptor.

RCB Residual Cancer Burden.

RECIST Response Evaluation Criteria in Solid Tumors.

ReLU Rectified Linear Unit.

RF Random Forest.

RFE Recursive Feature Elimination.

ROC Receiver Operating Characteristic.

ROI Region Of Interest.

RQS Radiomic Quality Score.

SubT1 Subtraction image.

SVM Support Vector Machine.

T Tumor.

T1-DCE T1-weighted Dynamic Contrast-Enhanced.

T1c First post-contrast DCE-MR image.

TILs Tumor-Infiltrating Lymphocytes.

TN Triple Negative.

VOI Volume Of Interest.

WS White-Stripe normalization.

Bibliography

[1] 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. issn: 1542-4863. doi: 10.3322/CAAC.21660.

[2] Gnant, Harbeck, and Thomssen. “St. Gallen 2011: Summary of the Consensus Discussion”. In: *Breast Care* 6.2 (2011), p. 136. issn: 16613791. doi: 10.1159/000328054.

[3] Charfare, Limongelli, and Purushotham. “Neoadjuvant chemotherapy in breast cancer”. In: *British Journal of Surgery* 92.1 (2005), pp. 14–23. issn: 0007-1323. doi: 10.1002/BJS.4840.

[4] Asaoka et al. “Neoadjuvant Chemotherapy for Breast Cancer: Past, Present, and Future.” In: *Breast cancer : basic and clinical research* 14 (2020), p. 117822342 0980377. issn: 1178-2234. doi: 10.1177/1178223420980377.

[5] Haque et al. “Response rates and pathologic complete response by breast cancer molecular subtype following neoadjuvant chemotherapy”. In: *Breast Cancer Research and Treatment* 170.3 (2018), pp. 559–567. issn: 15737217. doi: 10.1007/S10549-018-4801-3/TABLES/3.

[6] Wang-Lopez et al. “Can pathologic complete response (pCR) be used as a surrogate marker of survival after neoadjuvant therapy for breast cancer?” In: *Critical Reviews in Oncology/Hematology* 95.1 (2015), pp. 88–104. issn: 1040-8428. doi: 10.1016/J.CRITREVONC.2015.02.011.

[7] Lambin et al. “Radiomics: extracting more information from medical images using advanced feature analysis”. In: *European journal of cancer* 48.4 (2012), pp. 441–446. issn: 1879-0852. doi: 10.1016/J.EJCA.2011.11.036.

[8] Gillies, Kinahan, and Hricak. “Radiomics: Images are more than pictures, they are data”. In: *Radiology* 278.2 (2016), pp. 563–577. issn: 15271315. doi: 10.1148/radiol.2015151169.

[9] Mann, Cho, and Moy. “Breast MRI: State of the art”. In: *Radiology* 292.3 (2019), pp. 520–536. issn: 15271315. doi: 10.1148/RADIOL.2019182947/SUPPL_FILE/RY182947SUPPF2.JPG.

[10] Reig et al. “Role of MRI to Assess Response to Neoadjuvant Therapy for Breast Cancer”. In: *Journal of Magnetic Resonance Imaging* 52.6 (2020). issn: 15222586. doi: 10.1002/jmri.27145.

[11] Orlhac et al. “A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies”. In: *Journal of Nuclear Medicine* 63.2 (2022), pp. 172–179. issn: 0161-5505. doi: 10.2967/JNUMED.121.262464.

[12] Hussain et al. “Machine learning classification of texture features of MRI breast tumor and peri-tumor of combined pre- and early treatment predicts pathologic complete response”. In: *BioMedical Engineering Online* 20.1 (2021), pp. 1–23. issn: 1475925X. doi: 10.1186/S12938-021-00899-Z/FIGURES/5.

[13] Eun et al. “Texture Analysis with 3.0-T MRI for Association of Response to Neoadjuvant Chemotherapy in Breast Cancer”. In: *Radiology* 294.1 (2020), pp. 31–41. issn: 0033-8419. doi: 10.1148/radiol.2019182718.

[14] Pesapane et al. “Radiomics of MRI for the Prediction of the Pathological Response to Neoadjuvant Chemotherapy in Breast Cancer Patients: A Single Referral Centre Analysis”. In: *Cancers* 13.17 (2021), p. 4271. issn: 20726694. doi: 10.3390/CANCERS13174271.

[15] Nemeth et al. “Multicontrast MRI-based radiomics for the prediction of pathological complete response to neoadjuvant chemotherapy in patients with early triple negative breast cancer”. In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 34.6 (2021), pp. 833–844. issn: 1352-8661. doi: 10.1007/S10334-021-00941-0.

[16] Caballo et al. “Four-Dimensional Machine Learning Radiomics for the Pretreatment Assessment of Breast Cancer Pathologic Complete Response to Neoadjuvant Chemotherapy in Dynamic Contrast- Enhanced MRI”. In: *Journal of magnetic resonance imaging* (2022), pp. 1–14. doi: 10.1002/jmri.28273.

[17] Spak et al. “BI-RADS® fifth edition: A summary of changes”. In: *Diagnostic and Interventional Imaging* 98.3 (2017), pp. 179–190. issn: 2211-5684. doi: 10.1016/J.DIII.2017.01.001.

[18] Granzier et al. “Mri-based radiomics analysis for the pretreatment prediction of pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients: A multicenter study”. In: *Cancers* 13.10 (2021). doi: 10.3390/CANCERS13102447.

[19] Saint Martin et al. “A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study”. In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 34.3 (2021), pp. 355–366. issn: 13528661. doi: 10.1007/s10334-020-00892-y.

[20] Tustison et al. “N4ITK: Improved N3 bias correction”. In: *IEEE Transactions on Medical Imaging* 29.6 (2010), pp. 1310–1320. issn: 02780062. doi: 10.1109/TMI.2010.2046908.

[21] Nyúl and Udupa. “On standardizing the MR image intensity scale”. In: *Magnetic Resonance in Medicine* 42.6 (1999), pp. 1072–1081. issn: 1522-2594. doi: 10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

[22] Shah et al. “Evaluating intensity normalization on MRIs of human brain with multiple sclerosis”. In: *Medical Image Analysis* 15.2 (2011), pp. 267–282. issn: 13618415. doi: 10.1016/j.media.2010.12.003.

[23] Johnson, Li, and Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8 (2007), pp. 118–127. doi: 10.1093/biostatistics/kxj037.

[24] Isaksson et al. “Effects of MRI image normalization techniques in prostate cancer radiomics”. In: *Physica Medica* 71 (2020), pp. 7–13. issn: 1724191X. doi: 10.1016/j.ejmp.2020.02.007.

[25] Saint Martin et al. “Decrypting the information captured by MRI-radiomic features in predicting the response to neoadjuvant chemotherapy in breast cancer”. In: *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2022), pp. 3227–3230. doi: 10.1109/EMBC48229.2022.9871844.

[26] 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* In press (2022). doi: 10.1007/s00330-022-09113-7.

[27] Fidler, Soerjomataram, and Bray. “A global view on cancer incidence and national levels of the human development index”. In: *International Journal of Cancer* 139.11 (2016), pp. 2436–2446. issn: 1097-0215. doi: 10.1002/IJC.30382.

[28] 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. issn: 2523-3548. doi: 10.1002/CAC2.12207.

[29] Bray et al. “Global cancer transitions according to the Human Development Index (2008–2030): a population-based study”. In: *The Lancet Oncology* 13.8 (2012), pp. 790–801. issn: 1470-2045. doi: 10.1016/S1470-2045(12)70211-5.

[30] Łukasiewicz et al. “Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review”. In: *Cancers* 13.17 (2021), p. 4287. issn: 20726694. doi: 10.3390/CANCERS13174287.

[31] Huang et al. “Global incidence and mortality of breast cancer: a trend analysis”. In: *Aging* 13.4 (2021), p. 5748. issn: 19454589. doi: 10.18632/AGING.202502.

[32] Kamińska et al. “Breast cancer risk factors”. In: *Menopause Review* 14.3 (2015), p. 196. issn: 22990038. doi: 10.5114/PM.2015.54346.

[33] Islami et al. “Breastfeeding and breast cancer risk by receptor status—a systematic review and meta-analysis”. In: *Annals of Oncology* 26.12 (2015), pp. 2398–2407. issn: 0923-7534. doi: 10.1093/ANNONC/MDV379.

[34] Seiler et al. “Obesity, Dietary Factors, Nutrition, and Breast Cancer Risk”. In: *Current Breast Cancer Reports* 10.1 (2018), pp. 14–27. issn: 1943-4596. doi: 10.1007/S12609-018-0264-0.

[35] Gaudet et al. “Active Smoking and Breast Cancer Risk: Original Cohort Data and Meta-Analysis”. In: *Journal of the National Cancer Institute* 105.8 (2013), pp. 515–525. issn: 0027-8874. doi: 10.1093/JNCI/DJT023.

[36] Makki. “Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance.” In: *Clinical medicine insights. Pathology* 8.1 (2015), pp. 23–31. issn: 1179-5557. doi: 10.4137/CPath.S31563.

[37] Nascimento and Otoni. “Histological and molecular classification of breast cancer: what do we know?” In: *Mastology* 30 (2020), pp. 1–8. issn: 25945394. doi: 10.29289/25945394202020200024.

[38] Kalli et al. “American joint committee on cancer’s staging system for breast cancer, eighth edition: What the radiologist needs to know”. In: *Radiographics* 38.7 (2018), pp. 1921–1933. issn: 15271323. doi: 10.1148/rg.2018180056.

[39] Plichta et al. “Anatomy and Breast Cancer Staging: Is It Still Relevant?” In: *Surgical Oncology Clinics of North America* 27.1 (2018), pp. 51–67. issn: 15585042. doi: 10.1016/j.soc.2017.07.010.

[40] Ljuslinder et al. “LRIG1 expression in colorectal cancer.” In: *Acta oncologica* 46.8 (2007), pp. 1118–11122. doi: 10.1080/02841860701426823.

[41] Elston and ELLIS. “pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up”. In: *Histopathology* 19.5 (1991), pp. 403–410. issn: 13652559. doi: 10.1111/J.1365-2559.1991.TB00229.X.

[42] Rakha et al. “Breast cancer prognostic classification in the molecular era: The role of histological grade”. In: *Breast Cancer Research* 12.4 (2010), pp. 1–12. issn: 14655411. doi: 10.1186/BCR2607/TABLES/2.

[43] Schwartz et al. “Histologic grade remains a prognostic factor for breast cancer regardless of the number of positive lymph nodes and tumor size: a study of 161 708 cases of breast cancer from the SEER Program”. In: *Archives of pathology& laboratory medicine* 138.8 (2014), pp. 1048–1052. issn: 1543-2165. doi: 10.5858/ARPA.2013-0435-OA.

[44] Atanda et al. “Audit of nottingham system grades assigned to breast cancer cases in a Teaching Hospital”. In: *Annals of Tropical Pathology* 8.2 (2017), p. 104. issn: 2251-0060. doi: 10.4103/ATP.ATP_9_17.

[45] O’Sullivan, Loprinzi, and Haddad. “Updates in the Evaluation and Management of Breast Cancer”. In: *Mayo Clinic Proceedings* 93.6 (2018), pp. 794–807. issn: 19425546. doi: 10.1016/j.mayocp.2018.03.025.

[46] Perou et al. “Molecular portraits of human breast tumours”. In: *Nature* 406.6797 (2000), pp. 747–752. issn: 1476-4687. doi: 10.1038/35021093.

[47] Sørlie et al. “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”. In: *Proceedings of the National Academy of Sciences of the United States of America* 98.19 (2001), pp. 10869–10874. issn: 00278424. doi: 10.1073/PNAS.191367098/SUPPL_FILE/INDEX.HTML.

[48] Blows et al. “Subtyping of Breast Cancer by Immunohistochemistry to Investigate a Relationship between Subtype and Short and Long Term Survival: A Collaborative Analysis of Data for 10,159 Cases from 12 Studies”. In: *PLOS Medicine* 7.5 (2010), e1000279. issn: 1549-1676. doi: 10.1371/JOURNAL.PMED.1000279.

[49] Fragomeni, Sciallis, and Jeruss. “Molecular subtypes and local-regional control of breast cancer”. In: *Surgical oncology clinics of North America* 27.1 (2018), p. 95. issn: 15585042. doi: 10.1016/J.SOC.2017.08.005.

[50] Pernas and Tolaney. “HER2-positive breast cancer: new therapeutic frontiers and overcoming resistance.” In: *Therapeutic advances in medical oncology* 11 (2019), p. 1758835919833519. issn: 1758-8340. doi: 10.1177/1758835919833519.

[51] Lehmann et al. “Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies”. In: *The Journal of Clinical Investigation* 121.7 (2011), pp. 2750–2767. issn: 0021-9738. doi: 10.1172/JCI45014.

[52] Yin et al. “Triple-negative breast cancer molecular subtyping and treatment progress”. In: *Breast Cancer Research* 22.1 (2020), pp. 1–13. issn: 1465542X. doi: 10.1186/S13058-020-01296-5/TABLES/3.

[53] Lombardi et al. “The Proper Ki-67 Cut-Off in Hormone Responsive Breast Cancer: A Monoinstitutional Analysis with Long-Term Follow-Up”. In: *Breast Cancer: Targets and Therapy* 13 (2021), pp. 213–217. issn: 11791314. doi: 10.2147/BCTT.S305440.

[54] Amor et al. “215P Identifying the best Ki67 cut-off for determining luminal breast cancer subtypes using immunohistochemical analysis and PAM50 genomic classification”. In: *Annals of Oncology* 31 (2020), S327. issn: 0923-7534. doi: 10.1016/J.ANNONC.2020.08.337.

[55] Setyawati et al. “The Association between Molecular Subtypes of Breast Cancer with Histological Grade and Lymph Node Metastases in Indonesian Woman”. In: *Asian Pacific journal of cancer prevention* 19.5 (2018), pp. 1263–1268. issn: 2476-762X. doi: 10.22034/APJCP.2018.19.5.1263.

[56] Sujarittanakarn et al. “The case to case comparison of hormone receptors and HER2 status between primary breast cancer and synchronous axillary lymph node metastasis”. In: *Asian Pacific Journal of Cancer Prevention* 21.6 (2020), pp. 1559–1565. issn: 2476762X. doi: 10.31557/APJCP.2020.21.6.1559.

[57] Mieog, Van Der Hage, and Van De Velde. “Neoadjuvant chemotherapy for operable breast cancer”. In: *British Journal of Surgery* 94.10 (2007), pp. 1189–1200. issn: 0007-1323. doi: 10.1002/BJS.5894.

[58] Fisher et al. “Effect of preoperative chemotherapy on the outcome of women with operable breast cancer”. In: *Journal of clinical oncology* 16.8 (1998), pp. 2672–2685. issn: 0732-183X. doi: 10.1200/JCO.1998.16.8.2672.

[59] Van Nes et al. “Preoperative chemotherapy is safe in early breast cancer, even after 10 years of follow-up; clinical and translational results from the EORTC trial 10902”. In: *Breast cancer research and treatment* 115.1 (2009), pp. 101–113. issn: 1573-7217. doi: 10.1007/S10549-008-0050-1.

[60] Spring et al. “Pathologic Complete Response after Neoadjuvant Chemotherapy and Impact on Breast Cancer Recurrence and Survival: A Comprehensive Meta-analysis”. In: *Clinical Cancer Research* 26.12 (2020), pp. 2838–2848. issn: 15573265. doi: 10.1158/1078-0432.CCR-19-3492/75986/AM/PATHOLOGICAL-COMPLETE-RESPONSE-AFTER-NEOADJUVANT.

[61] Sahoo and Lester. “Pathology of Breast Carcinomas After Neoadjuvant Chemotherapy: An Overview With Recommendations on Specimen Processing and Reporting”. In: *Archives of Pathology& Laboratory Medicine* 133.4 (2009), pp. 633–642. issn: 0003-9985. doi: 10.5858/133.4.633.

[62] Lee et al. “Comparison of Pathologic Response Evaluation Systems after Anthracycline with/without Taxane-Based Neoadjuvant Chemotherapy among Different Subtypes of Breast Cancers”. In: *PLoS ONE* 10.9 (2015). issn: 19326203. doi: 10.1371/JOURNAL.PONE.0137885.

[63] Cortazar et al. “Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis”. In: *The Lancet* 384.9938 (2014), pp. 164–172. issn: 0140-6736. doi: 10.1016/S0140-6736(13)62422-8.

[64] Cortazar and Geyer. “Pathological Complete Response in Neoadjuvant Treatment of Breast Cancer”. In: *Annals of Surgical Oncology* 22.5 (2015), pp. 1441–1446. issn: 15344681. doi: 10.1245/S10434-015-4404-8/TABLES/1.

[65] Von Minckwitz et al. “Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes”. In: *Journal of Clinical Oncology* 30.15 (2012), pp. 1796–1804. issn: 0732183X. doi: 10.1200/JCO.2011.38.8595.

[66] *Pathological Complete Response in Neoadjuvant Treatment of High-Risk Early-Stage Breast Cancer: Use as an Endpoint to Support Accelerated Approval | FDA*.

[67] Kim et al. “Molecular subtypes and tumor response to neoadjuvant chemotherapy in patients with locally advanced breast cancer”. In: *Oncology* 79.5-6 (2011), pp. 324–330. issn: 00302414. doi: 10.1159/000322192.

[68] Kim et al. “Potential Benefits of Neoadjuvant Chemotherapy in Clinically Node-Positive Luminal Subtype- Breast Cancer”. In: *Journal of Breast Cancer* 22.3 (2019), p. 412. issn: 20929900. doi: 10.4048/JBC.2019.22.E35.

[69] Petrelli et al. “The value of platinum agents as neoadjuvant chemotherapy in triple-negative breast cancers: A systematic review and meta-analysis”. In: *Breast Cancer Research and Treatment* 144.2 (2014), pp. 223–232. issn: 15737217. doi: 10.1007/S10549-014-2876-Z/FIGURES/4.

[70] Colleoni et al. “A nomogram based on the expression of Ki-67, steroid hormone receptors status and number of chemotherapy courses to predict pathological complete remission after preoperative chemotherapy for breast cancer”. In: *European Journal of Cancer* 46.12 (2010), pp. 2216–2224. issn: 0959-8049. doi: 10.1016/J.EJCA.2010.04.008.

[71] Keam et al. “Nomogram predicting clinical outcomes in breast cancer patients treated with neoadjuvant chemotherapy”. In: *Journal of Cancer Research and Clinical Oncology* 137.9 (2011), pp. 1301–1308. issn: 01715216. doi: 10.1007/S00432-011-0991-3/FIGURES/4.

[72] Rouzier et al. “Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer”. In: *Journal of Clinical Oncology* 23.33 (2005), pp. 8331–8339. issn: 0732183X. doi: 10.1200/JCO.2005.01.2898.

[73] Xu et al. “Predictors of Neoadjuvant Chemotherapy Response in Breast Cancer: A Review”. In: *OncoTargets and Therapy* 13 (2020), pp. 5887–5899. issn: 11786930. doi: 10.2147/OTT.S253056.

[74] Lusho et al. “Platelet-to-Lymphocyte Ratio Is Associated With Favorable Response to Neoadjuvant Chemotherapy in Triple Negative Breast Cancer: A Study on 120 Patients”. In: *Frontiers in Oncology* 11 (2021), p. 678315. issn: 2234943X. doi: 10.3389/FONC.2021.678315.

[75] Salgado et al. “Tumor-Infiltrating Lymphocytes and Associations With Pathological Complete Response and Event-Free Survival in HER2-Positive Early-Stage Breast Cancer Treated With Lapatinib and Trastuzumab: A Secondary Analysis of the NeoALTTO Trial”. In: *JAMA Oncology* 1.4 (2015), pp. 448–455. issn: 2374-2437. doi: 10.1001/JAMAONCOL.2015.0830.

[76] Liu et al. “Optimal threshold for stromal tumor-infiltrating lymphocytes: its predictive and prognostic value in HER2-positive breast cancer treated with trastuzumab-based neoadjuvant chemotherapy”. In: *Breast Cancer Research and Treatment* 154.2 (2015), pp. 239–249. issn: 15737217. doi: 10.1007/S10549-015-3617-7/FIGURES/6.

[77] Morrow, Waters, and Morris. “MRI for breast cancer screening, diagnosis, and treatment”. In: *The Lancet* 378 (9805 2011), pp. 1804–1811. issn: 0140-6736. doi: 10.1016/S0140-6736(11)61350-0.

[78] 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. issn: 0007-9235. doi: 10.3322/CANJCLIN.57.2.75.

[79] Arnaout et al. “Use of Preoperative Magnetic Resonance Imaging for Breast Cancer: A Canadian Population-Based Study”. In: *JAMA Oncology* 1.9 (2015), pp. 1238–1250. issn: 2374-2437. doi: 10.1001/JAMAONCOL.2015.3018.

[80] Houssami et al. “Accuracy and surgical impact of magnetic resonance imaging in breast cancer staging: Systematic review and meta-analysis in detection of multifocal and multicentric cancer”. In: *Journal of Clinical Oncology* 26.19 (2008), pp. 3248–3258. issn: 0732183X. doi: 10.1200/JCO.2007.15.2108.

[81] Houssami and Hayes. “Review of preoperative magnetic resonance imaging (MRI) in breast cancer: should MRI be performed on all women with newly diagnosed, early stage breast cancer?” In: *CA: a cancer journal for clinicians* 59.5 (2009), pp. 290–302. issn: 1542-4863. doi: 10.3322/CAAC.20028.

[82] Houssami et al. “An individual person data meta-analysis of preoperative magnetic resonance imaging and breast cancer recurrence”. In: *Journal of Clinical Oncology* 32.5 (2014), pp. 392–401. issn: 15277755. doi: 10.1200/JCO.2013.52.7515.

[83] Thompson and Wright. “The role of breast MRI in newly diagnosed breast cancer: An evidence-based review”. In: *The American Journal of Surgery* 221.3 (2021), pp. 525–528. issn: 0002-9610. doi: 10.1016/J.AMJSURG.2020.12.018.

[84] Chen et al. “Breast cancer: Evaluation of response to neoadjuvant chemotherapy with 3.0-T MR imaging”. In: *Radiology* 261.3 (2011), pp. 735–743. issn: 15271315. doi: 10.1148/RADIOL.11110814/-/DC1.

[85] Bouzón et al. “Diagnostic accuracy of MRI to evaluate tumour response and residual tumour size after neoadjuvant chemotherapy in breast cancer patients”. eng. In: *Radiology and oncology* 50.1 (2016), pp. 73–79. issn: 1318-2099. doi: 10.1515/raon-2016-0007.

[86] O’Flynn and DeSouza. “Functional magnetic resonance: Biomarkers of response in breast cancer”. In: *Breast Cancer Research* 13.1 (2011). issn: 14655411. doi: 10.1186/BCR2815.

[87] Gordon et al. “Dynamic contrast-enhanced magnetic resonance imaging: fundamentals and application to the evaluation of the peripheral perfusion”. In: *Cardiovascular Diagnosis and Therapy* 4.2 (2014), p. 147. issn: 2223-3652. doi: 10.3978/J.ISSN.2223-3652.2014.03.01.

[88] Cheon et al. “Invasive breast cancer: Prognostic value of peritumoral edema identified at preoperative MR imaging”. In: *Radiology* 287.1 (2018), pp. 68–75. issn: 15271315. doi: 10.1148/RADIOL.2017171157/ASSET/IMAGES/LARGE/RADIOL.2017171157.TBL4.JPEG.

[89] Uematsu, Kasami, and Watanabe. “Can T2-weighted 3-T breast MRI predict clinically occult inflammatory breast cancer before pathological examination? A single-center experience”. In: *Breast cancer* 21.1 (2014), pp. 115–121. issn: 1880-4233. doi: 10.1007/S12282-012-0425-3.

[90] Le Bihan et al. “Artifacts and pitfalls in diffusion MRI”. In: *Journal of Magnetic Resonance Imaging* 24.3 (2006), pp. 478–488. issn: 1522-2586. doi: 10.1002/JMRI.20683.

[91] Woodhams et al. “ADC mapping of benign and malignant breast tumors”. In: *Magnetic resonance in medical sciences* 4.1 (2005), pp. 35–42. issn: 1347-3182. doi: 10.2463/MRMS.4.35.

[92] Milon et al. “Abbreviated breast MRI combining FAST protocol and high temporal resolution (HTR) dynamic contrast enhanced (DCE) sequence”. In: *European Journal of Radiology* 117 (2019), pp. 199–208. issn: 0720-048X. doi: 10.1016/J.EJRAD.2019.06.022.

[93] Moschetta et al. “Abbreviated Combined MR Protocol: A New Faster Strategy for Characterizing Breast Lesions”. In: *Clinical Breast Cancer* 16.3 (2016), pp. 207–211. issn: 19380666. doi: 10.1016/j.clbc.2016.02.008.

[94] Harvey et al. “An Abbreviated Protocol for High-Risk Screening Breast MRI Saves Time and Resources”. In: *Journal of the American College of Radiology* 13.4 (2016), pp. 374–380. issn: 1546-1440. doi: 10.1016/J.JACR.2015.08.015.

[95] Mann et al. “A novel approach to contrast-enhanced breast magnetic resonance imaging for screening: High-resolution ultrafast dynamic imaging”. In: *Investigative Radiology* 49.9 (2014), pp. 579–585. issn: 15360210. doi: 10.1097/RLI.0000000000000057.

[96] Ramtohul et al. “Prospective Evaluation of Ultrafast Breast MRI for Predicting Pathologic Response after Neoadjuvant Therapies”. In: *Radiology* (2022). issn: 0033-8419. doi: 10.1148/radiol.220389.

[97] Gutierrez et al. “BI-RADS lesion characteristics predict likelihood of malignancy in breast MRI for masses but not for nonmasslike enhancement”. In: *American journal of roentgenology* 193.4 (2009), pp. 994–1000. issn: 1546-3141. doi: 10.2214/AJR.08.1983.

[98] Wu et al. “Prediction of molecular subtypes of breast cancer using BI-RADS features based on a ”white box” machine learning approach in a multi-modal imaging setting”. In: *European Journal of Radiology* 114 (2019), pp. 175–184. issn: 0720-048X. doi: 10.1016/J.EJRAD.2019.03.015.

[99] Moffa et al. “Can MRI Biomarkers Predict Triple-Negative Breast Cancer?” In: *Diagnostics* 10.12 (2020), p. 1090. issn: 2075-4418. doi: 10.3390/DIAGNOSTICS10121090.

[100] Cen et al. “BI-RADS 3–5 microcalcifications: prediction of lymph node metastasis of breast cancer”. In: *Oncotarget* 8.18 (2017), p. 30190. issn: 19492553. doi: 10.18632/ONCOTARGET.16318.

[101] Lee et al. “Patterns of malignant non-mass enhancement on 3-T breast MRI help predict invasiveness: using the BI-RADS lexicon fifth edition”. In: *Acta Radiologica* 59.11 (2018), pp. 1292–1299. issn: 16000455. doi: 10.1177/0284185118759139.

[102] Harada et al. “Evaluation of Breast Edema Findings at T2-weighted Breast MRI Is Useful for Diagnosing Occult Inflammatory Breast Cancer and Can Predict Prognosis after Neoadjuvant Chemotherapy”. In: *Radiology* 299.1 (2021), pp. 53–62. issn: 1527-1315. doi: 10.1148/RADIOL.2021202604.

[103] Malhaire et al. “Association of pretherapeutic BI-RADS and breast edema MRI descriptors with breast cancer response after neoadjuvant chemotherapy: a systematic study”. In: *European Radiology* In revision (2022).

[104] Uematsu, Kasami, and Yuen. “Neoadjuvant chemotherapy for breast cancer: Correlation between the baseline MR imaging findings and responses to therapy”. In: *European Radiology* 20.10 (2010), pp. 2315–2322. issn: 09387994. doi: 10.1007/S00330-010-1813-8/FIGURES/3.

[105] Bae et al. “Pretreatment MR Imaging Features of Triple-Negative Breast Cancer: Association with Response to Neoadjuvant Chemotherapy and Recurrence-Free Survival”. In: *Radiology* 281.2 (2016), pp. 392–400. issn: 1527-1315. doi: 10.1148/RADIOL.2016152331.

[106] Michoux et al. “Texture analysis on MR images helps predicting non-response to NAC in breast cancer”. In: *BMC Cancer* 15.1 (2015), pp. 1–13. issn: 14712407. doi: 10.1186/S12885-015-1563-8/TABLES/5.

[107] Tahmassebi et al. “Impact of Machine Learning with Multiparametric Magnetic Resonance Imaging of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients”. In: *Investigative radiology* 54.2 (2019), p. 110. issn: 15360210. doi: 10.1097/RLI.0000000000000518.

[108] Ibrahim et al. “Radiomics for precision medicine: current challenges,future prospects, and the proposal of a new framework”. In: *Methods* 188 (2021), pp. 20–29. issn: 10462023. doi: 10.1016/j.ymeth.2020.05.022.

[109] Duron et al. “Can we use radiomics in ultrasound imaging? Impact of preprocessing on feature repeatability”. In: *Diagnostic and Interventional Imaging* 102.11 (2021), pp. 659–667. issn: 2211-5684. doi: 10.1016/J.DIII.2021.10.004.

[110] Ponsiglione et al. “Cardiac CT and MRI radiomics: systematic review of the literature and radiomics quality score assessment”. In: *European radiology* 32.4 (2022), pp. 2629–2638. issn: 1432-1084. doi: 10.1007/S00330-021-08375-X.

[111] Ranjbar et al. “Brain MR radiomics to differentiate cognitive disorders”. In: *The Journal of neuropsychiatry and clinical neurosciences* 31.3 (2019), p. 210. issn: 15457222. doi: 10.1176/APPI.NEUROPSYCH.17120366.

[112] Bickelhaupt et al. “Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography.” In: *Journal of Magnetic Resonance Imaging* 46.2 (2017), pp. 604–616. issn: 10531807. doi: 10.1002/jmri.25606.

[113] Wu et al. “Identifying relations between imaging phenotypes and molecular subtypes of breast cancer: Model discovery and external validation”. In: *Journal of Magnetic Resonance Imaging* 46.4 (2017), pp. 1017–1027. issn: 15222586. doi: 10.1002/jmri.25661.

[114] Wu et al. “Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology”. In: *Frontiers in oncology* 6.MAR (2016). issn: 2234-943X. doi: 10.3389/FONC.2016.00071.

[115] Bae et al. “Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction”. In: *Radiology* 289.3 (2018), pp. 797–806. issn: 1527-1315. doi: 10.1148/RADIOL.2018180200.

[116] Wang et al. “Radiomics features on radiotherapy treatment planning CT can predict patient survival in locally advanced rectal cancer patients”. In: *Scientific Reports* 9.1 (2019). issn: 20452322. doi: 10.1038/S41598-019-51629-4.

[117] Park et al. “Radiomics signature on magnetic resonance imaging: Association with disease-free survival in patients with invasive breast cancer”. In: *Clinical Cancer Research* 24.19 (2018), pp. 4705–4714. issn: 15573265. doi: 10.1158/1078-0432.CCR-17-3783.

[118] Lohmann et al. “Radiomics in neuro-oncology: Basics, workflow, and applications”. In: *Methods* 188 (2021), pp. 112–121. issn: 1046-2023. doi: 10.1016/J.YMETH.2020.06.003.

[119] Bianconi et al. “PET/CT radiomics in lung cancer: An overview”. In: *Applied Sciences* 10.5 (2020), pp. 1–11. issn: 20763417. doi: 10.3390/app10051718.

[120] Ye, Wang, and Yu. “The Application of Radiomics in Breast MRI: A Review”. In: *Technology in cancer research& treatment* 19.44 (2020), pp. 1–16. issn: 15330338. doi: 10.1177/1533033820916191.

[121] Nasief et al. “A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer”. In: *Precision Oncology* 3.1 (2019), pp. 1–10. issn: 2397-768X. doi: 10.1038/s41698-019-0096-z.

[122] Sellami et al. “Predicting response to radiotherapy of head and neck squamous cell carcinoma using radiomics from cone-beam CT images”. In: *Acta oncologica* 61.1 (2022), pp. 73–80. issn: 1651-226X. doi: 10.1080/0284186X.2021.1983207.

[123] Wesdorp et al. “Advanced analytics and artificial intelligence in gastrointestinal cancer: a systematic review of radiomics predicting response to treatment”. In: *European Journal of Nuclear Medicine and Molecular Imaging* 48.6 (2021), pp. 1785–1794. issn: 16197089. doi: 10.1007/S00259-020-05142-W/FIGURES/2.

[124] Trebeschi et al. “Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers”. In: *Annals of Oncology* 30.6 (2019), pp. 998–1004. issn: 0923-7534. doi: 10.1093/ANNONC/MDZ108.

[125] García-Figueiras et al. “Assessing Immunotherapy with Functional and Molecular Imaging and Radiomics”. In: *Radiographics* 40.7 (2020), pp. 1987–2010. issn: 1527-1323. doi: 10.1148/RG.2020200070.

[126] Sahiner et al. “Deep learning in medical imaging and radiation therapy”. In: *Medical Physics* 46.1 (2019), e1–e36. issn: 2473-4209. doi: 10.1002/MP.13264.

[127] Zhang et al. “Deep Learning With Radiomics for Disease Diagnosis and Treatment: Challenges and Potential”. In: *Frontiers in Oncology* 12 (2022), p. 276. issn: 2234943X. doi: 10.3389/FONC.2022.773840/BIBTEX.

[128] Wang et al. “Deep learning combined with radiomics may optimize the prediction in differentiating high-grade lung adenocarcinomas in ground glass opacity lesions on CT scans”. In: *European Journal of Radiology* 129 (2020), p. 109150. issn: 0720-048X. doi: 10.1016/J.EJRAD.2020.109150.

[129] Kobayashi et al. “Observing deep radiomics for the classification of glioma grades”. In: *Scientific Reports* 11.1 (2021), pp. 1–13. issn: 2045-2322. doi: 10.1038/s41598-021-90555-2.

[130] Wei et al. “Machine learning for radiomics-based multimodality and multiparametric modeling”. In: *The quarterly journal of nuclear medicine and molecular imaging* 63.4 (2019), pp. 323–338. issn: 1827-1936. doi: 10.23736/S1824-4785.19.03213-8.

[131] Li et al. “18F-FDG PET/CT radiomic predictors of pathologic complete response (pCR) to neoadjuvant chemotherapy in breast cancer patients”. In: *European Journal of Nuclear Medicine and Molecular Imaging* 47 (2020), pp. 1116–1126. doi: 10.1007/s00259-020-04684-3.

[132] Antunovic et al. “PET/CT radiomics in breast cancer: promising tool for prediction of pathological response to neoadjuvant chemotherapy”. In: *European Journal of Nuclear Medicine and Molecular Imaging* 46.7 (2019), pp. 1468–1477. issn: 16197089. doi: 10.1007/S00259-019-04313-8/FIGURES/5.

[133] Umutlu et al. “Multiparametric18F-FDG PET/MRI-Based Radiomics for Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer”. In: *Cancers* 14.7 (2022). issn: 20726694. doi: 10.3390/cancers14071727.

[134] Liu et al. “Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: A multicenter study”. In: *Clinical Cancer Research* 25.12 (2019), pp. 3538–3547. issn: 15573265. doi: 10.1158/1078-0432.CCR-18-3190.

[135] Traverso et al. “Repeatability and Reproducibility of Radiomic Features: A Systematic Review”. In: *International Journal of Radiation Oncology Biology Physics* 102.4 (2018), pp. 1143–1158. issn: 1879355X. doi: 10.1016/j.ijrobp.2018.05.053.

[136] Waugh et al. “The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms”. In: *Medical Physics* 38.9 (2011), pp. 5058–5066. issn: 00942405. doi: 10.1118/1.3622605.

[137] Saha et al. “Effects of MRI scanner parameters on breast cancer radiomics”. In: *Expert Systems with Applications* 87 (2017), pp. 384–391. issn: 09574174. doi: 10.1016/j.eswa.2017.06.029.

[138] Rai et al. “Multicenter evaluation of MRI-based radiomic features: A phantom study”. In: *Medical Physics* 47.7.0 (2020), pp. 3054–3063. issn: 00942405. doi: 10.1002/mp.14173.

[139] Ford et al. “Quantitative Radiomics: Impact of Pulse Sequence Parameter Selection on MRI-Based Textural Features of the Brain.” eng. In: *Contrast media& molecular imaging* 2018 (2018), p. 1729071. issn: 1555-4317 (Electronic). doi: 10.1155/2018/1729071.

[140] Chirra et al. *Empirical Evaluation of Cross-site Reproducibility and Discriminability of Radiomic Features for Characterizing Tumor Appearance on Prostate MRI*. Thesis realized in Case Western Reserve University, 2018.

[141] Song, Zheng, and He. “A review of Methods for Bias Correction in Medical Images”. In: *Biomedical Engineering Review* 1.1 (2017), p. 1. issn: 23759143. doi: 10.18103/bme.v3i1.1550.

[142] Uros, Franjo, and Bostjan. “A review of methods for correction of intensity inhomogeneity in MRI”. In: *IEEE Transactions on Medical Imaging* 26.3 (2007), pp. 405–421.

[143] 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. issn: 15271315. doi: 10.1148/radiol.2020191145.

[144] Braman et al. “Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI”. In: *Breast Cancer Research* 19.1 (2017). issn: 1465542X. doi: 10.1186/s13058-017-0846-1.

[145] Fan et al. “Radiomic analysis of DCE-MRI for prediction of response to neoadjuvant chemotherapy in breast cancer patients”. In: *European Journal of Radiology* 94 (2017), pp. 140–147. issn: 18727727. doi: 10.1016/j.ejrad.2017.06.019.

[146] Zhou et al. “Predicting the response to neoadjuvant chemotherapy for breast cancer: Wavelet transforming radiomics in MRI”. In: *BMC Cancer* 20.1 (2020), p. 100. issn: 14712407. doi: 10.1186/s12885-020-6523-2.

[147] Saha et al. “Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics”. In: *Medical physics* 43.8 (2016), pp. 4558–4564. issn: 2473-4209. doi: 10.1118/1.4955435.

[148] Saha, Harowicz, and Mazurowski. “Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors”. In: *Medical physics* 45.7 (2018), pp. 3076–3085. issn: 2473-4209. doi: 10.1002/MP.12925.

[149] Teruel et al. “Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer”. In: *NMR in Biomedicine* 27.8 (2014), pp. 887–896. issn: 09523480. doi: 10.1002/nbm.3132.

[150] Cain et al. “Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set”. In: *Breast Cancer Research and Treatment* 173.2 (2019), pp. 455–463. issn: 15737217. doi: 10.1007/s10549-018-4990-9.

[151] Kumar et al. “Radiomics: The process and the challenges”. In: *Magnetic Resonance Imaging* 30.9 (2012), pp. 1234–1248. issn: 0730725X. doi: 10.1016/j.mri.2012.06.010.

[152] Griethuysen et al. “Computational Radiomics System to Decode the Radiographic Phenotype”. In: *Cancer research* 77.21 (2017), e104–e107. issn: 1538-7445. doi: 10.1158/0008-5472.CAN-17-0339.

[153] Shrestha. “Detecting Multicollinearity in Regression Analysis”. In: *American Journal of Applied Mathematics and Statistics* 8.2 (2020), pp. 39–42. issn: 2328-7306. doi: 10.12691/ajams-8-2-1.

[154] Solorio-Fernández, Carrasco-Ochoa, and Martínez-Trinidad. “A review of unsupervised feature selection methods”. In: *Artificial Intelligence* 53.2 (2019), pp. 907–948. issn: 1573-7462. doi: 10.1007/S10462-019-09682-Y.

[155] Jović, Brkić, and Bogunović. “A review of feature selection methods with applications”. In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings* (2015), pp. 1200–1205. doi: 10.1109/MIPRO.2015.7160458.

[156] Radovic et al. “Minimum redundancy maximum relevance feature selection approach for temporal gene expression data”. In: *BMC Bioinformatics* 18.1 (2017), pp. 1–14. issn: 14712105. doi: 10.1186/S12859-016-1423-9/FIGURES/6.

[157] Chandrashekar and Sahin. “A survey on feature selection methods”. In: *Computers & Electrical Engineering* 40.1 (2014), pp. 16–28. issn: 0045-7906. doi: 10.1016/J.COMPELECENG.2013.11.024.

[158] Kursa and Rudnicki. “Feature Selection with the Boruta Package”. In: *Journal of Statistical Software* 36.11 (2010), pp. 1–13. issn: 1548-7660. doi: 10.18637/JSS.V036.I11.

[159] Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. issn: 1573-0565. doi: 10.1023/A:1010933404324.

[160] Tibshirani. “Regression Shrinkage and Selection Via the Lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1996), pp. 267–288. issn: 2517-6161. doi: 10.1111/J.2517-6161.1996.TB02080.X.

[161] Ogutu, Schulz-Streeck, and Piepho. “Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions”. In: *BMC Proceedings* 6 (SUPPL. 2 2012), pp. 1–6. issn: 17536561. doi: 10.1186/1753-6561-6-S2-S10/TABLES/1.

[162] Parmar et al. “Machine Learning methods for Quantitative Radiomic Biomarkers”. In: *Scientific Reports* 5.1 (2015), pp. 1–11. issn: 2045-2322. doi: 10.1038/srep13087.

[163] Bergstra and Bengio. “Random Search for Hyper-Parameter Optimization”. In: *Journal of Machine Learning Research* 13.null (2012), pp. 281–305. issn: 1532-4435.

[164] Yang and Shami. “On hyperparameter optimization of machine learning algorithms: Theory and practice”. In: *Neurocomputing* 415 (2020), pp. 295–316. issn: 0925-2312. doi: 10.1016/J.NEUCOM.2020.07.061.

[165] Hossin and Sulaiman. "A Review on Evaluation Metrics for Data Classification Evaluations". In: *International Journal of Data Mining & Knowledge Management Process* 5.2 (2015), pp. 01–11. issn: 2231007X. doi: 10.5121/ijdkp.2015.5201.

[166] Grandini, Bagli, and Visani. "Metrics for Multi-Class Classification: an Overview". In: *arXiv* (2020), pp. 1–17. eprint: 2008.05756.

[167] Golden et al. "Dynamic contrast-enhanced MRI-based biomarkers of therapeutic response in triple-negative breast cancer". In: *Journal of the American Medical Informatics Association* 20.6 (2013), pp. 1059–1066. issn: 10675027. doi: 10.1136/amiajnl-2012-001460.

[168] Banerjee et al. "Assessing treatment response in triple-negative breast cancer from quantitative image analysis in perfusion magnetic resonance imaging". In: *Journal of Medical Imaging* 5.01 (2017), p. 1. issn: 2329-4302. doi: 10.1117/1.jmi.5.1.011008.

[169] Thibault et al. "DCE-MRI Texture Features for Early Prediction of Breast Cancer Therapy Response". In: *Tomography* 3.1 (2017), pp. 23–32. issn: 23791381. doi: 10.18383/j.tom.2016.00241.

[170] Wu et al. "Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy". In: *Journal of Magnetic Resonance Imaging* 44.5 (2016), pp. 1107–1115. issn: 15222586. doi: 10.1002/jmri.25279.

[171] Lee, Park, and Ko. "Radiomics in Breast Imaging from Techniques to Clinical Applications: A Review". In: *Korean Journal of Radiology* 21.7 (2020), p. 779. issn: 12296929. doi: 10.3348/KJR.2019.0855.

[172] Nie et al. "Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI". In: *Academic radiology* 15.12 (2008), pp. 1513–1525. issn: 1878-4046. doi: 10.1016/J.ACRA.2008.06.005.

[173] Wang et al. "Computer-aided diagnosis of breast DCE-MRI using pharmacokinetic model and 3-D morphology analysis". In: *Magnetic resonance imaging* 32.3 (2014), pp. 197–205. issn: 1873-5894. doi: 10.1016/J.MRI.2013.12.002.

[174] Whitney et al. "Harmonization of radiomic features of breast lesions across international DCE-MRI datasets". In: *Journal of Medical Imaging* 7.01 (2020), p. 012707. issn: 2329-4310. doi: 10.1117/1.jmi.7.1.012707.

[175] Chan et al. "Eigentumors for prediction of treatment failure in patients with early-stage breast cancer using dynamic contrast-enhanced MRI: a feasibility study". In: *Physics in medicine and biology* 62.16 (2017), pp. 6467–6485. issn: 1361-6560. doi: 10.1088/1361-6560/AA7DC5.

[176] Moghadas-Dastjerdi et al. "A priori prediction of tumour response to neoadjuvant chemotherapy in breast cancer patients using quantitative CT and machine learning". In: *Scientific Reports* 10.1 (2020), pp. 1–11. issn: 2045-2322. doi: 10.1038/s41598-020-67823-8.

[177] Qi et al. “Multi-center evaluation of artificial intelligent imaging and clinical models for predicting neoadjuvant chemotherapy response in breast cancer”. In: *Breast Cancer Research and Treatment* 193.1 (2022), pp. 121–138. issn: 1573-7217. doi: 10.1007/S10549-022-06521-7.

[178] Humbert et al. “HER2-positive breast cancer: 18F-FDG PET for early prediction of response to trastuzumab plus taxane-based neoadjuvant chemotherapy”. In: *European Journal of Nuclear Medicine and Molecular Imaging* 41.8 (2014), pp. 1525–1533. issn: 16197089. doi: 10.1007/S00259-014-2739-1/FIGURES/3.

[179] Hatt et al. “Comparison Between 18F-FDG PET Image–Derived Indices for Early Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer”. In: *Journal of Nuclear Medicine* 54.3 (2013), pp. 341–349. issn: 0161-5505. doi: 10.2967/JNUMED.112.108837.

[180] DiCenzo et al. “Quantitative ultrasound radiomics in predicting response to neoadjuvant chemotherapy in patients with locally advanced breast cancer: Results from multi-institutional study”. In: *Cancer Medicine* 9.16 (2020), pp. 5798–5806. issn: 20457634. doi: 10.1002/cam4.3255.

[181] Cremoux et al. “18FDG-PET/CT and molecular markers to predict response to neoadjuvant chemotherapy and outcome in HER2-negative advanced luminal breast cancers patients”. In: *Oncotarget* 9.23 (2018), p. 16343. doi: 10.18632/ONCOTARGET.24674.

[182] Giannini et al. “A computer-aided diagnosis (CAD) scheme for pretreatment prediction of pathological response to neoadjuvant therapy using dynamic contrast-enhanced MRI texture features”. In: *British Journal of Radiology* 90.1077 (2017). issn: 00071285. doi: 10.1259/BJR.20170269/ASSET/IMAGES/LARGE/BJR.20170269.G003.JPEG.

[183] Henderson et al. “Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer”. In: *European Radiology* 27.11 (2017), pp. 4602–4611. issn: 14321084. doi: 10.1007/s00330-017-4850-8.

[184] Chamming’s et al. “Features from Computerized Texture Analysis of Breast Cancers at Pretreatment MR Imaging Are Associated with Response to Neoadjuvant Chemotherapy”. In: *Radiology* 286.2 (2018), pp. 412–420. issn: 0033-8419. doi: 10.1148/radiol.2017170143.

[185] Machireddy et al. “Early Prediction of Breast Cancer Therapy Response using Multiresolution Fractal Analysis of DCE-MRI Parametric Maps”. In: *Tomography* 5.1 (2019), pp. 90–98. issn: 2379139X. doi: 10.18383/j.tom.2018.00046.

[186] Drukker et al. “Breast MRI radiomics for the pretreatment prediction of response to neoadjuvant chemotherapy in node-positive breast cancer patients”. In: *Journal of medical imaging* 6.3 (2019), p. 1. issn: 2329-4302. doi: 10.1117/1.JMI.6.3.034502.

[187] Braman et al. “Association of Peritumoral Radiomics With Tumor Biology and Pathologic Response to Preoperative Targeted Therapy for HER2 (ERBB2)-Positive Breast Cancer”. In: *JAMA network open* 2.4 (2019), e192561. issn: 25743805. doi: 10.1001/jamanetworkopen.2019.2561.

[188] Bitencourt et al. “MRI-based machine learning radiomics can predict HER2 expression level and pathologic response after neoadjuvant therapy in HER2 overexpressing breast cancer”. In: *EBioMedicine* 61 (2020), p. 103042. issn: 23523964. doi: 10.1016/j.ebiom.2020.103042.

[189] Chen et al. “Machine Learning-Based Radiomics Nomogram Using Magnetic Resonance Images for Prediction of Neoadjuvant Chemotherapy Efficacy in Breast Cancer Patients”. In: *Frontiers in Oncology* 10 (2020), p. 1410. issn: 2234943X. doi: 10.3389/fonc.2020.01410.

[190] Chen et al. “Combining Dynamic Contrast-Enhanced Magnetic Resonance Imaging and Apparent Diffusion Coefficient Maps for a Radiomics Nomogram to Predict Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer Patients”. In: *Journal of Computer Assisted Tomography* 44.2 (2020), pp. 275–283. issn: 0363-8715. doi: 10.1097/RCT.0000000000000978.

[191] Xiong et al. “Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy”. In: *Clinical and Translational Oncology* 22.1 (2019), pp. 50–59. issn: 1699-3055. doi: 10.1007/S12094-019-02109-8.

[192] Fusco et al. “Textural radiomic features and time-intensity curve data analysis by dynamic contrast-enhanced MRI for early prediction of breast cancer therapy response: preliminary data”. In: *European Radiology Experimental* 4.1 (2020). issn: 25099280. doi: 10.1186/s41747-019-0141-2.

[193] Bian et al. “Radiomic signatures derived from multiparametric MRI for the pretreatment prediction of response to neoadjuvant chemotherapy in breast cancer”. In: *The British journal of radiology* 93.1115 (2020), p. 20200287. issn: 1748880X. doi: 10.1259/bjr.20200287.

[194] Sutton et al. “A machine learning model that classifies breast cancer pathologic complete response on MRI post-neoadjuvant chemotherapy”. In: *Breast Cancer Research* 22.1 (2020), pp. 1–11. issn: 1465542X. doi: 10.1186/S13058-020-01291-W/FIGURES/4.

[195] 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. 1289. issn: 2234943X. doi: 10.3389/FONC.2021.630780/BIBTEX.

[196] “MRI texture features from tumor core and margin in the prediction of response to neoadjuvant chemotherapy in patients with locally advanced breast cancer”. In: *Oncotarget* 12.14 (2021), p. 1354. issn: 19492553. doi: 10.18632/ONCOTARGET.28002.

[197] Yoshida et al. “Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using radiomics of pretreatment dynamic contrast-enhanced MRI”. In: *Magnetic Resonance Imaging* 92 (2022), pp. 19–25. issn: 0730-725X. doi: 10.1016/J.MRI.2022.05.018.

[198] Jimenez et al. “A model combining pretreatment MRI radiomic features and tumor-infiltrating lymphocytes to predict response to neoadjuvant systemic therapy in triple-negative breast cancer”. In: *European Journal of Radiology* 149 (2022), p. 110220. issn: 0720-048X. doi: 10.1016/J.EJRAD.2022.110220.

[199] Li et al. “A Noninvasive Tool Based on Magnetic Resonance Imaging Radiomics for the Preoperative Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer”. In: *Annals of Surgical Oncology 2022* (2022), pp. 1–9. issn: 1534-4681. doi: 10.1245/S10434-022-12034-W.

[200] Peng et al. “Pretreatment DCE-MRI-Based Deep Learning Outperforms Radiomics Analysis in Predicting Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer.” In: *Frontiers in Oncology* 12 (2022), pp. 846775–846775. issn: 2234-943X. doi: 10.3389/FONC.2022.846775.

[201] Huynh, Antropova, and Giger. “Comparison of breast DCE-MRI contrast time points for predicting response to neoadjuvant chemotherapy using deep convolutional neural network features with transfer learning”. In: *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* 10134 (2017), 101340U. doi: 10.1117/12.2255316.

[202] Ha et al. “Prior to Initiation of Chemotherapy, Can We Predict Breast Tumor Response? Deep Learning Convolutional Neural Networks Approach Using a Breast MRI Tumor Dataset”. In: *Journal of Digital Imaging* 32.5 (2018), pp. 693–701. issn: 1618-727X. doi: 10.1007/S10278-018-0144-1.

[203] Ravichandran et al. “A deep learning classifier for prediction of pathological complete response to neoadjuvant chemotherapy from baseline breast DCE-MRI”. In: *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* 10575 (2018). Ed. by Petrick and Mori, p. 105750C. doi: 10.1117/12.2294056.

[204] Adoui, Drisis, and Benjelloun. “Predict Breast Tumor Response to Chemotherapy Using a 3D Deep Learning Architecture Applied to DCE-MRI Data”. In: *Lecture Notes in Computer Science* 11466 LNBI (2019), pp. 33–40. doi: 10.1007/978-3-030-17935-9_4.

[205] Liu et al. “A novel CNN algorithm for pathological complete response prediction using an I-SPY TRIAL breast MRI database”. In: *Magnetic resonance imaging* 73 (2020), pp. 148–151. issn: 1873-5894. doi: 10.1016/J.MRI.2020.08.021.

[206] Braman et al. “Deep learning-based prediction of response to HER2-targeted neoadjuvant chemotherapy from pre-treatment dynamic breast MRI: A multi-institutional validation study”. In: *ArXiv* (2020). doi: 10.48550/arxiv.2001.08570. arXiv: 2001.08570.

[207] Qu et al. “Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using a deep learning (DL) method”. In: *Thoracic Cancer* 11.3 (2020), pp. 651–658. issn: 1759-7714. doi: 10.1111/1759-7714.13309.

[208] Duanmu et al. “Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer Using Deep Learning with Integrative Imaging, Molecular and Demographic Data”. In: *Lecture Notes in Computer Science* 12262 (2020), pp. 242–252. issn: 16113349. doi: 10.1007/978-3-030-59713-9_24/FIGURES/3.

[209] Joo et al. “Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer”. In: *Scientific Reports* 11.1 (2021), pp. 1–8. issn: 2045-2322. doi: 10.1038/s41598-021-98408-8.

[210] Massafra et al. “Robustness Evaluation of a Deep Learning Model on Sagittal and Axial Breast DCE-MRIs to Predict Pathological Complete Response to Neoadjuvant Chemotherapy”. In: *Journal of Personalized Medicine* 12.6 (2022), p. 953. issn: 2075-4426. doi: 10.3390/JPM12060953.

[211] Granzier et al. “MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability”. In: *Scientific Reports* 10.1 (2020). issn: 20452322. doi: 10.1038/s41598-020-70940-z.

[212] Mallat. “Understanding deep convolutional networks”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016). issn: 1364503X. doi: 10.1098/RSTA.2015.0203. arXiv: 1601.04920.

[213] Lambin et al. “Radiomics: the bridge between medical imaging and personalized medicine”. In: *Nature Reviews Clinical Oncology 2017 14:12* 14.12 (2017), pp. 749–762. issn: 1759-4782. doi: 10.1038/nrclinonc.2017.141.

[214] Hylton et al. “Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy–results from ACRIN 6657/I-SPY TRIAL”. In: *Radiology* 263.3 (2012), pp. 663–672. issn: 1527-1315. doi: 10.1148/RADIOL.12110748.

[215] Jafari and Ansari-Pour. “Why, When and How to Adjust Your P Values?” In: *Cell Journal* 20.4 (2019), p. 604. issn: 22285814. doi: 10.22074/CELLJ.2019.5992.

[216] Wang et al. “Prognostic value of residual cancer burden and Miller-Payne system after neoadjuvant chemotherapy for breast cancer”. In: *Gland Surgery* 10.12 (2021), p. 3211. issn: 22278575. doi: 10.21037/GS-21-608.

[217] DeLong, DeLong, and Clarke-Pearson. “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.” eng. In: *Biometrics* 44.3 (1988), pp. 837–845. issn: 0006-341X (Print).

[218] Ding et al. “Breast density quantification using magnetic resonance imaging (MRI) with bias field correction: A postmortem study”. In: *Medical Physics* 40.12 (2013), p. 122305. issn: 2473-4209. doi: 10.1118/1.4831967.

[219] Vovk, Pernus, and Likar. “A Review of Methods for Correction of Intensity Inhomogeneity in MRI”. In: *IEEE transactions on medical imaging* 26 (2007), pp. 405–421. doi: 10.1109/TMI.2006.891486.

[220] Wachowicz. "Evaluation of active and passive shimming in magnetic resonance imaging". In: *Research and Reports in Nuclear Medicine* 4 (2014), pp. 1–12. doi: 10.2147/RRNM.S46526.

[221] Li et al. "MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, oncotype DX, and PAM50 gene assays". In: *Radiology* 281.2 (2016), pp. 382–391. issn: 15271315. doi: 10.1148/radiol.2016152110.

[222] Granzier et al. "Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review". In: *European Journal of Radiology* 121 (2019), p. 108736. issn: 18727727. doi: 10.1016/j.ejrad.2019.108736.

[223] Buch et al. "Quantitative variations in texture analysis features dependent on MRI scanning parameters: A phantom model". In: *Journal of Applied Clinical Medical Physics* 19.6 (2018), pp. 253–264. issn: 15269914. doi: 10.1002/acm2.12482.

[224] Frackiewicz et al. "The evaluation of correction algorithms of intensity nonuniformity in breast MRI images: a phantom study". In: *Tenth International Conference on Machine Vision (ICMV 2017)* 10696 (2018), p. 15. doi: 10.1117/12.2309464.

[225] Lin et al. "A new bias field correction method combining N3 and FCM for improved segmentation of breast density on MRI". In: *Medical Physics* 38.1 (2010), pp. 5–14. issn: 00942405. doi: 10.1118/1.3519869.

[226] Shinohara et al. "Statistical normalization techniques for magnetic resonance imaging". In: *NeuroImage: Clinical* 6 (2014), pp. 9–19. issn: 22131582. doi: 10.1016/j.nicl.2014.08.008.

[227] Fortin et al. "Removing inter-subject technical variability in magnetic resonance imaging studies". In: *NeuroImage* 132 (2016), pp. 198–212. issn: 10959572. doi: 10.1016/j.neuroimage.2016.02.036.

[228] Goya-Outi et al. "Computation of reliable textural indices from multimodal brain MRI: Suggestions based on a study of patients with diffuse intrinsic pontine glioma". In: *Physics in Medicine and Biology* 63.10 (2018), p. 105003. issn: 13616560. doi: 10.1088/1361-6560/aabd21.

[229] Lacroix et al. "Correction for Magnetic Field Inhomogeneities and Normalization of Voxel Values Are Needed to Better Reveal the Potential of MR Radiomic Features in Lung Cancer". In: *Frontiers in Oncology* 10 (2020), p. 43. issn: 2234-943X. doi: 10.3389/fonc.2020.00043.

[230] Um et al. "Impact of image preprocessing on the scanner dependence of multiparametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets." eng. In: *Physics in medicine and biology* 64.16 (2019), p. 165011. issn: 1361-6560 (Electronic). doi: 10.1088/1361-6560/ab2f44.

[231] Moradmand, Aghamiri, and Ghaderi. "Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma". In: *Journal of Applied Clinical Medical Physics* 21.1 (2020), pp. 179–190. issn: 1526-9914. doi: 10.1002/acm2.12795.

[232] Shiradkar et al. “Radiomic features from pretreatment biparametric MRI predict prostate cancer biochemical recurrence: Preliminary findings”. In: *Journal of Magnetic Resonance Imaging* 48.6 (2018), pp. 1626–1636. issn: 10531807. doi: 10.1002/jmri.26178.

[233] Orlhac et al. “A postreconstruction harmonization method for multicenter radiomic studies in PET”. In: *Journal of Nuclear Medicine* 59.8 (2018), pp. 1321–1328. issn: 2159662X. doi: 10.2967/jnumed.117.199935.

[234] Orlhac et al. “Validation of a method to compensate multicenter effects affecting CT radiomics”. In: *Radiology* 291.1 (2019), pp. 53–59. issn: 15271315. doi: 10.1148/radiol.2019182023.

[235] Nioche et al. “Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity”. In: *Cancer Research* 78.16 (2018), pp. 4786–4789. issn: 15387445. doi: 10.1158/0008-5472.CAN-18-0125.

[236] Madabhushi and Udupa. “Interplay between intensity standardization and inhomogeneity correction in MR image processing”. In: *IEEE Transactions on Medical Imaging* 24.5 (2005), pp. 561–576. issn: 02780062. doi: 10.1109/TMI.2004.843256.

[237] Reinhold et al. “Evaluating the Impact of Intensity Normalization on MR Image Synthesis”. In: *arXiv* (2018), p. 126. eprint: 1812.04652.

[238] Fortin et al. “Harmonization of multi-site diffusion tensor imaging data”. In: *NeuroImage* 161 (2017), p. 116541. issn: 149-170. doi: 10.1016/j.neuroimage.2017.08.047.

[239] Fortin et al. “Harmonization of cortical thickness measurements across scanners and sites”. In: *NeuroImage* 167 (2018), pp. 104–120. issn: 10959572. doi: 10.1016/j.neuroimage.2017.11.024.

[240] Chatterjee et al. “Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization”. In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 3.2 (2019), pp. 210–215. issn: 2469-7311. doi: 10.1109/trpms.2019.2893860.

[241] Castaldo et al. “The impact of normalization approaches to automatically detect radiogenomic phenotypes characterizing breast cancer receptors status”. In: *Cancers* 12.2 (2020). issn: 20726694. doi: 10.3390/cancers12020518.

[242] Bianchini et al. “PETER PHAN: An MRI phantom for the optimisation of radiomic studies of the female pelvis”. In: *Physica Medica* 71 (2020), pp. 71–81. issn: 1724191X. doi: 10.1016/j.ejmp.2020.02.003.

[243] Valladares, Beyer, and Rausch. “Physical imaging phantoms for simulation of tumor heterogeneity in PET, CT, and MRI: An overview of existing designs”. In: *Medical Physics* 47.4 (2020), pp. 2023–2037. issn: 00942405. doi: 10.1002/mp.14045.

[244] Li et al. “Impact of preprocessing and harmonization methods on the removal of scanner effects in brain mri radiomic features”. In: *Cancers* 13.12 (2021). issn: 20726694. doi: 10.3390/cancers13123000.

[245] Fatania et al. “Intensity standardization of MRI prior to radiomic feature extraction for artificial intelligence research in glioma—a systematic review”. In: *European Radiology* (2022), pp. 1–12. issn: 14321084. doi: 10.1007/S00330-022-08807-2/TABLES/3.

[246] Destito et al. “The effect of MRI signal intensity normalization for radiomics analysis on PCNSL patients”. In: *Proceedings of ESMRMB* (2021).

[247] Wrobel et al. “Intensity warping for multisite MRI harmonization”. In: *NeuroImage* 223 (2020). issn: 10959572. doi: 10.1016/j.neuroimage.2020.117242.

[248] Houssami et al. “Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy”. In: *European Journal of Cancer* 48.18 (2012), pp. 3342–3354. issn: 0959-8049. doi: 10.1016/J.EJCA.2012.05.023.

[249] Liang, Yu, and Gao. “Machine learning with magnetic resonance imaging for prediction of response to neoadjuvant chemotherapy in breast cancer: a systematic review and meta-analysis”. In: *European Journal of Radiology* 150 (2022), p. 110247. issn: 0720048X. doi: 10.1016/j.ejrad.2022.110247.

[250] Wallis and Buvat. “Clever Hans effect found in a widely used brain tumour MRI dataset”. In: *Medical Image Analysis* 77 (2022), p. 102368. issn: 1361-8415. doi: 10.1016/J.MEDIA.2022.102368.

[251] Orlhac et al. “How can we combat multicenter variability in MR radiomics? Validation of a correction procedure”. In: *European Radiology* (2020), pp. 1–9. issn: 0938-7994. doi: 10.1007/s00330-020-07284-9.

[252] Nadrljanski and Milosevic. “Tumor texture parameters of invasive ductal breast carcinoma in neoadjuvant chemotherapy: early identification of non-responders on breast MRI”. In: *Clinical Imaging* 65 (2020), pp. 119–123. issn: 18734499. doi: 10.1016/j.clinimag.2020.04.016.

[253] Da-ano et al. “Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies”. In: *Scientific Reports* 10.1 (2020), p. 10248. issn: 2045-2322. doi: 10.1038/s41598-020-66110-w.

[254] Mann et al. “Breast MRI: guidelines from the European Society of Breast Imaging”. In: *European Radiology* 18.7 (2008), p. 1307. issn: 09387994. doi: 10.1007/S00330-008-0863-7.

[255] Hosny et al. “Artificial intelligence in radiology”. In: *Nature reviews. Cancer* 18.8 (2018), pp. 500–510. issn: 1474-1768. doi: 10.1038/S41568-018-0016-5.

[256] Ronneberger, Fischer, and Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Lecture Notes in Computer Science* 9351 (2015), pp. 234–241. issn: 16113349. arXiv: 1505.04597.

[257] Sharma and Bhatt. “Importance of Deep Learning Models to Perform Segmentation on Medical Imaging Modalities”. In: *Lecture Notes in Networks and Systems* 238 (2022), pp. 593–603. issn: 23673389. doi: 10.1007/978-981-16-2641-8_56.

[258] Isensee et al. “No New-Net”. In: *Lecture Notes in Computer Science* 11384 LNCS (2018), pp. 234–244. issn: 16113349. doi: 10.1007/978-3-030-11726-9_21. arXiv: 1809.10483.

[259] Çiçek et al. “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation”. In: *Lecture Notes in Computer Science* 9901 LNCS (2016), pp. 424–432. issn: 16113349. doi: 10.1007/978-3-319-46723-8_49. arXiv: 1606.06650.

[260] Antonelli et al. “The Medical Segmentation Decathlon”. In: *arXiv* (2021). eprint: 2106.05735.

[261] Menze et al. “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)”. In: *IEEE transactions on medical imaging* 34.10 (2015), pp. 1993–2024. issn: 1558-254X. doi: 10.1109/TMI.2014.2377694.

[262] Bilic et al. “The Liver Tumor Segmentation Benchmark (LiTS)”. In: *arXiv* (2019). issn: 2331-8422. arXiv: 1901.04056.

[263] Michael et al. “Breast Cancer Segmentation Methods: Current Status and Future Potentials”. In: *BioMed Research International* 2021 (2021). issn: 23146141. doi: 10.1155/2021/9962109.

[264] El Adoui et al. “MRI Breast Tumor Segmentation Using Different Encoder and Decoder CNN Architectures”. In: *Computers* 8.3 (2019), p. 52. issn: 2073431X. doi: 10.3390/COMPUTERS8030052.

[265] Zhang et al. “Hierarchical Convolutional Neural Networks for Segmentation of Breast Tumors in MRI With Application to Radiogenomics”. In: *IEEE Transactions on Medical Imaging* 38.2 (2019), pp. 435–447. issn: 1558254X. doi: 10.1109/TMI.2018.2865671.

[266] Wang et al. “Annotation-efficient deep learning for automatic medical image segmentation”. In: *Nature Communications 2021 12:1* 12.1 (2021), pp. 1–13. issn: 2041-1723. doi: 10.1038/s41467-021-26216-9.

[267] Hirsch et al. “Radiologist-Level Performance Using Deep Learning for Segmentation of Breast Cancers on MRI”. In: *Radiology, Artificial intelligence* 4.1 (2021), p. 200231. issn: 2638-6100. doi: 10.1148/RYAI.200231.

[268] Kamnitsas et al. “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation”. In: *Medical image analysis* 36 (2017), pp. 61–78. issn: 1361-8423. doi: 10.1016/J.MEDIA.2016.10.004. arXiv: 1603.05959.

[269] Badrinarayanan, Kendall, and Cipolla. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2015), pp. 2481–2495. issn: 01628828. doi: 10.1109/TPAMI.2016.2644615. arXiv: 1511.00561.

[270] Zhang et al. “Deep-learning method for tumor segmentation in breast DCE-MRI”. In: *SPIE Medical Imaging* 10954 (2019), pp. 97–102. issn: 16057422. doi: 10.1117/12.2513090.

[271] Khaled et al. “A U-Net Ensemble for breast lesion segmentation in DCE MRI”. In: *Computers in Biology and Medicine* 140 (2022), p. 105093. issn: 0010-4825. doi: 10.1016/J.COMPBIOMED.2021.105093.

[272] Piantadosi et al. “DCE-MRI breast lesions segmentation with a 3TP U-net deep convolutional neural network”. In: *IEEE 32nd International Symposium on Computer-Based Medical Systems* (2019), pp. 628–633. issn: 10637125. doi: 10.1109/CBMS.2019.00130.

[273] Newell, Giess, and Argus. “Practice Parameter for the Performance of Contrast-Enhanced Magnetic Resonance Imaging (CE-MRI) of the Breast”. In: *American College of Radiology* 1076 (2018), pp. 1–11.

[274] Rahimpour et al. “Cross-Modal Distillation to Improve MRI-Based Brain Tumor Segmentation With Missing MRI Sequences”. In: *IEEE transactions on bio-medical engineering* 69.7 (2022). issn: 1558-2531. doi: 10.1109/TBME.2021.3137561.

[275] Ma et al. “Loss odyssey in medical image segmentation”. In: *Medical image analysis* 71 (2021). issn: 1361-8423. doi: 10.1016/J.MEDIA.2021.102035.

[276] Hylton et al. “Neoadjuvant Chemotherapy for Breast Cancer: Functional Tumor Volume by MR Imaging Predicts Recurrence-free Survival-Results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL”. In: *Radiology* 279.1 (2016), pp. 44–55. issn: 1527-1315. doi: 10.1148/RADIOL.2015150013.

[277] Escobar et al. “Voxel-wise supervised analysis of tumors with multimodal engineered features to highlight interpretable biological patterns”. In: *Medical physics* 49.6 (2022), pp. 3816–3829. issn: 2473-4209. doi: 10.1002/MP.15603.

List of publications

Articles

Marie-Judith Saint Martin, Fanny Orlhac, Pia Akl, Fahad Khalid, Christophe Nioche, Irène Buvat, Caroline Malhaire, Frédérique Frouin. "A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study". In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 34.3 (2021), pp. 355-366. issn: 13528661. doi: 10.1007/s10334-020-00892-y.

Masoomeh Rahimpour*, **Marie-Judith Saint Martin***, Frédérique Frouin, Pia Akl, Fanny Orlhac, Michel Koole*, Caroline Malhaire*."Visual ensemble selection of deep convolutional neural networks for 3D segmentation of breast tumors on dynamic contrast enhanced MRI". In: *European Radiology*. In press. doi:10.1007/s00330-022-09113-7

* equal contribution

Caroline Malhaire, Fatine Selhane, **Marie-Judith Saint Martin**, Vincent Cockenpot, Pia Akl, Enora Laas, Audrey Bellesoeur, Catherine Ala-Eddine, Melodie Bereby-Kahane, Julie Manceau, Delphine Sebbag-Sfez, Jean-Yves Pierga, Fabien Reyal, Anne Vincent-Salomon, Hervé Brisse, Frédérique Frouin. "Association of pretherapeutic BI-RADS and breast edema MRI descriptors with breast cancer response after neoadjuvant chemotherapy: a systematic study". Submitted to *European Radiology*.

Confererence paper

M.-J. Saint Martin, F. Frouin, C. Malhaire and F. Orlhac, "Decrypting the information captured by MRI-radiomic features in predicting the response to neoadjuvant chemotherapy in breast cancer", *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 3227-3230, doi: 10.1109/EMBC48229.2022.9871844.

Selected as finalist for the Student Paper Competition

Conference abstracts

Marie-Judith Saint Martin, Fanny Orlhac, Pia Akl, Fahad Khalid, Christophe Nioche, Caroline Malhaire, Frédérique Frouin."Importance of applying ComBat besides image standardis-

ation in multicentre breast MRI radiomics: a phantom study". ESMRMB, S139, September 30 - October 2 2020, online.

Marie-Judith Saint Martin, Caroline Malhaire, Pia Akl, Delphine Sebbag Sfez, Frédérique Frouin, Fanny Orlhac."Estimation of a robust multi-scanner radiomic signature to predict the response to neoadjuvant chemotherapy in breast cancer using MRI". ESMRMB, S43, 7-9 October 2021, online.

Conference poster

Marie-Judith Saint Martin, Fanny Orlhac, Caroline Malhaire, Frédérique Frouin. "It is possible to recover the visual assessment of BI-RADS classification based on MRI using radiomic features?". ECR, C-13583, 13-17 July 2022, Vienna, Austria.

Masoomeh Rahimpour, **Marie-Judith Saint Martin**, Pia Akl, Delphine Sebbag Sfez, Fanny Orlhac, Caroline Malhaire, Michel Koole, Frédérique Frouin. "Clinical evaluation of CNN models for MRI-based breast lesion segmentation". ECR, 13-17 July 2022, Vienna, Austria.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

Titre: Modèles de prédiction de la réponse à la chimiothérapie néoadjuvante à partir d'examens d'IRM mammaire

Mots clés: Radiomique, IRM mammaire, apprentissage automatique, traitement d'images, apprentissage profond

Résumé: La chimiothérapie néoadjuvante (CNA) est devenue le traitement de référence des cancers agressifs ou localement avancés. Cependant, seulement 20 à 30% des patientes obtiennent une réponse pathologique complète (pCR). Être capable d'identifier les lésions chimiorésistantes avant le début du traitement améliorerait considérablement la prise en charge des patients. Dans cette perspective, la radiomique cherche à mieux exploiter les images et extrait des indices de forme, des indices issus de l'histogramme ou de texture pour construire des modèles d'aide à la décision. L'objectif de ce travail de thèse a été d'améliorer la prédiction de la réponse à la CNA en s'intéressant notamment aux problématiques de normalisation des images et d'exportabilité des modèles de prédiction. Nous avons travaillé sur une base clinique rétrospective multicentrique de 136 IRM mammaires constituée à l'Institut Curie et composée d'images pondérées en T1 après injection de produit de contraste et d'images pondérées en T2. La qualité des études radiomiques en IRM mammaire est sujette à trois limitations : le champ de biais magnétique affectant la distribution des intensités au sein du champ de vue, l'arbitraire de l'intensité dans les images et les variations d'intensité liées aux paramètres d'acquisition (machine, antenne, séquences...), appelées « effet scanner ». Une étude multi-machine réalisée sur deux fantômes de sein acquis suivant le protocole utilisé en clinique a mis en évidence la nécessité d'adapter pour le sein l'algorithme de correction de biais N4. L'intérêt d'harmoniser les indices radiomiques avec la méthode ComBat, après une étape de normalisation des images, a aussi été démontré. Cette chaîne de traitement a ensuite été adaptée à la base des patientes. Des analyses statistiques ont été menées pour identifier les indices robustes à la segmentation inter-radiologue. Nous avons aussi proposé une méthode de segmentation automatique des tumeurs par apprentissage profond, utilisant la fusion d'images pondérées en T1 après contraste et d'images de soustraction, dans l'objectif de réduire la charge de travail des radiologues et de rendre cette tâche plus robuste. Une chaîne de sélection de caractéristiques radiomiques a été proposée pour construire des modèles multiparamétriques. Les résultats ont montré l'intérêt d'associer les paramètres radiomiques issus de la région tumorale classique, de la tumeur binarisée placée dans une boîte englobante et d'une boîte de taille fixe située à l'intérieur de la tumeur. Ces modèles ont été testés sur une base indépendante multicentrique, harmonisée de façon originale pour pallier les limites de ComBat dans le cas de petits échantillons, ce qui a permis d'améliorer les performances dans 73% des expériences.

Title: Definition of predictive models to assess the response to neoadjuvant chemotherapy from breast magnetic resonance images

Keywords: radiomics, breast MRI, machine learning, image processing, deep learning

Abstract: Neoadjuvant chemotherapy (NAC) has become the standard treatment for locally advanced or invasive breast cancer, but with only 20 to 30% of patients achieving pathological complete response (pCR). Being able to predict non-responders to NAC would greatly improve patient care. In this context, the field of radiomics considers images as sources of a large amount of data and extracts shape, histogram-based and texture features to build decision-making tools. The goal of this thesis is to improve the prediction of pCR to NAC, with a particular focus on the normalization of images and the exportability of radiomic models. We used a retrospective multicentric database of 136 patients treated at Institut Curie, using T1-weighted dynamic contrast-enhanced and T2 images. In the literature, radiomic studies suffer from three main drawbacks: the bias field inhomogeneity creating regional intensity variations, the arbitrary units in which MR signal is expressed and the influence of acquisition parameters on feature values, called the "scanner effect". A multi-scanner study based on two breast phantoms imaged using the routine clinical protocol, highlighted the need to adapt the bias field correction N4 algorithm for the breast area. The need for further harmonization of features, using the ComBat method, after image normalization was also underscored. This pre-processing pipeline was then applied to patient data. Statistical analyses were carried out to identify features robust to inter-radiologist segmentation variabilities. A deep learning-based automatic segmentation approach using combined post-contrast T1-weighted and subtraction images was developed to reduce radiologists' workload and improve segmentation robustness. A pipeline to build multiparametric radiomic models was proposed. Results showed that combining features extracted from the standard tumor segmentation, from a bounding box on the binarized tumor images and from a constant box placed inside the tumor, improved performances. Models were tested on an independent multicentric test set, harmonized using an original method to overcome the limitations of the ComBat method in small datasets, that improved performances in 73% of experiments.