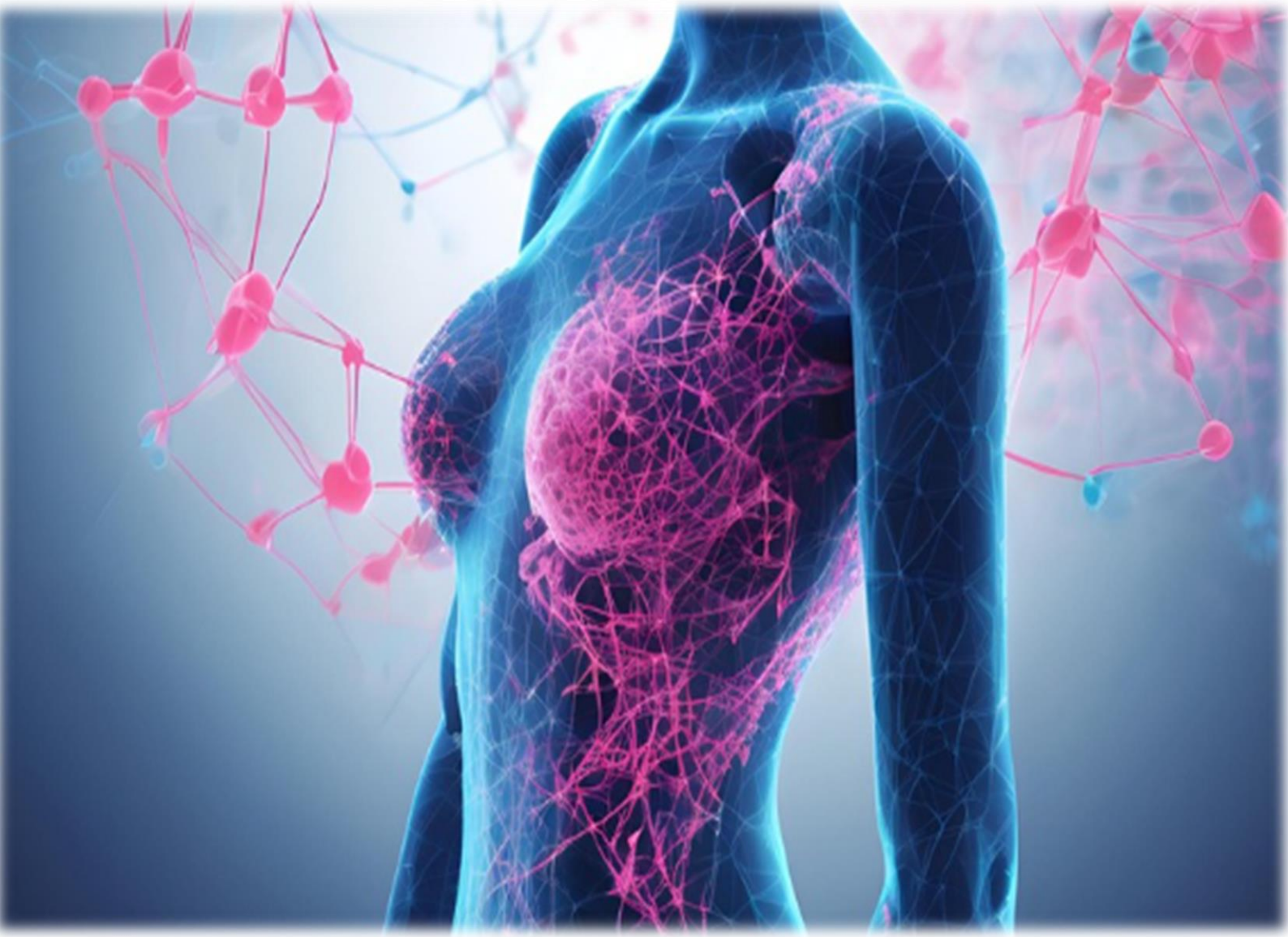

Harnessing AI to Predict a Pathologic Response in Breast Cancer Patients Undergoing Neoadjuvant Chemotherapy

J.A.N. Snoeijink

Harnessing AI to Predict a Pathologic Response in Breast Cancer Patients Undergoing Neoadjuvant Chemotherapy

Author:
J.A.N. Snoeijink, BSc

Supervision:
Prof. dr. C. Brune
Dr. A.L. Th. Imholz
Dr. J.M. Wolterink
MSc. I. van der Loo
MSc. E. Walter

Thesis – Master Technical Medicine
February 2025

UNIVERSITY OF TWENTE. Deventer ziekenhuis

Abstract

Introduction Breast cancer can be treated with neoadjuvant chemotherapy, which involves the use of chemotherapy before surgery. The main goal of this type of chemotherapy is to shrink the tumor and enhance the likelihood of achieving a pathologic complete response (pCR) following surgery. Despite the improvement of neoadjuvant chemotherapy in recent years, which has led to increased pCR rates in especially the Her2+ and triple-negative subgroups, reliable non-invasive biomarkers or imaging methods for the prediction of pCR are currently lacking.

Methods A total of 291 patients from Deventer Hospital between June 2005 and August 2023 were retrospectively enrolled. DCE-MRI features during the neoadjuvant chemotherapy were collected at three different time points, as well as clinical-pathological variables. For this purpose, a subset of DCE-MRI scans was used to evaluate two automatic, deep learning-based segmentation models: Zhang network and MAMA-MIA network. The radiological features were extracted with Pyradiomics, along with one intensity feature and delta features. All extracted features were used to train and test machine learning models for binary classification (pCR vs no pCR) and multi-class classification of the residual cancer burden score (RCB) (RCB-0 vs. RCB-1 vs. RCB-2 vs. RCB-3). For the binary classification, the model performance was assessed based on sensitivity, specificity, and the area under the curve (AUC). Accuracy and Cohen's kappa was used for the multi-class classification model, along with sensitivity, specificity and AUC for a one-vs-all classification (RCB-0 vs RCB-1, RCB-2, RCB-3).

Results This study showed that the MAMA-MIA network segments the breast tumors with the highest accuracy. Manual correction of the MAMA-MIA network segmentations was required for use in both clinical and research settings. The optimal trade-off between sensitivity and specificity for pCR and RCB-0 was achieved using the first two MRIs during the neoadjuvant chemotherapy, along with clinical and radiological data. For pCR and RCB-0 the model showed sensitivities of 0.75 and 0.81, and specificities of 0.83 and 0.80, respectively. The multi-class classification model for RCB showed an accuracy of 0.69 and a Cohen's kappa of 0.51.

Conclusion This study highlights the potential of combining clinical and radiological features with machine learning to predict pCR in breast cancer patients undergoing neoadjuvant chemotherapy. The publicly available deep learning segmentation networks are not robust enough for the Deventer Hospital MRIs, and it is therefore recommended to explore retraining them using the hospital's data. To optimize the prediction of pCR and RCB, it is recommended to expand the dataset, apply external validation, and explore deep learning approaches. Such advancements could lead to more accurate predictions of pCR, ultimately enabling more personalized breast cancer care.

Keywords Breast cancer, Artificial intelligence, Prediction models, Multimodal prediction

Graduation Committee

Chair
Prof. dr. C. Brune
Mathematics of Imaging & AI (MIA), University of Twente, Enschede, The Netherlands

Clinical supervisor
dr. A.L. Th. Imholz
Internal Medicine, Oncologist, Deventer Ziekenhuis, Deventer, The Netherlands

Technical supervisor
dr. J.M. Wolterink
Mathematics of Imaging & AI (MIA), University of Twente, Enschede, The Netherlands

Technical medicine supervisor
MSc. I. van der Loo
Technical Medicine, Research Agency, Deventer Ziekenhuis, Deventer, The Netherlands

Process supervisor
MSc. E. Walter
Technical Medicine, University of Twente, Enschede, The Netherlands

External member
dr. R.F.M. van Doremalen
Technical Medicine, Innovation lab and 3D lab, Medisch Spectrum Twente, Enschede, The Netherlands

Acknowledgements

Over the past seven years, I have enjoyed my Technical Medicine study at the University of Twente. This study continuously fueled my curiosity in the medical-technical sector and challenged me to explore and learn about new topics. It has shaped my perspective towards innovations in healthcare, allowing me to contribute meaningfully to the society.
During my study, I developed an interest in artificial intelligence. However, the curriculums of my study did not perfectly lend itself to applying this interest. I am therefore grateful that I was able to pursue this interest during my TG internships, particularly in my M3 year. During this M3 year, I enjoyed working on using AI in breast cancer care, hoping to enable more personalized care in the future.

This M3 year would not have been possible without the passionate and professional guidance of Dr. A.L. Th. Imholz and MSc. I. van der Loo at Deventer Hospital. I want to thank Alex and Iris for their time and for creating a workspace that allowed me to bring out the best in myself. The approachable environment and the professional input were key elements in this process. I would also like to thank AIOS Demi van der Oord–Hekman. It was a pleasure working together both during the research and in the clinical setting.

Additionally, I would like to thank Prof. Dr. C. Brune and Dr. J.M. Wolterink from the University of Twente for their technical support. Your insights helped me maintain a technical perspective on the subject, making this research and thesis more complete. Through you, I also had access to the cluster, which provided me with sufficient computing power to bring my ideas to life.

Furthermore, I would like to thank MSc. E. Walter, who guided me in my professional development during my M2 and M3 years. I am grateful for Elyse's time, support, and guidance in my journey of self-discovery, which allowed me to grow further.

Finally, I would like to thank my parents, Henk and Zwanet. You have always encouraged me to follow my interests and dreams. I would not have reached this milestone without your endless support.

I hope you enjoy reading this thesis.

Contents

List of Abbreviations

AI	Artificial intelligence
ANOVA	Analysis of variance
ASA score	American society of anesthesiologist physical status
AUC	Area under the curve
BMI	Body mass index
BI-RADS	Breast imaging reporting % data systems
CNN	Conventional neural network
CS	Concentric shrinkage
DD	Diffuse decrease
DIO	Decrease of intensity only
DCE	Dynamic contrast-enhanced
DSC	Dice similarity coefficient
DT	Decision trees
EUSOBI	European society of breast imaging
ER	Estrogen
FISH	Fluorescence in situ hybridization
GPU	Graphics processing unit
Grad-CAM	Gradient-weighted class activation mapping
HER2	Human epidermal growth factor
Her2+	Her2-positive breast cancer
HSD	Hausdorff distance
HR+	Hormone receptor positive breast cancer
KNN	K-Nearest neighbors
MAI	Mitotic activity index
microRNA	Micro ribonucleic acid
MRI	Magnetic resonance imaging
NB	Naïve bayes
nnU-Net	no-new U-Net
pCR	Pathologic complete response
PACS	Picture archiving and communications system
RCB	Residual cancer burden
RF	Random forest
RHPC	Radiological, histopathological, personal and clinical
ROI	Region of interest
ROC	Receiver operating characteristics
RFE	Recursive feature elimination
RFS	Recurrence free survival
SCC	Spearman correlation coefficient
SD	Stable disease
SGD	Stochastic gradient descent
SVM	Support vector machines
TNBC	Triple negative breast cancer
TPOT	Tree-based pipeline optimization tool
TSP	Tumor shrinkage patterns
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting

List of Figures

List of Tables

1 INTRODUCTION

1.1 Breast Cancer

1.1.1 Classification of Breast Cancer

Breast cancer ranks as the most commonly diagnosed cancer in women worldwide [1]. The Netherlands recorded 15,634 new cases of invasive breast cancer in 2023 [2]. Breast cancer can be subdivided into various subtypes based on the expression of hormone receptors such as estrogen (ER) and progesterone (PR), as well as the human epidermal growth factor receptor 2 (HER2). Depending on the presence or absence of the receptors, breast cancer can be grouped into three primary subtypes: hormone receptor-positive breast cancer (HR+), Her2-positive breast cancer (Her2+), and triple-negative breast cancer (TNBC) (Table 1).

Table 1 *Subtypes of breast cancer. ER: estrogen receptor, PR: progesterone receptor, Her2: human epidermal growth factor, HR+: hormone receptor-positive breast cancer, Her2+: Her2-positive breast cancer, TNBC: triple-negative breast cancer, FISH: fluorescence in situ hybridization (test for assessing the DNA of the breast tumor for amplification of HER2 gene) [3,4]*

Subtypes	***ER***	***PR***	***Her2***
HR+	+ or -	+ or -	-
Her2+	+ or -	+ or -	+ or FISH amplification
TNBC	-	-	-

1.1.2 Imaging Modalities in Breast Cancer

The initial diagnostic imaging modality for breast cancer is a mammography. On mammography, radiologists assess abnormalities, such as architectural distortion, calcifications, asymmetry, and mass. Subsequently, an ultrasound-guided biopsy may be performed to obtain tissue samples for pathological examination [5]. This procedure is essential for not only confirming the diagnosis but also for evaluating the receptor status of the tumor. Determining the receptor status is crucial for guiding treatment decisions for the patient. Additionally, although breast magnetic resonance imaging (MRI) can help measure tumor size and assess treatment response, it is not conducted for every breast patient [6]. If an MRI is performed, the guidelines of the European Society of Breast Imaging (EUSOBI) recommend using a T1-weighted Dynamic Contrast-Enhanced (DCE) sequence. [7] In this sequence, a paramagnetic contrast agent is intravenously injected, altering the recovery time of water molecules in the body after exposure to a magnetic field, which adjusts the contrast in T1-weighted images [8]. Analyzing changes in tissue contrast over time allows for determining the extent of tissue vascularization, interstitial space composition and lesion differentiation [6].

1.1.3 Therapy

The primary treatments for breast cancer can be divided into two categories: local- and systemic treatments [9]. The local treatments target a specific organ or a confined area of the body, such as surgery and radiotherapy [10]. On the other hand, systemic therapy refers to treatments that circulate through the bloodstream to reach and affect cells all over the body. These treatments are usually divided into three categories: (a) conventional cytotoxic chemotherapy, (b) hormonal agents, and (c) targeted therapy or immunotherapy [11]. The clinical indication for a specific treatment is determined based on several factors, for example, the tumor's molecular subtype, the grade of the tumor, and the stage of the tumor [5,12,13].

One important application of systemic therapy is neoadjuvant chemotherapy, which involves the use of chemotherapy before surgery [14]. This form of therapy is mainly considered in two specific types of breast tumors: Her2+ and TNBC. The main goal of this type of chemotherapy is to shrink the tumor and enhance the likelihood of achieving a pathologic complete response

(pCR) following surgery. A pCR signifies the absence of residual cancer in both the breast tissue and the lymph nodes. The chance of achieving a pCR is approximately 40% [15] for Her2+ tumors and 40-50% [16] for triple-negative tumors. The benefits of neoadjuvant systemic therapy, besides increased likelihood of reaching pCR, include a greater probability of breast-conserving surgery and the removal of smaller tumor volumes during these procedures, leading to better cosmetic results [17].

1.1.4 Therapy Response Evaluation

1.1.4.1 Radiological Complete Response

The response evaluation assesses the effectiveness of the treatment regimen. In the case of neoadjuvant chemotherapy, this evaluation considers both the radiological and pathological response assessment. A radiological complete response (rCR) indicates the absence of visible tumor lesions on radiological imaging following neoadjuvant systemic therapy and is assessed by the radiologist. Patients with rCR presented a 3-year recurrence-free survival (RFS) of 92.8% in all subtypes, in contrast to the 74.8% observed in the absence of rCR, across the entire breast cancer population [18]. Similarly, other research showed a 2.4-fold increase in risk of recurrence for patients without rCR compared to patients with rCR [19].

1.1.4.2 Pathologic Complete Response

Neoadjuvant chemotherapy was initially developed to reduce tumor size, making surgical interventions easier and increasing the likelihood of complete tumor removal. In addition to this, achieving a pCR after chemotherapy is associated with improved survival outcomes. A pCR indicates the absence of residual cancer in breast tissue and lymph nodes after neoadjuvant chemotherapy upon examination of the resected tissue by the pathologist. Gampenrieder et al, suggest that patients achieving pCR generally experience better outcomes during follow-up. In a retrospective analysis encompassing all subtypes of breast cancer, the patients who achieved pCR had a significantly lower risk of recurrence of death, with a 3-year RFS of 94.4% compared to 78.3% for those without pCR [18]. Besides this, in HR+ breast cancer, the concurrence between rCR and pCR, is in general, about 30%, whereas it reaches approximately 50% in Her2+ breast cancer. Additionally, within Her2+ breast carcinomas, there are subgroups that respond differently to neoadjuvant chemotherapy, impacting rCR and pCR. For Her2+/HR- cases, indications suggest an 87% concurrence between rCR and pCR. In the case of Her2+/HR+, this concurrence is about 53% [20]. For TNBC, agreement between rCR and pCR is attained in nearly 30% [18].

Various prognostic factors are recognized for guiding decisions regarding neoadjuvant chemotherapy. These factors encompass patient-related variables, such as age, performance status, and body mass index (BMI), as well as tumor-related variables, including tumor size, hormone receptor status, Her2 status, tumor grade, and lymph node status. Moreover, high levels of the Ki-67 protein [21], an increased number of micro ribonucleic acid (microRNAs) [22], and an increased number of tumor-infiltrating lymphocytes [23] are all associated with a greater likelihood of achieving pCR following neoadjuvant chemotherapy.

1.2 Study Aim

Despite the improvement of neoadjuvant chemotherapy in recent years, which has led to increased pCR rates, especially in the Her2+ and triple-negative subgroups, reliable non-invasive biomarkers or imaging methods for the prediction of pCR are currently lacking [24]. Additionally, there is only a 30% concordance between rCR and pCR in HR+ and TNBC breast cancer. Nowadays, the conformation of pCR can only be obtained through surgical resection of the breast tissue followed by a histopathological examination. For these reasons, this research will investigate the main question: *How can artificial intelligence be used to predict a pathologic response based on clinical-pathological and radiological information among patients diagnosed with stadium I-III breast cancer receiving neoadjuvant chemotherapy?*

An artificial intelligence (AI) model capable of reliably predicting the pathological response in breast cancer patients after neoadjuvant treatment could further personalize the care of these patients. In the future, it may even be possible to safely omit surgery for some patients with a positive pCR prediction after chemotherapy. This approach would alleviate the patient's physical and emotional burden. Additionally, it would lead to a reduction in healthcare costs. Moreover, during neoadjuvant chemotherapy, patients currently receive a fixed number of chemotherapy cycles; the assessment of radiological response during treatment does not alter this regimen. With the help of the AI model, the pathologic response can be monitored throughout treatment, enabling adjustments to the treatment plan and potentially reducing the number of chemotherapy cycles for each individual patient [20].

2 BACKGROUND

2.1 Cancer

The ability of cells to repair and renew themselves is essential for maintaining healthy tissues and organs in the human body. This process is regulated by a network of signaling pathways that control the cell cycle, DNA repair, and apoptosis. When these regulatory mechanisms are disrupted, it can lead to uncontrolled cell division, resulting in cancer [25].
Mutations in specific genes, namely proto-oncogenes and tumor suppressor genes, play an important role in this process, leading to uncontrolled cell division. Proto-oncogenes are genes responsible for producing proteins that stimulate cell growth and division. A mutation in a proto-oncogene results in an oncogene. This oncogene causes abnormal cell proliferation even without the presence of growth factors. Continuous stimulation of cell proliferation can lead to cell accumulation and tumor formation. On the other hand, tumor suppressor genes play a crucial role in inhibiting cell growth and promoting DNA repair or apoptosis in case of severe cell damage. Mutations in tumor suppressor genes can cause these genes to lose their regulatory functions. This allows damaged cells to survive and divide uncontrollably, contributing to tumor formation [25].

The concept of cancer has been further refined by the identification of the 'hallmarks of cancer', which describe the essential characteristics of cancer cells. According to Hanahan and Weinberg [26], these hallmarks include the following eight biological capacities (Figure 1):

1. Sustaining proliferative signaling: cancer cells can produce their own growth signals or bypass growth signals.
2. Evading growth suppressors: cancer cells ignore signals that would stop normal cells from growing.
3. Resisting cell death: cancer cells can avoid apoptosis, allowing damaged cells to stay alive and multiply.
4. Enabling replicative immortality: cancer cells can divide indefinitely by increasing telomerase activity, for example.
5. Inducing or accessing vasculature: cancer cells promote the formation of new blood vessels to supply themselves with nutrients and oxygen.
6. Activating invasion and metastasis: cancer cells can invade surrounding tissues and spread to other parts of the body.
7. Reprogramming cellular metabolism: cancer cells alter the energy production and metabolic pathways to support rapid growth and survival, even under conditions that would be unfavorable to normal cells. This often involves a shift to aerobic glycolysis, allowing cancer cells to produce energy even in low-oxygen environments.
8. Avoiding immune destruction: cancel cells have the ability to evade detection and elimination by the body's immune system.

In addition to the eight core hallmarks of cancer, two enabling characteristics have been identified [26]. These characteristics facilitate the emergence and persistence of the core hallmarks, helping to drive cancer progression and adaptation (Figure 1):

1. Genome instability and mutation: Genomic instability fuels the mutation process, allowing cancer cells to rapidly acquire genetic changes that drive tumor progression. These mutations can affect key genes that control cell growth, death, and differentiation, providing cancer cells with the flexibility to adapt and survive under various conditions, such as during metastasis or treatment resistance.
2. Tumor-promoting inflammation: Chronic inflammation in the tumor microenvironment acts as a key driver of cancer progression. Inflammatory cells and cytokines can promote tumor growth by supplying bioactive molecules such as growth factors, survival signals,

and pro-angiogenic factors. This inflammation aids in nearly every stage of cancer development, including proliferation, invasion, and metastasis.

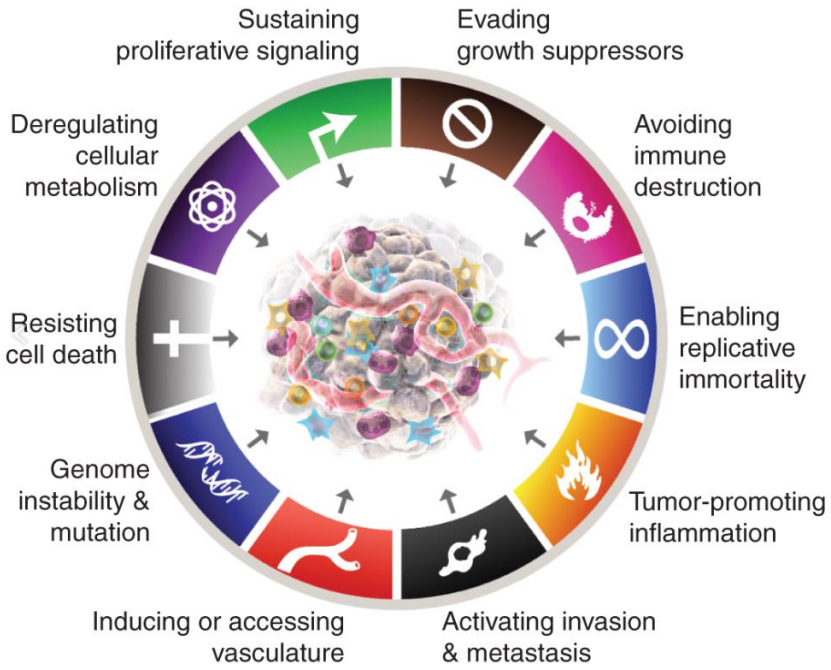

Figure 1 *Hallmarks of Cancer. Figure taken from [26].*

In 2023, there were 128,927 new cases of cancer in the Netherlands [2]. The most commonly diagnosed types of cancer are breast, lung, prostate, colon, and rectal- cancer. Among women, breast cancer is the most frequently diagnosed cancer, with an incidence of 15,634 cases [2]. The four most prevalent types of breast cancer are [27]:

1. *Ductal Carcinoma In Situ (DCIS):* DCIS is an early stage of breast cancer where abnormally growing epithelial cells are localized within the ducts of the breast. The cancer cells remain at their original location and have not invaded to other parts of the breast or body .
2. *Invasive Ductal Carcinoma (IDC):* IDC originates in the ducts and is invasive. This means the cancer cells invaded to the surrounding breast tissue. IDC has the potential to invade further to other parts of the body via blood vessels or lymph nodes.
3. *Lobular Carcinoma In Situ (LCIS):* LCIS originates in the lobules of the breast and is non-invasive.
4. *Invasive Lobular Carcinoma (ILC):* ILC, like LCIS, starts in the lobules of the breast but has invaded into the surrounding tissue.

For the early detection of breast cancer, the Netherlands uses a national breast cancer screening program, which is available to women aged 50-75 years. Every two years, these women are invited to undergo a mammogram, which is assessed by a radiologist using the Breast Imaging Reporting and Data System (BI-RADS) [28]. This classification includes the amount of fibroglandular tissue, background parenchymal enhancement, and the assessment category. The use of BI-RADS aids in the systematic evaluation of mammography images and supports decision-making regarding additional tests or treatments.

2.2 Clinical and Radiological Variables

Neoadjuvant chemotherapy followed by surgical removal of pathological tissue is one of the treatment options for breast cancer. The goal of neoadjuvant chemotherapy is to shrink the tumor as much as possible, making surgical removal easier and increasing the likelihood of complete tumor removal. In addition to this, achieving a pCR after chemotherapy is associated with improved surgical outcomes. This study examined how clinical and radiological variables can be integrated with artificial intelligence (AI) to predict pCR in breast cancer patients after neoadjuvant chemotherapy. Developing a predictive model that integrates clinical and radiological factors associated with pCR could support the personalization of treatment strategies. This study included clinical parameters, such as patient characteristics and tumor type, as well as radiological features derived from DCE-MRI.

2.2.1 Clinical variables

2.2.1.1 Prognostic Factors

Various prognostic factors are recognized for guiding decisions regarding neoadjuvant chemotherapy. These factors can encompass patient-related variables, such as age, performance status, menopausal status, and BMI [29]. The performance status can be indicated by the American Society of Anesthesiologists Physical Status (ASA score). This scoring system works with a score of 1 to score 5, where score 1 represents a healthy patient and score 5 describes a patient who, without treatment, dies within 24 hours. In addition, tumor-related variables such as tumor size, hormone receptor status, Her2 status, tumor grade, and lymph node status are considered prognostic [30]. Mainly, Her2+ and triple-negative subtypes show better prognosis and favorable response rates after neoadjuvant chemotherapy than HR+ subtypes [31].

2.2.1.2 Blood profile

The patient's blood profile may also influence the treatment's effectiveness. For example, platelets not only contribute to tumor growth and spread but also play a role in reducing the efficacy of chemotherapy. They promote tumor proliferation by releasing growth factors, enabling tumors to counteract the effects of chemotherapy [32]. This decreases the effectiveness of the treatment and makes tumors more resistant to therapies. Furthermore, a reduced hemoglobin concentration can lead to hypoxia in tumors. Hypoxia stimulates the overexpression of genes involved in drug resistance, making tumors more resilient to chemotherapy. This occurs because hypoxia increases the activity of efflux pumps, which actively expel chemotherapeutic agents from tumor cells [33]. As a result, intracellular drug concentrations drop, further reducing the effectiveness of the treatment.

2.2.1.3 Ki-67 and microRNA

Yerushalmi et al. describe the Ki-67 protein as a prognostic factor. The Ki-67 protein is associated with cellular proliferation. As a result, a high level of Ki-67 correlates with a worse prognosis and it can be used to predict pCR after neoadjuvant chemotherapy [34,35]. Among histological grade characteristics, only the Mitotic Activity Index (MAI) was proven to be of prognostic value [36]. The MAI is the most widely used method to estimate mitotic activity and is defined as the number of mitotic figures in a given area of the tumor [37]. This, along with the extent of tube formation and nuclear pleomorphism is incorporated into the Bloom-Richardson grading system.

Also, sixty of the 123 microRNAs seem to have a possible association with prognosis and neoadjuvant response [22]. MicroRNAs are small, non-coding RNA molecules, which play a role in the regulation of gene expression. They bind to messenger RNA (mRNA) and can inhibit

the translation of mRNA to proteins or degrade the mRNA. As a result, they affect various biological processes such as differentiation, proliferation, and apoptosis [38]. Lastly, neutrophil count, lymphocyte count, neutrophil/lymphocyte ratio, and thrombocytes/lymphocytes ratio seem to have predictive value for pCR [32,39,40].

2.2.2 Radiological features

Radiomics is a widely used approach for defining features in advance for a predictive model [41]. Radiomics aims to extract quantitative information from diagnostic images. These radiomics features are calculated based on the segmentation of the tumor. The features can be divided into three different categories [42] (Figure 2):

1. Intensity features: first-order statistics features
2. Shape features: shape-based features
3. Texture features: gray level co-occurrence matrix features, gray level run length matrix features, gray level size zone matrix features, neighboring gray-tone difference matrix features, and gray level dependence matrix features.

A systematic review and meta-analysis using MRI radiomics features for the prediction of pCR in breast cancer patients undergoing neoadjuvant chemotherapy shows a mean area under the curve (AUC) of 0.78 [43]. Nardone et al. conducted a systematic review that highlights several studies predicting pCR with delta radiomics, analyzing features at different acquisition time points, often before and after therapy (the so-called delta features) [44]. Guo et al. included 140 patients who underwent DCE-MRI both before and after the first cycle of chemotherapy. This article showed that a prediction model with the features before the start of chemotherapy, the features after the first cycle of chemotherapy, and the delta features combined achieved an AUC of 0.87 [45].

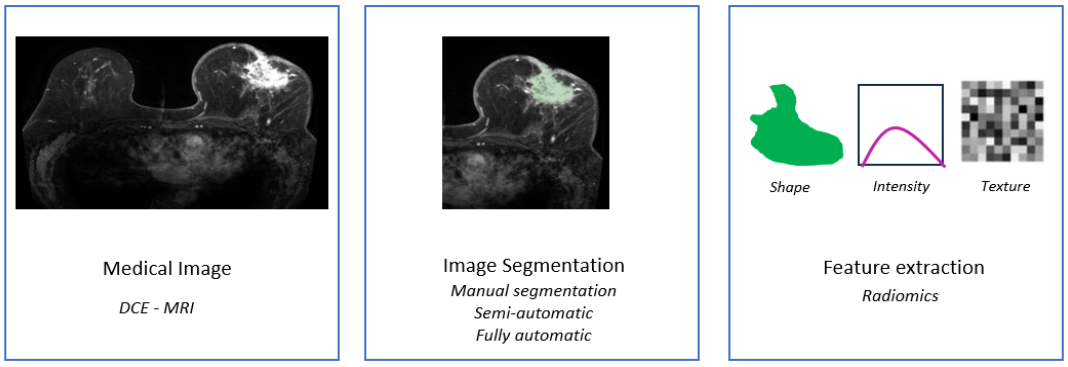

Figure 2 *Radiomics workflow: starting with image acquisition (in this study: DCE-MRI), followed by image segmentation and feature extraction.*

Additionally, a recent article demonstrates that MRI-based tumor shrinkage patterns (TSP) during neoadjuvant chemotherapy are associated with pCR [46]. This study included 362 patients, with TSP classified into four categories: concentric shrinkage (CS), diffuse decrease (DD), decrease of intensity only (DIO) and stable disease (SD). Furthermore, the CS pattern was further specified as: simple CS, CS to small foci, and CS plus decreased enhancement. The DD was also further specified as: concentric shrinkage with surrounding lesions and residual multinodular lesions. TSP determination in this study was conducted by two breast radiologists. Results suggested that a DD pattern in HR+/Her2- patient strongly predicts pCR, while an SD pattern in Her2+ patients and triple negative patients suggested a non-pCR.

2.3 Prediction Models

2.3.1 Supervised learning

Prediction models based on machine learning use predefined features to identify patterns in the data, which are then used to estimate the outcome variable—in this case, pCR in breast cancer patients. Machine learning algorithms can identify patterns through various approaches, including supervised, unsupervised, and semi-supervised learning. This study utilized supervised learning, where a function is optimized to map input features to outcome variables based on sample input-output pairs (Figure 3). To optimize this function, the data is typically divided into training and test sets (Figure 3). The training set is used to train the model, allowing it to learn underlying patterns and relationships between the input features and the outcome variable. The test set remains untouched during the training process and is used solely to evaluate the model's performance. This ensures an unbiased assessment of the model's ability to generalize to new, unseen data. In addition to the train- and test set, the training set is often further subdivided into a primary training subset and a validation subset. The validation set plays a role in hyperparameter tuning. Hyperparameters—such as the number of neighbors in a K-nearest neighbors (KNN) model or the maximum depth of a tree in a Decision Tree (DT) model—are predefined settings. By using the validation set, different hyperparameter configurations can be tested to identify the optimal combination for achieving the best model performance [47].

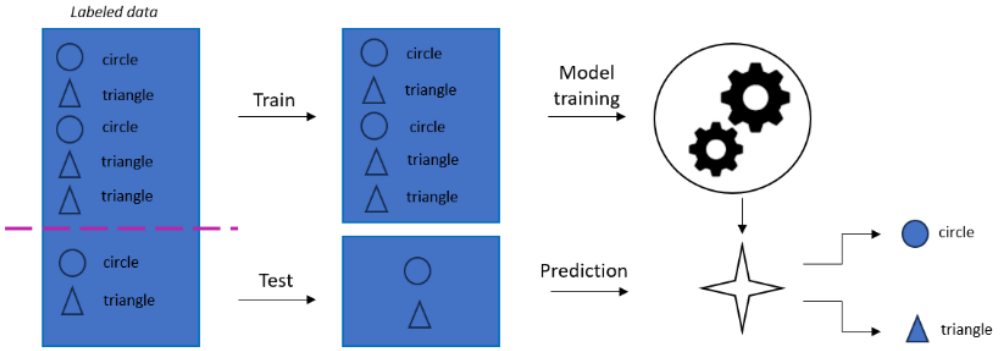

Figure 3 *Supervised learning with splitting data into training set and test set.*

2.3.2 Machine learning algorithms

Various machine learning models can be fit on the training set. Some commonly used approaches are [48]:

- Naive Bayes:
  Naive Bayes (NB) is a probabilistic classifier based on Bayes' theorem, which is a fundamental principle in probability theory (Equation 1) that quantifies the likelihood of a hypothesis by incorporating prior knowledge and updating it with new information. Besides this, the NB classifier used the assumption that features are conditionally independent given the class label.

$$P\ (A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

Equation 1 P(A|B): posterior probability – probability of event A occurring given event B, P(B|A): likelihood – probability of event B occurring given event A, P(A): prior probability – initial probability of event A occurring, P(B): marginal likelihood – total probability of event B occurring, regardless of event A

Common variants of the NB classifier are, for example, the Gaussian NB classifier or the Bernoulli NB classifier. The Gaussian NB classifier works well for continuous data and assumes the features follow a Gaussian distribution. It uses the probability density function of a Gaussian distribution to calculate the probability of each feature given a class (Equation 2). The total probability for the class is calculated by multiplying the probabilities for all features. These feature probabilities, along with the prior probability of the class and the marginal likelihood, are combined using Bayes' Theorem. This process is repeated for each class, and the class with the highest posterior probability P(A|B) is selected as the predicted class.

$$P\ (B|A) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) * \frac{(e^{-}(x-\mu)^2}{2\sigma^2} \quad (2)$$

Equation 2 P(B|A): likelihood – probability of event B occurring given event A, μ: mean of feature, σ: standard deviation of feature

On the other hand, the Bernoulli NB classifier is designed for binary data and assumes that each feature follows a Bernoulli distribution. This classifier follows the same workflow as the Gaussian Naive Bayes classifier, but instead of using the probability density function for the Gaussian distribution, it uses the probability mass function for the Bernoulli distribution (Equation 3).

$$P\ (B|A) = p^B(1-p)^{1-B} \quad (3)$$

Equation 3 P(B|A): likelihood – probability of event B occurring given event A, p^B: indicator function

- <u>Logistic Regression:</u>
  Logistic regression typically uses a sigmoid function to estimate the probabilities of a class, given the features (Equation 4). This sigmoid function transforms the linear combination of input features into a value between 0 and 1, which can be interpreted as probability. The goal of training this model is to adjust the weights (b_0, b_1, b_2 ….., b_n) such that the predicted probabilities accurately represent the likelihood of the class. The predicted probability can be compared to a threshold to predict the actual class.

$$P(A|B) = \frac{1}{1 + \ e^{-(b0,+b1x1+b2x2+\cdots+bnxn)}} \quad (4)$$

Equation 4 P(A|B): probability of event A occurring given event B, b0: bias, b1,b2…bn: weights, x1,x2…x2n: input features

- K-Nearest Neighbors:
  K-Nearest Neighbors (KNN) is a supervised learning algorithm used for classification and regression. In KNN classification, a data point is classified based on the majority class among its *k* closest neighbors in the features space (Figure 4). For regression, it predicts the value based on the average of the values of its nearest neighbors.

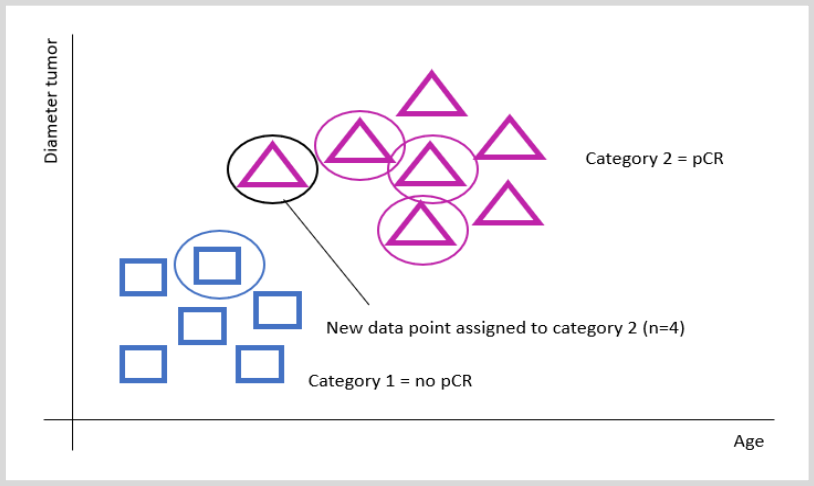

Figure 4 *Example of KNN with hypothetical data. In this case, the new data point is classified based on its four closest neighbors, resulting in a prediction of category 2. pCR = pathologic complete respons.*

- Decision Trees:
  Decision Trees (DT) are supervised learning models commonly used for both classification and regression. In a decision tree, the model learns to make decisions by iteratively splitting the data into branches based on specific feature values. Each split is guided by decision rules that aim to separate the data in a way that increases the homogeneity of the resulting group. The tree structure begins at a root node and splits down through branches, where each internal node represents a feature-based decision. This process continues until the model reaches leaf nodes, which represent final predictions or outcomes (Figure 5).

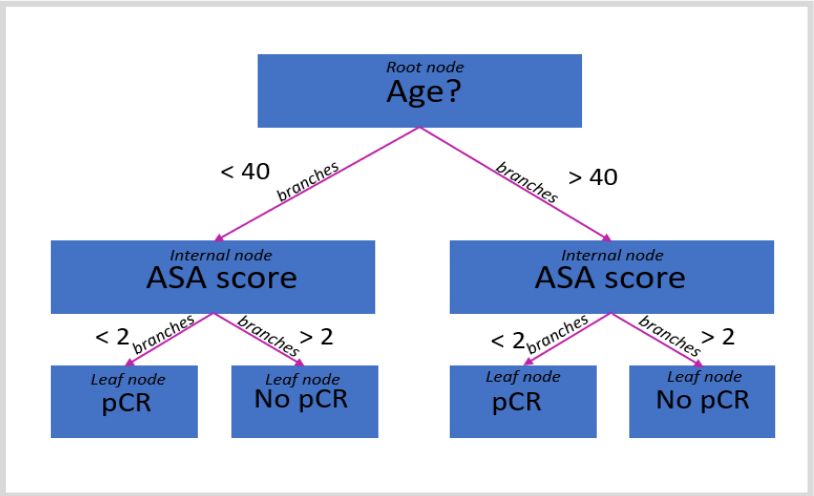

Figure 5 *Simplified example of a decision tree with hypothetical data. Each internal node represents a decision based on a feature, while the branches indicate the possible outcome of that decision. The leaf nodes represent the predicted class, depending on the input variable. ASA score = American Society of Anesthesiologists physical status, pCR = pathologic complete response.*

- Random Forest:
  Random Forest (RF) is an ensemble learning method that combines multiple decision trees to improve prediction accuracy. Ensemble learning is a technique in machine learning where multiple models are trained to solve the same problem. By combining the predictions from several models, ensemble learning aims to produce a more accurate and stable prediction than any single model could achieve. In the case of a random forest, each decision tree is built on a random subset of the data and features, which introduces diversity among the trees. This approach employs 'parallel ensembling', where multiple decision tree classifiers are trained simultaneously on different subsets of the dataset. The final prediction is made by aggregating the outcomes from all trees, typically through majority voting for classification or averaging for regression (Figure 6).

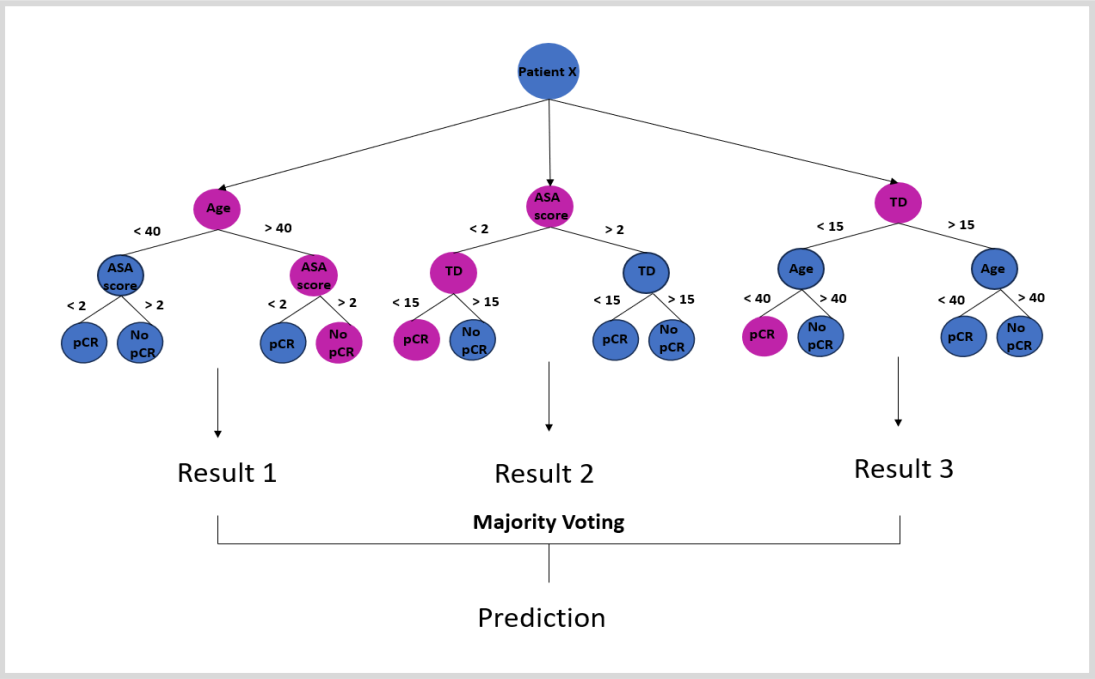

Figure 6 *Example of a random forest, where the outcome is determined by majority voting. ASA score = American society of anesthesiologist physical status, TD = tumor diameter, pCR = pathologic complete response.*

- Support Vector Machines:
  Support Vector Machines (SVM) is a supervised learning model and works by finding the optimal hyperplane that best separates the different classes in the feature space (Figure 7). This hyperplane is essentially a boundary that maximizes the margin between classes, where the margin is defined as the distance between the hyperplane and the nearest data point of each class. The nearest points are called support vectors, as they are the critical elements that define the margin and, therefore, the placement of the hyperplane. In cases where classes are not linearly separable, SVM can use a kernel trick to transform the feature space into higher dimensions, making it possible to find a separating hyperplane in this transformed space. This flexibility allows SVM to perform well even with complex and non-linearly separable data.

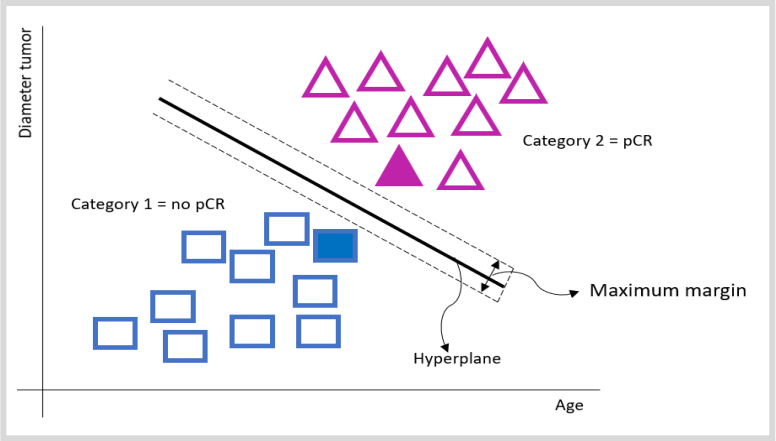

Figure 7 *Example of a support vector machine. pCR = pathologic complete response.*

- Gradient Boosting:
  Gradient Boosting is a supervised learning technique widely used for both classification and regression tasks. In gradient boosting, the model is built iteratively, where each new model (usually a decision tree) is trained to correct the residual errors—the differences between the predicted and actual values—from previous models. The process begins with an initial model, typically a simple decision tree, that provides a rough set of predictions. Each subsequent model then attempts to refine this by "boosting" the performance, focusing specifically on areas where the previous models performed poorly. This process of correcting errors continues iteratively, with the ensemble of models collectively improving prediction accuracy.

During the training of machine learning models, optimization algorithms are used to optimize the model's parameters, such as the weights in a logistic regression model. One example of an optimization algorithm is Stochastic Gradient Descent (SGD). SGD is an optimization algorithm widely used in machine learning to minimize a loss function, which measures a model's performance. In SGD, the gradient represents the slope of the loss function, guiding how the model parameters should adjust to reduce error. Unlike traditional gradient descent, which uses the full dataset to calculate the gradient, SGD randomly selects a small batch or even a single data point at each iteration. This makes SGD faster but less stable, as it can cause the algorithm to oscillate around the optimal solution before converging. Another example of optimization algorithms is extreme gradient boosting (XGBoost), a variant of gradient boosting. This is an optimized version of gradient boosting that introduces several enhancements to improve speed, accuracy, and scalability. While it retains the iterative process of gradient boosting, XGBoost incorporates features like L1- and L2 regularization. L1 regularization adds a penalty proportional to the absolute values of the model's parameters. This can drive some parameters to zero, effectively performing feature selection by eliminating less important features. L2 regularization, on the other hand, adds a penalty proportional to the square of the parameters, which encourages smaller values but doesn't necessarily drive parameters to zero. This helps prevent the model from becoming too complex and improves its generalization ability to new data. Together, L1 and L2 regularization help improve model performance by making it more robust and less likely to overfit.

2.3.3 Tree-Based Pipeline Optimization Tool

The accuracy of classification tasks can vary significantly due to the wide range of model configurations and hyperparameter settings in machine learning. The Tree-based Pipeline Optimization Tool (TPOT) addresses this challenge by automating the search for the optimal model and hyperparameters. TPOT uses genetic programming (GP), an evolutionary algorithm, to optimize machine learning pipelines automatically (Figure 8). It begins by generating a random population of pipelines, each combining feature preprocessing operators (like the min-max scaler and robust scaler) to modify the dataset, followed by supervised classification operators (such as logistic regression, k-nearest neighbor classifiers, decision trees, and random forests). TPOT also incorporates feature selection techniques, such as variance threshold, select percentile, and select family-wise error rate, to optimize the feature space. Besides this, TPOT also optimizes hyperparameters within the pipelines. Each pipeline is evaluated based on its performance (fitness), and the top performers are selected. These selected pipelines undergo crossover (combining elements of the best pipelines) and mutation (introducing random changes) to create new variations. This process is repeated across multiple generations, progressively evolving through natural selection principles to identify the most effective pipeline for the given problem. The primary goal of TPOT is to maximize classification accuracy by systematically evaluating and refining combinations of feature selectors, preprocessing techniques, models, and their respective hyperparameters [49,50].

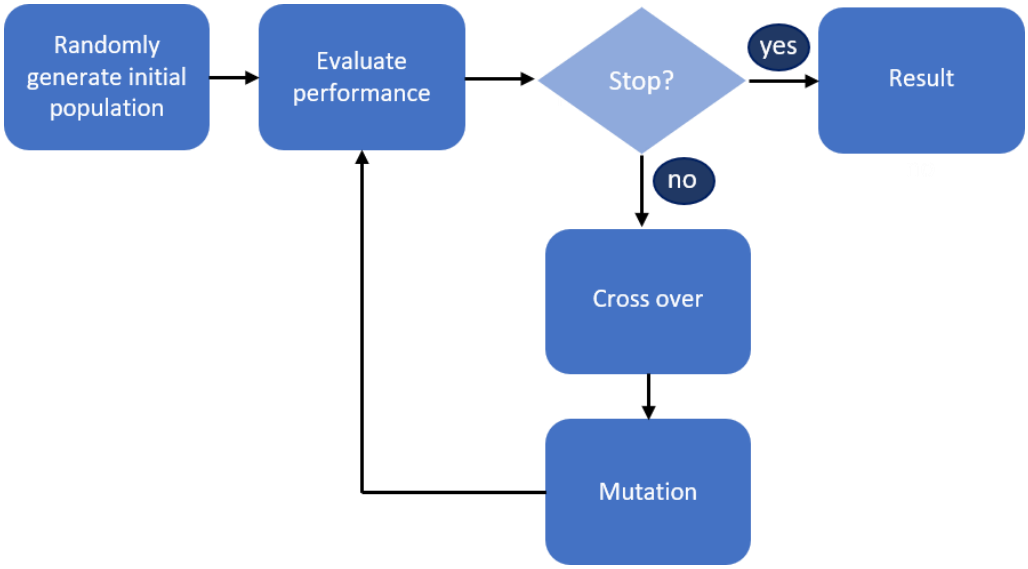

Figure 8 *Genetic programming, an evolutionary algorithm, used by TPOT for optimizing machine learning pipelines.*

2.3.4 Deep learning models

In addition to conventional machine learning approaches, deep learning models can also be used in predicting pCR and performing tasks such as tumor segmentation, feature extraction, and classification. Khan et al. described in a systematic review that various types of deep learning model architectures are used in the prediction of pCR in breast cancer patients, such as convolutional neural networks (CNN) like AlexNet, VGG13, VGGNet, and ResNet-50. These networks achieve an accuracy range of 77.2%-92.3% [51]. CNN is one of the most commonly used deep-learning models in medical imaging [51]. This network utilizes various layers: convolutional layers, pooling layers, and fully connected layers (Figure 9) [52]. The convolutional layer performs feature extraction; the pooling layer downsamples the dimensions in the features maps, reducing the number of features and minimizing the sensitivity to the exact spatial location of those features. The fully connected layers translate the extracted features into an output of the model [52]. In a CNN, it is possible to use different types of data as input,

known as a multiple-input CNN. This approach can incorporate both DCE-MRI from different time points and clinical data [53,54]. This can be achieved by using two parallel comparable sub-architectures for convolution and pooling. Subsequently, a flattened layer and concatenation can be applied before utilizing the fully connected layer [53].

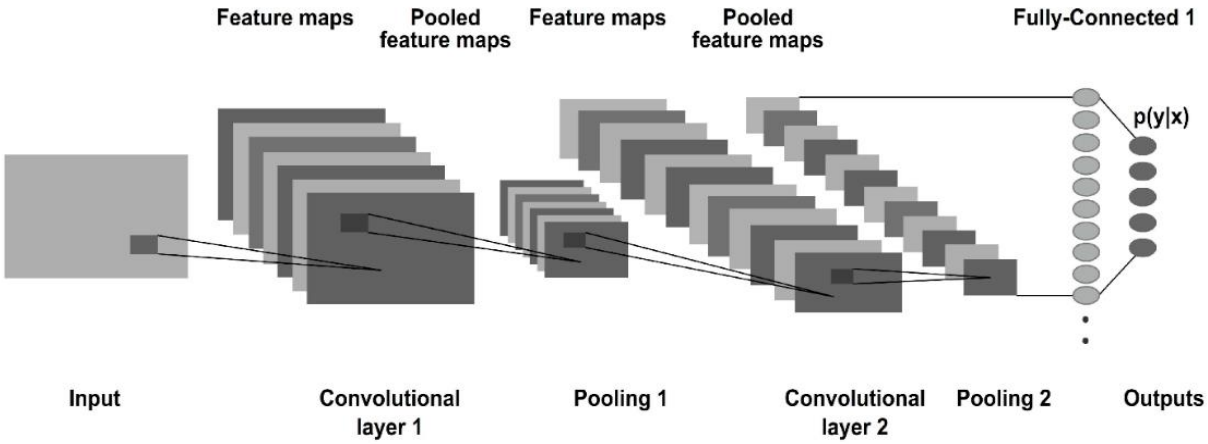

Figure 9 *Convolutional Neural Network. The convolutional layers extract features, the pooling layers downsamples feature maps, reducing the number of features and minimizing spatial sensitivity, while the fully connected layers map the features to the model's output. Figure taken from [52].*

However, medical experts have concerns regarding the black box nature of AI. Additionally, patients have the right to get an understandable explanation of how a decision is made [55]. For these reasons, explainable artificial intelligence (XAI) can be used. This may involve techniques such as the use of a hot map, which highlights pixels relevant to the output of the model [56]. Figure 10 shows an example of a saliency map. This saliency map is initially used for tumor segmentation with the use of a neural network and was generated with the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm. This algorithm examines how the outcome changes when the activations in the network are adjusted. In this way, it can determine which areas in the image have the most influence on the segmentation.

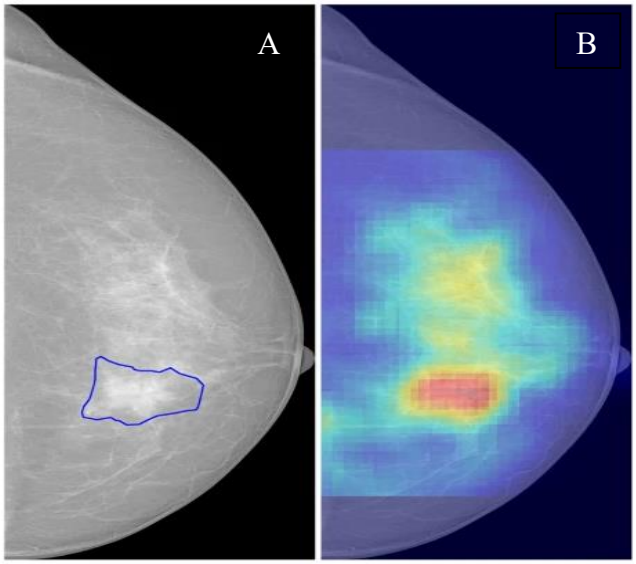

Figure 10 *Saliency analysis. A) mammogram with manually segmented lesion. B) saliency map for tumor segmentation. Figure taken from [56].*

3 DATA COLLECTION AND PREPROCESSING

3.1 Introduction

The development and evaluation of a predictive model for pCR in breast cancer patients depends on both the quantity and quality of the data. It is expected that integrating multiple data sources will enhance the prediction outcome of pCR [57]. For this reason, both clinical-pathological and radiological data will be collected. Radiological variables, such as tumor diameter and volume, can be calculated from segmented tumor lesions. Tumor segmentation can be performed manually using software like Slicer or ITK-SNAP [58,59]. However, these methods are time-consuming and prone to high inter- and intra-observer variability, depending on the tumor type and the observer's experience [60]. Such challenges have prompted the development of neural networks for automatic tumor lesion segmentation, leveraging deep learning architectures like U-Net [61–64].

Given these considerations, this chapter aims to address the following sub-questions:

1. *What data can be obtained within Deventer Hospital to develop a predictive model for pCR in breast cancer patients?*
2. *Which publicly available automatic tumor segmentation tool is suitable for the MRIs acquired at Deventer Hospital?*

3.2 Material and Methods

3.2.1 Study Population

This retrospective single-centre analysis aimed to include all patients with stadium I-III breast cancer receiving neoadjuvant chemotherapy and surgical resection at Deventer Hospital. The exclusion criteria included: (1) patients with either no DCE-MRI of only one DCE-MRI, (2) irrecoverable missing values in clinicopathologic data, and (3) the presence of tumor in both breasts. All clinical- and pathological data was extracted from the electronic health records (HiX, version 6.2), and all radiology data was extracted from the Picture Archiving and Communications System (PACS).

3.2.2 Clinicopathologic Data

For all included patients the following clinicopathologic data was collected, if available:

- ER, PR, and Her2+ receptor status
- Tumor type
- Tumor grade
- TNM stage before treatment
- Age at the time of diagnosis
- Mean length, mean weight, and weight change relative to the starting weight before neoadjuvant chemotherapy
- Menopausal status before neoadjuvant chemotherapy
- ASA score before neoadjuvant chemotherapy
- Number of cycli chemotherapy
- Given cytostatic during neoadjuvant chemotherapy
- Leucocytes, thrombocytes, and hemoglobin before neoadjuvant chemotherapy
- Pathologic response

All the clinicopathologic data were converted to numerical values using label encoding. One-hot encoding was used for categorical variables that could belong to multiple categories within the same variable. This approach transformed these variables into separate binary columns.

The receptor status was determined after biopsy by pathological examination; the removed tissue was identified as positive for ER or PR receptor when >10% tumor staining for ER or PR was seen. Her2 was determined as positive by a 3+ score with immunohistochemistry, possibly confirmed by fluorescence in situ hybridization (FISH).

Most clinical data were manually obtained from electronic health records. For the administered oncologic medicines per patient, raw data was acquired from the pharmacotherapeutic portal, which included all oncological medicines given to each patient. This data was filtered based on the date of the first MRI and the surgery date. The mean height, mean weight, and weight change relative to the starting weight were obtained from the institutional data desk of Deventer Hospital. The raw data were subsequently filtered by the date of the first and last administration of oncologic medicines per patient. Finally, raw data was obtained from the Deventer Hospital institutional data desk regarding hematological laboratory values. The laboratory values before the first administration of oncologic medicines were included in the database.

It was not possible to determine the menopausal status of each patient with the information from the electronic health record before the start of neoadjuvant chemotherapy. For these patients, an estimate of the menopausal status was made using the following criteria [65]:

- < 45 years = pre-menopausal
- 45-55 years = peri-menopausal
- > 55 years = post-menopausal

For each patient, at least a binary pathologic response was available: presence or absence of pCR (1 vs 0). However, the new guideline residual cancer burden (RCB) (2020) allows for the categorization into four categories: RCB-0 (pCR), RCB-I (minimal residual disease), RCB-II (moderate residual disease), and RCB-III (extensive residual disease) [66]. The categorization depends on factors such as the size of the residual tumor bed and the number of positive lymph nodes. Using a newly developed method (Appendix A), a part of the responses was reclassified according to the most recent guideline RCB based on the already existing pathological reports.

3.2.3 Radiological Data

All MRI examinations before, around the midpoint, and after completing neoadjuvant chemotherapy were included. The image acquisition followed the standard clinical protocol of Deventer Hospital. The image data was automatically exported from the PACS system of Deventer Hospital as DICOM files.

3.2.4 Tumor Segmentation

For tumor segmentation in DCE-MRI, two publicly available deep learning networks were tested: the AI assistant tool developed by Zhang et al. [67] and the MAMA-MIA tool [68]. These deep learning segmentation networks were tested in a compute cluster with Graphics Processing Unit (GPU) at the University of Twente.

The online available pre-trained network tool developed by Zhang et al. is based on 13,167 DCE-MRI volumes obtained from seven medical centers. Zhang et al. reported a Dice Similarity Coefficient (DSC) of 0.72 [67]. The Zhang model was built on an Asian population, which tends to have denser breasts compared to the Western population. The model preprocessed the scans automatically; resampling to a voxel size of 1.0x1.0x1.0 mm^3 and performing intensity normalization. The intensity normalization was performed by removing the top 1% of outliers from all image phases. Min-max normalization was applied to phase 2, which typically shows the strongest contrast enhancement. The minimum and maximum values

from phase 2 were used to normalize the other phases, scaling the intensity values between 0 and 1 while preserving temporal intensity changes. The network could handle an arbitrary number of phases.

The MAMA-MIA network is an online available pre-trained no-new U-Net (nnU-Net), trained using 1,506 DCE-MRIs with expert segmentations. The preprocessing steps included resampling to 1.0x1.0x1.0 mm^3 and z-score normalization with the mean and standard deviation of all its phases. These preprocessing steps were not automatically done by the network. The DSC of this model as reported by its authors was 0.70 [68]. In this study, only contrast phase 2 was used further because this phase was available for every patient and provided the clearest distinction of the breast tumor.

For the validation of the two different segmentation networks, a visual inspection was initially performed by a technical physician in training. During the visual inspection, the focus was primarily on whether the model segmented the breast tumor at all and which structures the model over-segmented. Based on this inspection, the segmentations from the model with the best results were evaluated by a radiologist from Deventer Hospital. The radiologist assessed an initial 52 segmentations. These 52 scans were selected using the following criteria: 1) one patient per year and 2) for each year one patient with and one patient without pCR. The segmentation assessment was facilitated by a segmentation review module [69] in Slicer [70]. This review made it possible to score the segmentation between 1 (acceptable, no changes) and 5 (bad image). In this segmentation review module, it was also possible to manually adjust the segmentation, resulting in a ground truth segmentation by the radiologist. This ground truth segmentation was used to compute a DSC score and Hausdorff distance (HSD) between the radiologist and the segmentation network. The DSC is a statistical measure that quantifies the overlap between two segmentations masks, and the HSD is another metric used to evaluate the spatial accuracy of the segmentation boundaries. The outcome of these scores determines if manual adjustments to the tumor segmentations are necessary. In consultation with the radiologist, the automatic segmentation model had to achieve a DSC score of at least 0.80; otherwise, a manual review and correction of every segmentation would be performed. Beside this, the volume of the radiologist's segmentations is plotted against the volume of the segmentations from the Zhang network, the MAMA-MIA network, and the manually adjusted MAMA-MIA network. The Pearson correlation coefficient (r) was also calculated to determine the linear correlation.

3.3 Results

3.3.1 Patient characteristics

This retrospective single-centre analysis initially included 331 patients diagnosed with stage I - III breast cancer at Deventer Hospital between June 2005 and August 2023. A total of 291 patients with complete clinicopathological records and at least two DCE MRI scans were finally included in this study (Figure 11). For 26 patients, the menopausal status was estimated based on age. The patient characteristics for the study population are listed in Table 2. For 187 patients, it was possible to reclassify the binary response (pCR and no pCR) to RCB-score (RCB-0, RCB-1, RCB-2, and RCB-3).

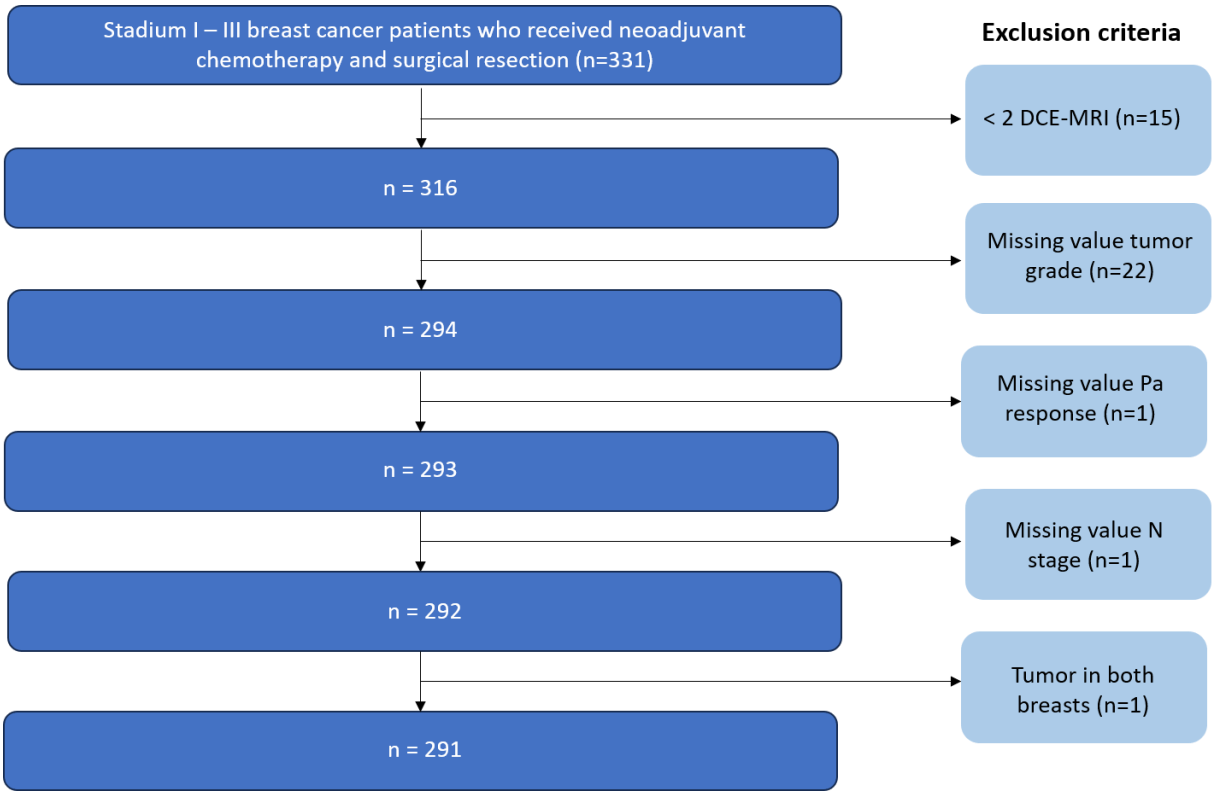

Figure 11 *The cohort consisted of 331 patients diagnosed with breast cancer grade I - III at Deventer Hospital between 2005 and 2023. Only patients with complete clinical-pathological data, at least two DCE-MRI scans and a single-breast tumor were included in the study.*

Table 2 *Patient characteristics. y = years, n = number, cm = centimeter, kg = kilogram, nL = nanoliter, mmol/L = mmol per liter*

Characteristics	Value
Age at the time of diagnosis (y)	
Median	51
Min – Max	27 - 79
ER Receptor Status, n (%)	
Positive	180 (62%)
Negative	111 (38%)
PR Receptor Status, n (%)	
Positive	131 (45%)
Negative	160 (55%)
Her2+ Receptor Status, n (%)	
Positive	70 (24%)
Negative	221 (76%)
Tumor Type, n (%)	
No Special Type	163 (56%)
Invasive Ductal Carcinoma	81 (28%)
Invasive Lobular Carcinoma	30 (10%)
Other	17 (6%)
Tumor Grade, n (%)	
I	25 (8%)
II	139 (48%)
III	127 (44%)
T stage before Neoadjuvant Chemotherapy, n (%)	
T1	28 (10%)
T2	164 (56%)
T3	90 (31%)
T4	8 (3%)

N stage before Neoadjuvant Chemotherapy, n (%)	
N0	132 (45%)
N1	127 (44%)
N2	12 (4%)
N3	20 (7%)
Menopausal Status before Neoadjuvant Chemotherapy, n (%)	
Pre-menopausal	123 (42%)
Peri-menopausal	41 (14%)
Post-menopausal	127 (44%)
ASA score before Neoadjuvant Chemotherapy, n (%)	
1	107 (37%)
2	162 (56%)
3	22 (7%)
Mean Length During Neoadjuvant Chemotherapy (cm)	
Median	170
Min – Max	150-183
Start Weight (kg)	
Median	73
Min – Max	48-130
Mean Weight Change Relative to Start Weight before Neoadjuvant Chemotherapy (kg)	
Median	0.0
Min – Max	-20 - 18
Number of chemotherapy cycli (n)	
Median	13
Min – Max	3 – 26
Given oncologic medicines, n (%)	
Atezolizumab	4 (1%)
Bevacizumab	4 (1%)
Carboplatin	119 (41%)
Cyclophosphamide	228 (78%)
Docetaxel	81 (27%)
Doxorubicin	222 (76%)
Epirubicin	5 (2%)
Gemcitabin	1 (0.3%)
Paclitaxel	207 (71%)
Pertuzumab	52 (18%)
Transtuzumab	71 (25%)
Vinorelbine	4 (1%)
Leucocytes before Neoadjuvant Chemotherapy (/nL)	
Median	7.5
Min – Max	3.2 – 19
Thrombocytes before Neoadjuvant Chemotherapy (/nL)	
Median	277
Min – Max	151- 548
Hemoglobin before Neoadjuvant Chemotherapy (mmol/L)	
Median	8.5
Min – Max	5.1 – 10

3.3.2 Radiological data

Images were acquired on either GE Healthcare or Philips MRIs with a field strength of 1.5 or 3 Tesla. For the DCE images, the gadolinium-based contrast agents Omniscan (GE Healthcare) or Clariscan (GE Healthcare) were used. For 33/291 (11%) patients, two DCE MRIs were available instead of three. The number of post-contrast phases following the intravenous injection of the contrast agent ranged from four to eight. The acquisition parameters of the MRIs are listed in Table 3.

Table 3 *Acquisition parameters. All values are described in the following way: mode (frequency).*

Parameter	DCE
Pixel spacing (mm)	[0.70, 0.70] mm (68%)
Slice thickness (mm)	2.2 mm (67%)
Field strength	1.5 (82%)
Echo time (s)	2.25 ms (50%)
Repetition time (s)	4.77 ms (42%)
Flip angle	10 ° (100%)

3.3.3 Tumor Segmentation – Visual Inspection

3.3.3.1 Zhang Network

Figure 12 visualizes the tumor segmentation results for two cases using the Zhang network. In Figure 12A, the tumor in case 1 is indicated within the red circle. In Figure 12B, the tumor in case 1 is segmented adequately by the network, although a small area at the dorsal margin has been missed (see blue arrow in Figure 12B). Additionally, there appeared to be an over-segmentation on the lateral side of the tumor (see orange arrow in Figure 12B). In Figure 12C, the tumor in case 2 is encircled with red. Conversely, Figure 12D demonstrates case 2, where the algorithm failed to segment the tumor and instead segmented a part of the pectoralis muscle (see yellow arrow in Figure 12D). The Zhang network was more likely to miss a tumor in the segmentation compared to the segmentation with the MAMA-MIA network. Additionally, in nearly every segmentation by the Zhang network, a nipple, breast contour, or part of the chest wall was segmented as the tumor.

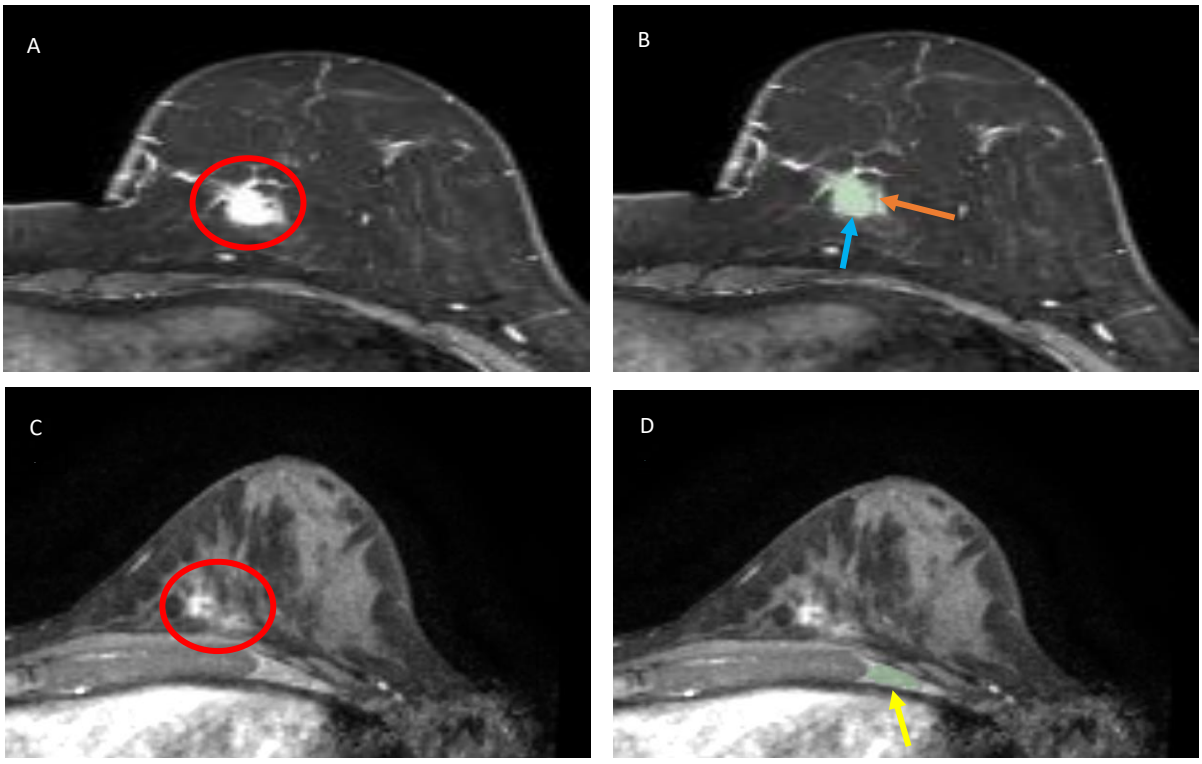

Figure 12 *Tumor segmentation by Zhang Network. A) Case 1 with red circle around tumor B) Case 1 with tumor segmentation by the network in green, blue arrow indicates under segmentation and orange arrow indicates over segmentation by the network C) Case 2 with red circle around tumor D) Case 2 with segmentation by the network in green (pectoralis muscle instead of tumor).*

3.3.3.2 MAMA-MIA Network

Figure 13 presents the tumor segmentation results for two cases using the MAMA-MIA network. In Figure 13A, the tumor in case 1 is indicated within the red circle. Figure 13B shows that the tumor in case 1, which was not segmented by the Zhang network, was successfully segmented by the MAMA-MIA network. However, it appears that the MAMA-MIA network slightly over-segmented towards the ventral side (see blue arrow in Figure 13B). Figure 13C shows the tumor in case 2 encircled in red. Figure 13D shows an accurate segmentation of the tumor in case 2. However, this case also showed the segmentation of normal breast tissue in the contralateral breast by the network (see orange arrow in Figure 13D).

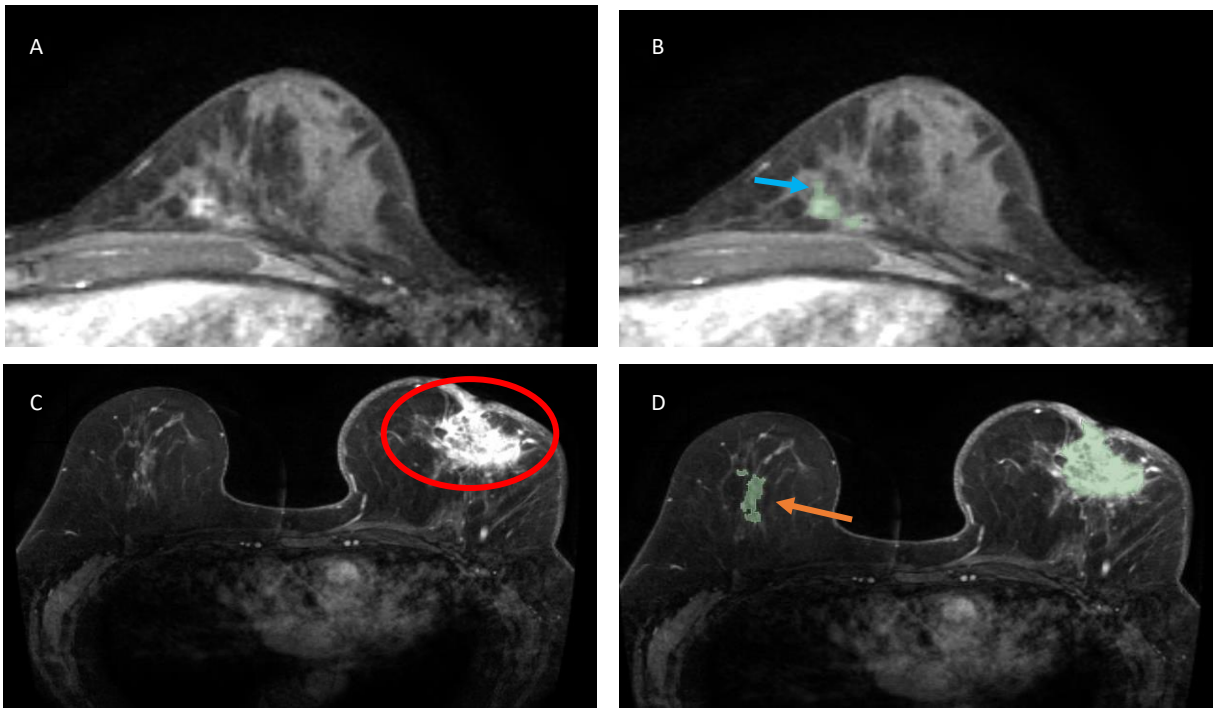

Figure 13 *Tumor segmentation by MAMA-MIA Netwerk. A) Case 1 with red circle around tumor B) Case 1 with tumor segmentation by the network in green, blue arrow indicates over segmentation by the network C) Case 2 with red circle around tumor D) Case 2 with segmentation by the network in green, orange arrow indicates over segmentation in contra lateral breast.*

3.3.4 Evaluation of Tumor Segmentation by Radiologist

During the segmentation review by the radiologist, it turned out that the pre-selection of the DCE-MRIs was not executed properly, leading to some patients' MRIs being included after postoperative chemotherapy. As a result, the radiologist evaluated a total of 42 segmentations from the MAMA-MIA network in Slicer. The results are presented in Figure 14. The radiologist assessed 24/42 (57%) segmentations as acceptable, which means no or only minor adjustments were required. The remaining 16 segmentations were scored as unacceptable, requiring significant changes in the tumor segmentation by the radiologist.

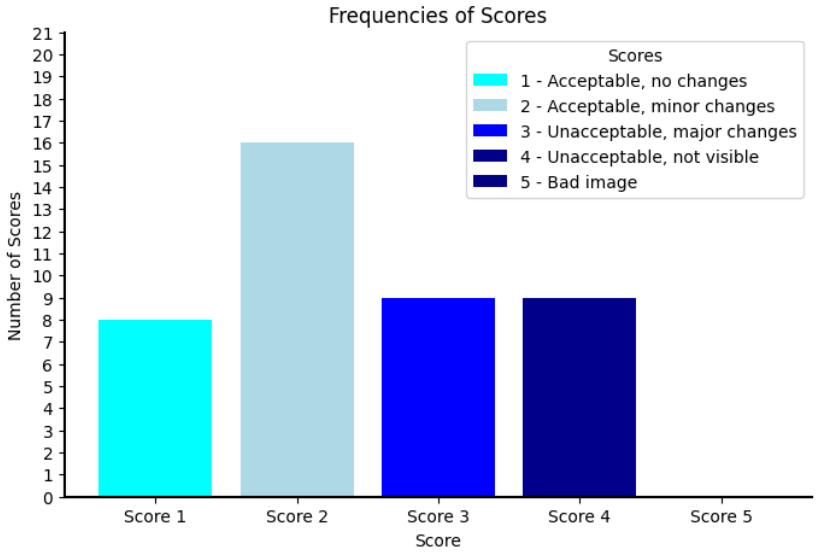

Figure 14 *Review of segmentations by radiologist as assessed on a 5-point likert scale.*

The DSC between the ground truth from the radiologist and the segmentations made by the Zhang network was 0.27, excluding 10 empty ground truth segmentations (these patients were classified as having a pCR). In these pCR cases, 10% of the network's segmentations were empty. The HSD between the radiologist and the Zhang network was 173 mm. On the other hand, the DSC between the ground truth from the radiologist and the MAMA-MIA network was 0.69, excluding the same 10 empty ground truth segmentations. In these empty cases, no segmentations from the network were empty. The HSD between the radiologist and the MAMA-MIA network was 129 mm.

Based on these results, the segmentations from the MAMA-MIA network were manually adjusted by a technical physician in training, leading to an DSC of 0.85 between the radiologist's ground truth segmentation and the manually adjusted segmentation from the MAMA-MIA network. In 10 empty ground thruth segmentations, 50% of the segmentations that were manually adjusted, were also empty. The HSD between the radiologist and the manually adjusted MAMA-MIA network segmentations was 39 mm. Table 4 lists these results.

Table 4 *Segmentation review by radiologist. DSC = Dice Similarity Coefficient, HSD = Hausdorff Distance.*

Network	DSC	HSD (mm)
Zhang	0.27	173
MAMA-MIA	0.69	129
Manually adjusted MAMA-MIA	0.85	39

Figure 15 presents the volume correlation curve, showing the segmentation volumes from the radiologist plotted against the segmentation volumes from the Zhang network, the MAMA-MIA network, and the manually adjusted segmentations of the MAMA-MIA network. Figure 15A showed that the Zhang network produced most of the time over-segmentation ($r = 0.59$). This over-segmentation was also observed in the segmentations from the MAMA-MIA network, with a correlation of $r = 0.64$ (Figure 15B). Figure 15C revealed a linear relationship between the radiologist's segmentations and the manually adjusted segmentations from the MAMA-MIA network ($r = 0.91$). In this case, six over-segmentations and one under-segmentation were noted.

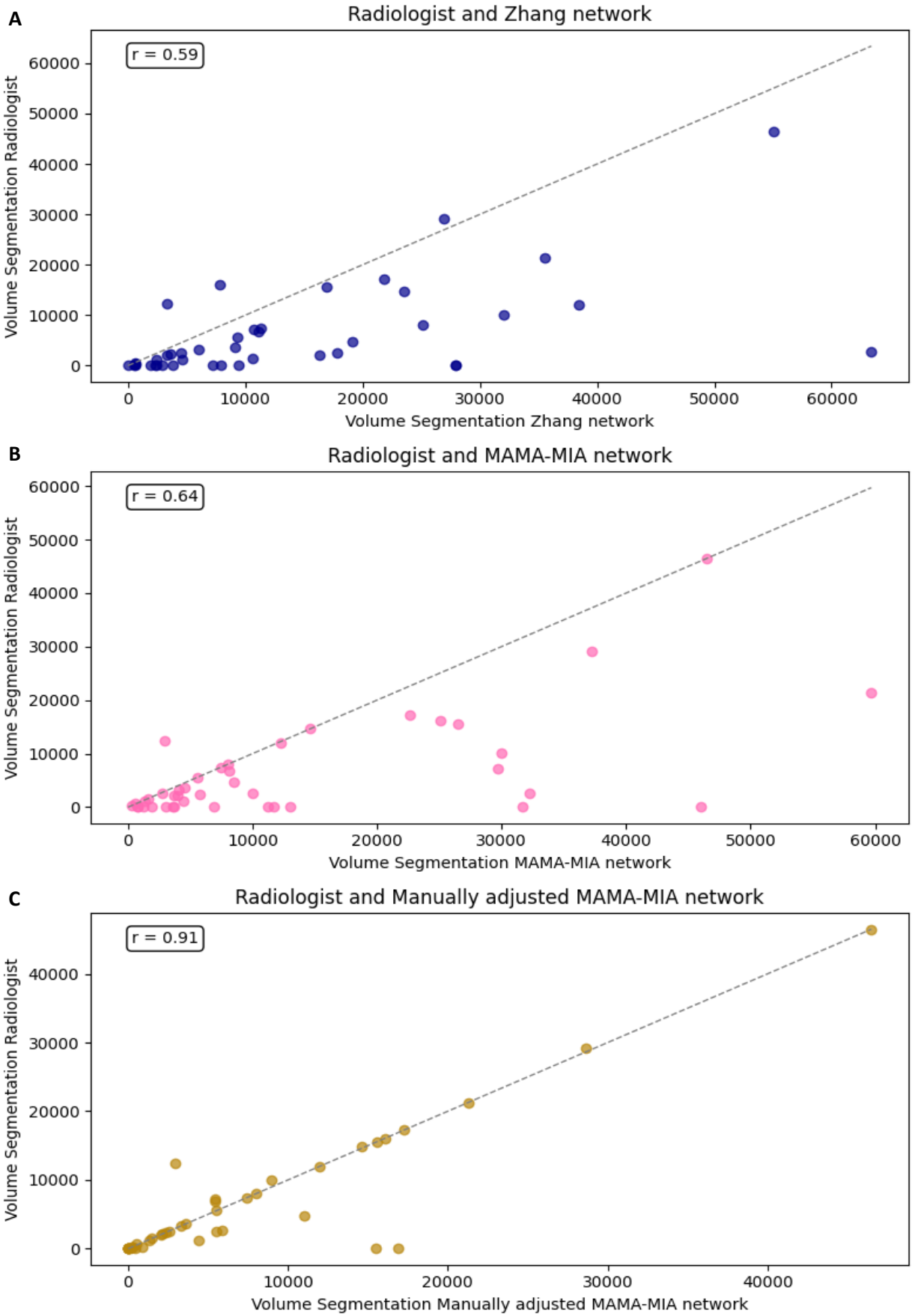

Figure 15 *Volume correlation curve between A) Radiologist and Zhang Network, B) Radiologist and MAMA-MIA network, and C) Radiologist and manually adjusted MAMA-MIA network. r = Pearson correlation coefficient.*

3.4 Discussion

3.4.1 Clinicopathologic data

A total of 18 clinicopathological variables were included in this study. However, other clinicopathological variables could also have an impact on predicting pCR in breast cancer patients undergoing neoadjuvant chemotherapy. For instance, a high level of Ki-67 correlates with a worse prognosis and a lower chance of pCR [34,35]. Additionally, sixty of the 123 microRNAs appear to be associated with prognosis and neoadjuvant response [22]. However, these variables could not be retrospectively determined from the Deventer Hospital data. Furthermore, data regarding the patient's blood profile was collected. Not all this data was included in this study because it was difficult to determine the duration and number of chemotherapy cycles for each oncological treatment. The blood profile is determined for each patient before each chemotherapy cycle, blood test are typically not conducted after the final cycle. If the start date of the last cycle for each patient can be determined, it will be possible to match the blood values to these cycles. This would enable the analysis of changes in the blood profile over time, which could potentially serve as a predictive parameter for pCR. To analyse the relationship between blood values over time and the prediction of pCR, a mixed effects model could be applied, for instance. Additionally, it may be interesting to investigate whether a dose reduction impacts the likelihood of pCR. A dose reduction is typically performed when the blood profile for example shows neutropenia or thrombocytopenia. In such cases, it might be that the chemotherapy is more aggressive in the body, causing damage to healthy cells as well. These patients would always receive a dose reduction because the patient's well-being takes priority over the outcome of the treatment. However, the impact of dose reduction on tumor response is not well understood in the literature, and it is uncertain whether dose reduction affects the likelihood of achieving a pCR after treatment. For instance, it could also be the case that a dose reduction, in combination with the other clinical-pathological variables, does not affect the chance of pCR, meaning that a dose reduction could be applied to this group of patients regardless. A dose reduction can help reduce the side effects of oncological treatments, which is beneficial for the patient's quality of life. However, it was hard to determine in the short term which dose belonged to each oncological treatment, making it impossible to calculate whether a patient received a dose reduction. With the already received data, the expertise of Deventer Hospital, and some extra time, it seems possible to determine the dose reduction per patient retrospectively.

3.4.2 Radiological data

In 33 patients, one DCE-MRI sequence is missing, which means that these patients only have two MRIs during the neoadjuvant chemotherapy. This may be due to a different scanning protocol used for these patients, primarily among those from 2005, 2006, or 2007. This scanning protocol does not include a DCE sequence. Another reason for the missing DCE-MRI sequences lies in the fact that for some patients the second MRI was performed shortly before the final chemotherapy cycle. In these cases, a third MRI is not conducted as it would be too close to the previous one. Additionally, the number of chemotherapy cycles varies per patient. The number of chemotherapy cycles per patient depends on the breast cancer subtype and the applicable guidelines at that time. Additionally, patients may participate in studies which can lead to variation in the number of chemotherapy cycles. In the TRAIN-3 study, for example, the number of chemotherapy cycles was determined based on the radiological response. When a radiological response was observed, the neoadjuvant treatment was discontinued, resulting in a minimum of three neoadjuvant chemotherapy cycles [20]. Furthermore, the development of toxicity during the neoadjuvant treatment can also influence the number of chemotherapy

cycles. Consequently, patients who underwent fewer chemotherapy cycles also had fewer MRIs.

3.4.3 Segmentation networks

Currently, the gold standard for tumor segmentation involves manual delineation by experienced radiologists [71]. This method is often associated with a high inter-observer variability and tends to be time intensive. In response to these challenges, various types of automatic deep learning segmentation models have been developed for breast tumor segmentation, with a DSC ranging from 0.61-0.98 [62,63,72]. Such models are trained to recognize patterns and features indicative of tumor presence, potentially providing faster and more consistent results compared to human experts. A notable drawback of the deep learning networks described in the literature is their limited availability, often due to financial constraints and privacy concerns regarding patients. Besides this, prior literature highlights the ongoing complexity of breast tumor segmentation, particularly on the DCE-MRI sequence. One significant challenge arises from the fact that contrast enhancement can occur not only within tumor regions but also in non-tumor tissues such as normal breast tissue and vessels [73]. This non-specific enhancement complicates the identification of tumor, especially when this tissue is situated adjacent to the tumor. Moreover, breast tumors exhibit considerable variability in their morphology and size; the tumor can be irregular in shape and can be heterogeneous in texture [62]. This variability complicates the segmentation task, as automated systems must be robust enough to handle a wide range of tumor characteristics.

The Zhang network, when applied to Deventer Hospital data, does not meet the mentioned DSC range in the literature of 0.61 – 0.98 (DSC Zhang network 0.27). The visual inspection of the Zhang network's segmentation results revealed that the network often failed to segment clearly visible tumors. Additionally, in almost every segmentation, the nipple, breast contour, or part of the chest wall was erroneously included. It was expected that these structures would be segmented by the Zhang network because these structures appeared hyperintense on the DCE-MRI. Due to this erroneous over-segmentation, the radiologist did not evaluate the Zhang network’s segmentations. The segmentation network has learned associations and features based on the training data. It is expected that the data from Deventer Hospital differs too much from this training data, preventing the network from effectively using the learned associations and features for tumor segmentation. This could be due to population differences, as the Zhang model was developed for an Asian population, which tends to have denser breast tissue compared to the Western population [67]. Furthermore, differences in imaging protocols could also contribute to this discrepancy, as the Zhang network was trained with a maximum of 6 contrast phases, whereas Deventer Hospital's imaging data includes up to 9 contrast phases. Since the radiologist's evaluation was performed on the segmentations from the MAMA-MIA network, a ground truth segmentation was obtained. This ground truth was used to calculate the DSC and HSD between the radiologist and the Zhang network (Table 4).

In contrast to the Zhang network, both the MAMA-MIA network and the MAMA-MIA network with manual adjustments, applied to Deventer Hospital data, achieve DSC values within the mentioned range of 0.61-0.98 in the literature (DSC MAMA-MIA network 0.69, DSC MAMA-MIA network with manual adjustments 0.85). The radiologist evaluated the results of the segmentations performed by the MAMA-MIA network. Initially, the radiologist was supposed to assess 52 segmentations, with one patient having a pCR and another not, from each year. However, it turned out that the pre-selection of the DCE-MRIs was not executed properly, leading to some patients’ MRIs being included after postoperative chemotherapy. As a result, only 42 correct segmentations were evaluated. From the radiologist's evaluation, it was found

that 57% of the segmentations were acceptable (Figure 14). It was particularly notable that non-tumor tissue in the contralateral breast was often included in the segmentation. Additionally, in cases of pCR, some tissue was also segmented. This can be explained by the fact that the training data always included tumor tissue, leading to a segmentation in every scan. For these reasons, the segmentations were manually corrected by a technical physician in training. This was done using the radiological reports, which highlighted the subjectivity in how radiologists classify breast tumors, depending on the radiologist's experience. Moreover, the radiologist cannot classify individual tumor cells on a DCE-MRI either. In addition, the manual adjustments of the tumor segmentations were performed by a trainee technical physician, not by a radiologist, so it cannot be expected that the segmentations meet the golden standard. As shown in Figure 15C, the comparison between the segmentations made by the radiologist and the trainee technical physician revealed six instances of over-segmentation and one instance of under-segmentation.

3.4.4 Clinical relevance

The findings emphasize that, while the Zhang network and especially the MAMA-MIA network show potential, they are not yet reliable for breast cancer segmentation in Deventer Hospital's DCE-MRIs without human oversight. The MAMA-MIA network is the closest to being applicable in both clinical and research settings, although the need for manual corrections still limits its efficiency and practicality. Given that it is uncertain whether online available networks are robust enough to perform effectively on data from Deventer Hospital, it may be worthwhile to consider retraining the available segmentation networks with the Deventer data. The segmentation networks could be retrained using the ground truth established by radiologists during the evaluation of the segmentation networks.

3.5 Conclusion

In this study, a total of 291 patients with complete clinicopathological records and with at least two DCE MRI scans were included. Examples of these clinicopathological variables include age, receptor status, menopausal status, ASA score, and given oncologic medicines. In addition to this clinical-pathological data, there was most of the time also access to three DCE-MRI scans per patient during the neoadjuvant chemotherapy process. Using these MRIs, this study also assessed the performance of two segmentation networks —the Zhang et al. network and the MAMA-MIA network—in detecting and segmenting breast tumors in DCE-MRI sequences. The results suggest that the Zhang et al. network was not accurate enough to consistently and reliably segment breast tumors in the Deventer Hospital data, often resulting in missed tumors (DSC 0.27, HSD 173 mm). The MAMA-MIA network, however, demonstrated a better performance for tumor segmentation (DSC 0.69, HSD 129 mm). Nevertheless, manual adjustments were necessary because the network tends to segment tissue in the contralateral breast. Further optimization will still be required for the clinical application of automatic tumor segmentation.

4 MACHINE LEARNING BASED PREDICTION OF PCR

4.1 Introduction

The development of a predictive model for pCR in breast cancer patients using machine learning can include a wide array of methodologies and approaches. A key element in this process is the configuration of the dataset. Therefore, effectively addressing missing values [74] and making decisions about data preprocessing—such as normalization and class balancing [75] —are crucial for success. Alongside dataset management, various machine learning models can be trained, ranging from simple regression techniques to more complex algorithms such as decision trees, random forests, and support vector machines. However, the manual evaluation and fine-tuning of these models can be both time-consuming and resource intensive. To streamline this process, automated tools like tree-based pipeline optimization tool (TPOT) can be utilized. TPOT leverages genetic programming to optimize model selection and hyperparameter tuning [49]. Given the diverse options available for dataset settings and machine learning models, this chapter aims to address the following sub-question: *Which dataset configuration and associated machine learning model achieve the best outcome for the prediction of pCR in breast cancer patients?*

4.2 Material and Methods

4.2.1 Study Population

Initially, this retrospective single-centre study used the same inclusion and exclusion criteria described in Section *3.2.1.* For this study, the following additional exclusion criterion was applied: < 3 DCE-MRI.

4.2.2 Feature Extraction

The included 18 clinical-pathological features were listed in Section *3.2.2*. For each included patient, the image data from the DCE-MRIs were exported from the PACS system of Deventer Hospital as DICOM files. For every included patient, three MRIs were available: MRI1, taken before the start of neoadjuvant chemotherapy; MRI2, taken during neoadjuvant chemotherapy; and MRI3, taken after neoadjuvant chemotherapy but before surgical tumor removal. The tumor was segmented in the DCE-MRI using the MAMA-MIA network. Each segmentation performed by the deep learning network was manually checked and, if necessary, corrected by a technical physician in training. Using the Python module PyRadiomics, 14 3D shape features were extracted from each tumor segmentation (Appendix B). During the manual segmentation review, a part of the sternum at the level of rib 5-6 was also manually segmented. The segmentation of the sternum was used to calculate the ratio in intensity between the tumor and the sternum, which was included as one intensity feature (Equation 5). When three MRIs were used for the classification model, the delta-radiomics for each feature were calculated by subtracting the feature value at MRI moment 1 from the value at MRI moment 3 (MRI1 – MRI3). When only two MRIs were used for the classification model, the delta-radiomics were calculated by subtracting the value of the first available MRI moment from the value of the second available MRI moment.

$$IR = \frac{M1}{M2} \tag{5}$$

Equation 5 IR: Intensity Ratio, M1: mean of tumor mask, M2: mean of reference mask

4.2.3 Statistical Analysis

The statistical analysis was based on the variables before the start of the neoadjuvant chemotherapy. The Spearman Correlation Coefficient (SCC) was calculated to assess the strength and direction of the monotonic relationship between pairs of variables. Only variables with a SCC between -0.7 and 0.7 were included in the statistical analysis [76]. The normal distribution of the variables was assessed using the Shapiro-Wilk test for continuous variables. Subsequently, the Mann-Whitney U Test was used for the normally distributed continuous variables and the Student's T-Test for the non-normally distributed continuous variables. For the categorical variables, Fisher's exact test or Chi-squared Test was used, based on the frequency of each category in the cross table [77]. The minimum frequency threshold for each category was set at five. The Chi-squared test was used when the frequency in each cell was five or more, while Fisher's exact test was applied when the expected frequency in one or more cells was less than five [78]. In the univariate analysis, correction for multiple testing was performed using the Bonferroni method. Besides this, a multivariate analysis was conducted using multivariate logistic regression with a 5-fold cross-validation. For this purpose, the dataset was divided into training and test sets using an 80/20 ratio. Stratification was applied during this process, based on the pathological response (pCR or no pCR) and the date of the first DCE MRI (based on 3-year intervals). In the logistic regression, the class weight was set to balanced, which automatically adjusts weights for each class inversely proportional to their frequencies in the data. The logistic regression was evaluated with the sensitivity, specificity, and receiver operating characteristics (ROC) curve with AUC. For all statistical analyses, a p-value smaller than 0.05 was considered statistically significant. The statistical analyses were carried out in Python 3.7 with SciPy and Statsmodels packages.

4.2.4 Model Development and Validation

For the development of the prediction model, the data for each model was divided into training and test sets using an 80/20 ratio. Stratification was applied based on the pathological response and the date of the first available MRI (based on 3-year intervals). Various models were trained using TPOT (Python 3.9.19), including binary classification (pCR vs no pCR) and multi-class RCB classification (RCB-0 vs. RCB-1 vs. RCB-2 vs. RCB-3). For this purpose, the compute cluster with GPUs from the University of Twente was used. In all analyses, the default parameters of TPOT were used.

Different dataset configurations were applied for predicting both binary outcomes and multi-class RCB outcomes. In model 1 and model 5 through model 9, three MRIs per patient were included (MRI1, MRI2, and MRI3). Model 5 normalized the radiological features, which ensured standardization of the values within a consistent range. This approach was expected to enhance the performance of the machine learning algorithms, as it prevented variables with larger scales from having an undue influence on the model and enhanced the model's ability to learn patterns in the data. In model 6 through model 9, class balancing is performed. Class balancing was expected to help prevent overfitting on the majority class. In model 6, class balancing was performed using SMOTE, which added synthetic samples to the minority class. However, SMOTE relies on interpolation, which can create synthetic samples that closely resemble the original data. As a result, the model may have performed poorly on new data. For this reason, class balancing was also performed by random undersampling (model 8). Nevertheless, random undersampeling could increase the risk of losing important information or patterns in the data. Additionally, the removal of samples could resulted in the data failing to accurately represent the original distribution. In models 7 and 9, the effect of combining normalization with different class balancing methods was examined. Furthermore, the literature does not clarify whether all three MRIs taken during the chemotherapy trajectory are necessary

for predicting pCR in every breast cancer subgroup. For this reason, in models 2 through 4, only radiological features from two specific MRI time points were included (MRI1 and MRI2, MRI1 and MRI3, or MRI2 and MRI3).

All models were trained and tested using clinical data, radiological data, and a combination of both. The best-performing dataset configurations were visualized with an ROC curve for the binary classification models (pCR vs no pCR). Model evaluation was based on sensitivity, specificity, and AUC. Binary classification models were tested on the entire dataset and within the three breast cancer subgroups: (1) HR+, (2) Her2+, and (3) TNBC. For the multi-class RCB classification models, evaluation metrics included accuracy and Cohen's kappa. In the case of RCB-0 classification, the data was restructured into a binary one-vs-all format (RCB-0 vs RCB-1, RCB-2, RCB-3). The best-performing configuration for RCB-0 was also visualized with an ROC curve and evaluated using sensitivity, specificity, and AUC. The optimal model for both binary classification (pCR vs no pCR) and RCB-0 classification was selected based on a trade-off between sensitivity and specificity.

4.3 Results

4.3.1 Study Population

This retrospective single-centre analysis included the same 291 patients with clinical-pathological variables described in Section *3.3.1.* Various imputation methods were explored to estimate missing radiological features (n=33) of the patients with < 3 DCE-MRI (Appendix C). These methods proved insufficiently reliable and is not used in this study. Consequently, the extra exclusion criteria resulted in a cohort of 258 patients.

4.3.2 Statistical Analysis

The statistical analyses were conducted on the dataset after removing all patients with missing values in either clinical-pathological or radiological features.

4.3.2.1 Clinical Data

The SCC of the clinical data is presented in a heatmap in Appendix D. The highest correlation was found between doxorubicin and trastuzumab (-0.94) and between doxorubicin and cyclophosphamide (0.94). Following this, there was a correlation of 0.93 between the presence of the Her2+ receptor and trastuzumab. All the other clinical variables known before neoadjuvant chemotherapy had an SCC between -0.7 and 0.7.

From the univariate analysis, the following significant values emerge: ER receptor, PR receptor, Her2 receptor, tumor type, and tumor grade. These features show an odds ratio in the multivariate logistic regression of 0.13, 0.81, 19.8, 0.92, and 2.5, respectively (Table 5).

Table 5 *Statistical analysis of clinical features. OR = odds ratio, CI = confidence interval.*

Characteristics	Univariate analysis	Multivariate analysis	
	P value	OR	CI(95%)
ER receptor	*< 0.001*	0.13	0.035, 0.46
PR receptor	*< 0.001*	0.81	0.24, 2.8
Her2 receptor	*< 0.001*	19.8	6.9, 56.8
Age	1.0	1.0	0.95, 1.1
Tumor Type	*< 0.001*	0.92	0.77, 1.1
Tumor Grade	*< 0.001*	2.5	1.1, 5.7
T before therapy	0.8257	1.4	0.9, 2.3
N before therapy	0.2903	0.97	0.6, 1.5
Menopausal state	1.0	0.89	0.4, 1.9
ASA score	1.0	0.97	0.44, 2.1
Mean length	1.0	0.98	0.92, 1.0
Start weight	1.0	0.99	0.95, 1,0
Leucocytes	1.0	1.0	0.82, 1.2
Thrombocytes	1.0	0.99	0.99, 1.0
Hemoglobin	1.0	0.73	0.42, 1.3

The sensitivity and specificity in the logistic regression for the clinical features were 0.63 and 0.75, respectively. The ROC curve showed an AUC of 0.77 (Figure 16).

4.3.2.2 Radiological Data

The SCC of the radiological data is presented in a heatmap in Appendix E. The highest correlation was logically observed between the voxel volume and mesh volume, with a coefficient of 0.99. This correlation was also found between the delta radiomics voxel volume and delta radiomics mesh volume. Additionally, an SCC of 0.99 was noted between the major axis length and maximum 3D diameter at MRI time point 3. Furthermore, there was a correlation of 0.98 between the voxel volume and surface area at MRI time point 3. Of the radiological variables known before neoadjuvant chemotherapy, only elongation, least axis length, and surface volume ratio had an SCC between -0.7 and 0.7.

From the univariate analysis, the following significant values emerge: elongation and surface volume ratio. In the multivariate logistic regression, the elongation and surface volume ratio show odds ratios of 8.7 and 0.072, respectively (Table 6).

Table 6 *Statistical analysis of radiological features. OR = odds ratio, CI = confidence interval.*

Characteristics	Univariate analysis	Multivariate analysis	
	P value	OR	CI(95%)
Elongation	*0.0001*	8.7	1.1, 67
Least axis length	0.92	0.99	0.96, 1.0
Surface volume ratio	*0.0089*	0.072	0.0061, 0.85
Ratio tumor sternum	0.8282	0.89	0.71, 1.1

The sensitivity and specificity in the logistic regression for the radiological features were 0.69 and 0.44, respectively. The ROC curve showed an AUC of 0.55 (Figure 16).

4.3.2.3 Clinical- and Radiological Data

The heatmap in Appendix F illustrates the SCC between clinical and radiological features. The strongest correlation was observed between the surface area measured on MRI at timepoint 1 and the T stage before chemotherapy, with a coefficient of 0.59. Additionally, a correlation of 0.58 was found between the maximum 2D diameter measured on MRI at timepoint 1 and the T stage before chemotherapy. The correlations between each clinical- and radiological feature assessed, before neoadjuvant chemotherapy, ranged from -0.7 to 0.7. The odds ratios for these clinical and radiological features derived from the multivariate logistic regression are present in Table 7.

Table 7 *Statistical analysis of clinical and radiological features. OR = odds ratio, CI = confidence interval.*

Characteristics	**Multivariate analysis**	
	OR	**CI(95%)**
ER receptor	0.19	0.048, 0.73
PR receptor	0.53	0.14, 2.0
Her2 receptor	21	6.9, 61
Age	1.0	0.94, 1.1
Tumor Type	0.98	0.80, 1.2
Tumor Grade	2.4	0.97, 5.9
T before therapy	1.8	1.0, 3.2
N before therapy	1.1	0.70, 1.8
Menopausal state	1.2	0.52, 2.6
ASA score	1.0	0.44, 2.4
Mean length	0.98	0.91, 1.0
Start weight	1.0	0.96, 1.0
Leucocytes	0.99	0.81, 1.2
Thrombocytes	0.99	0.99, 1.0
Hemoglobin	0.71	0.39, 1.3
Elongation	29	1.9, 449
Least axis length	0.96	0.91, 1.0
Surface volume ratio	0.31	0.0078, 0.89
Ratio tumor sternum	0.85	0.59, 1.2

The sensitivity and specificity in the logistic regression for the clinical features in combination with the radiological features were 0.63 and 0.75, respectively. The ROC curve showed an AUC of 0.73 (Figure 16).

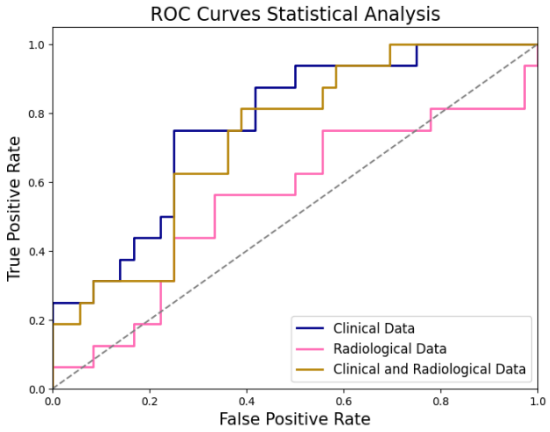

Figure 16 *ROC Curve clinical data, radiological data, clinical and radiological data.*

4.3.3 Model Development and Validation

4.3.3.1 Model Development

Different dataset configurations were applied for predicting both binary outcomes (Table 8) and RCB outcomes (Table 9).

Table 8 *Data configuration for different prediction models with binary outcome (pCR vs no pCR). MRI1 was obtained before neoadjuvant chemotherapy, MRI2 was obtained during neoadjuvant chemotherapy, and MRI3 was obtained after neoadjuvant chemotherapy. N = normalization, OS = oversampling with SMOTE, US = randomly undersampling.*

Model	Number of patients	pCR	No pCR	MRI 1	MRI 2	MRI 3	N	OS	US
1	258	77	181	X	X	X			
2	258	77	181	X	X				
3	258	77	181	X		X			
4	258	77	181		X	X			
5	258	77	181	X	X	X	X		
6	342	161	181	X	X	X		X	
7	342	161	181	X	X	X	X	X	
8	168	74	94	X	X	X			X
9	168	74	94	X	X	X	X		X

Table 9 *Data configuration for different prediction models with RCB outcome (RCB-0 vs RCB-1 vs RCB-2 vs RCB-3). MRI1 was obtained before neoadjuvant chemotherapy, MRI2 was obtained during neoadjuvant chemotherapy, and MRI3 was obtained after neoadjuvant chemotherapy. N = normalization, OS = oversampling with SMOTE, US = randomly undersampling.*

Model	Number of patients	RCB-0	RCB-1	RCB-2	RCB-3	MRI 1	MRI 2	MRI 3	N	OS
1	177	77	12	62	26	X	X	X		
2	177	77	12	62	26	X	X			
3	177	77	12	62	26	X		X		
4	177	77	12	62	26		X	X		
5	177	77	12	62	26	X	X	X	X	
6	280	77	63	74	66	X	X	X		X
7	280	77	63	74	66	X	X	X	X	X

4.3.3.2 Binary Classification Models – Total cohort

The results of the binary classification model are presented in Figure 17, 18 and appendix G. In these results, the entire test set is used for evaluation. Figure 17 shows that model 2 had the best ROC curve. Figure 18 illustrates that in models 1 through 7, regardless of which data type is used, specificity is higher than sensitivity. Models 5 through 9 show increased sensitivity compared to model 1 when using both clinical and radiological data. In models 8 and 9, the sensitivity and specificity for each data type were closer to each other compared to models 1 through 7. Besides this, the sensitivity in the different models ranged from 0.25 to 0.75, while specificity ranged from 0.67 to 1.00. The best trade-off between sensitivity and specificity was achieved with model 2, which used both clinical and radiological data, yielding sensitivity and specificity values of 0.75 and 0.83, respectively. The AUC for model 2 was 0.79. This model first applied an SGD classifier with a modified Huber loss. This classifier combines elements of logistic regression and SVM with SGD learning. The outcome of this classifier is scaled with a standard scaler and used in a new SGD classifier with a Hinge loss. This classifier applied a linear model similar to an SVM. The output of this second classifier is used as input to a Bernoulli NB classifier for final classification.

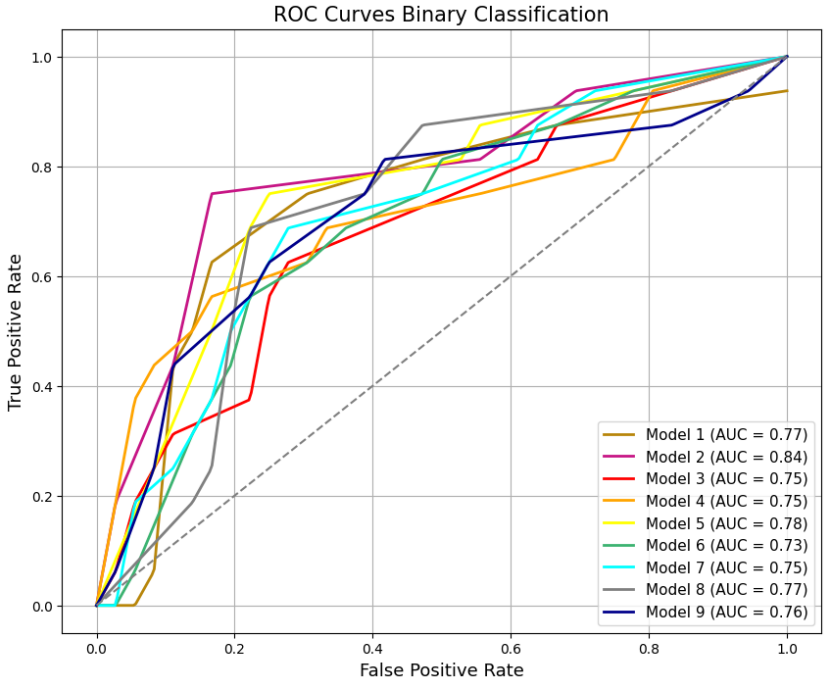

Figure 17 *Best ROC curves (clinical, radiological, or clinical and radiological data) for binary classification.*

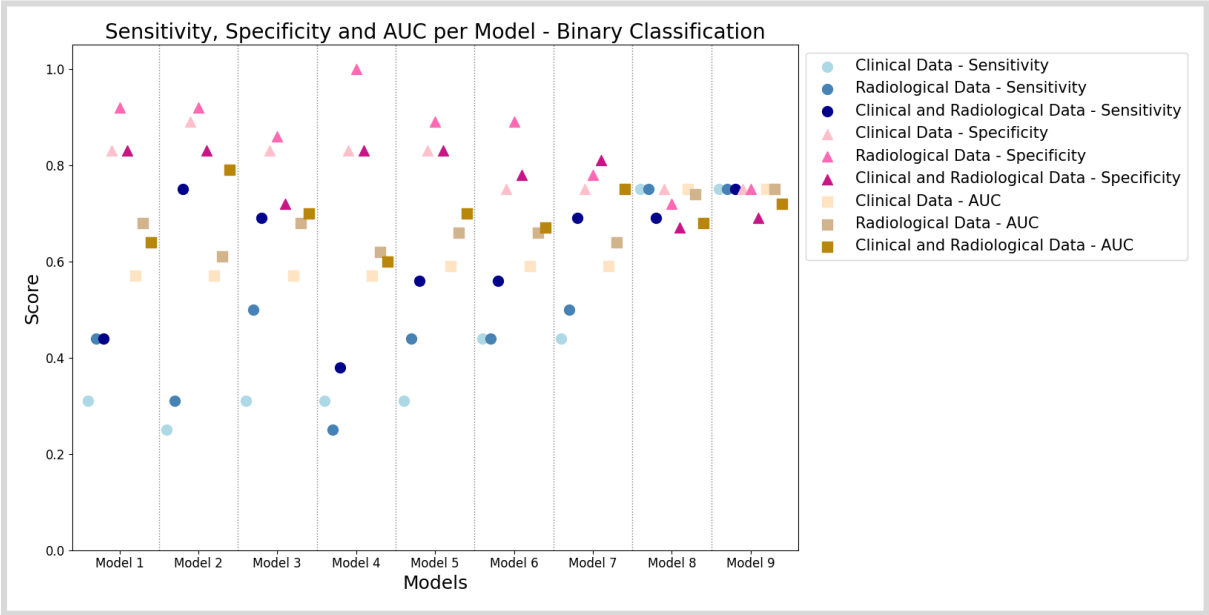

Figure 18 *Sensitivity, specificity, and AUC of the binary classification model tested on the test set with the total cohort, categorized by clinical data (circle), radiological data (triangle), and the combination of clinical data and radiological data (square).*

In this study, the binary classification model was also tested on the different breast cancer subtypes. All the results are presented in appendix H. There were only two pCR samples of the Hr+ subgroup, so no analysis was performed on this group. The sensitivity and specificity in the Her2+ subgroup ranged from 0.25 to 1.0. The best trade-off between sensitivity and specificity was achieved with model 6 or 7 which used both clinical and radiological data. This model achieved a sensitivity of 1.0 and a specificity of 0.75. The AUC for this model was 0.88. For the TNBC subgroup, the sensitivity ranges from 0.0-1.0 and the specificity ranged from 0.30 – 1.0. The best trade-off between sensitivity and specificity for this subgroup was obtained with model 2, with clinical and radiological data. This model achieved a sensitivity and specificity of respectively 0.88 and 0.80. The AUC for this model was 0.84.

4.3.3.3 RCB-score Models

The highest accuracy and Cohen's kappa, respectively 0.69 and 0.51, for the multi-class RCB classification model were achieved by model 2 (appendix I). For RCB-0 classification, a one (RCB-0) vs. all (RCB-1, RCB-2, RCB3) approach was used. Figure 19 shows that model 2 had the best ROC curve for RCB-0 classification. Figure 20 and appendix J show the results of the RCB-0 classification model. Across all data types, sensitivity is generally higher than specificity. In models 6 and 7, sensitivity for all data types is lower compared to model 1. For RCB-0, sensitivity ranged from 0.50 to 1.00, and specificity ranged from 0.50 to 0.80. The best trade-off between sensitivity and specificity was achieved by model 2, using clinical and radiological data, with a sensitivity of 0.81, specificity of 0.80, and an AUC of 0.81. Model 2 used a gradient boosting classifier for the classification. Appendix K (Figure 33) presents the confusion matrix for RCB-0 prediction with model 2 based on clinical and radiological data. There were four false positive samples. There were also three false negative samples: two were predicted as RCB-2, and one as RCB-3.

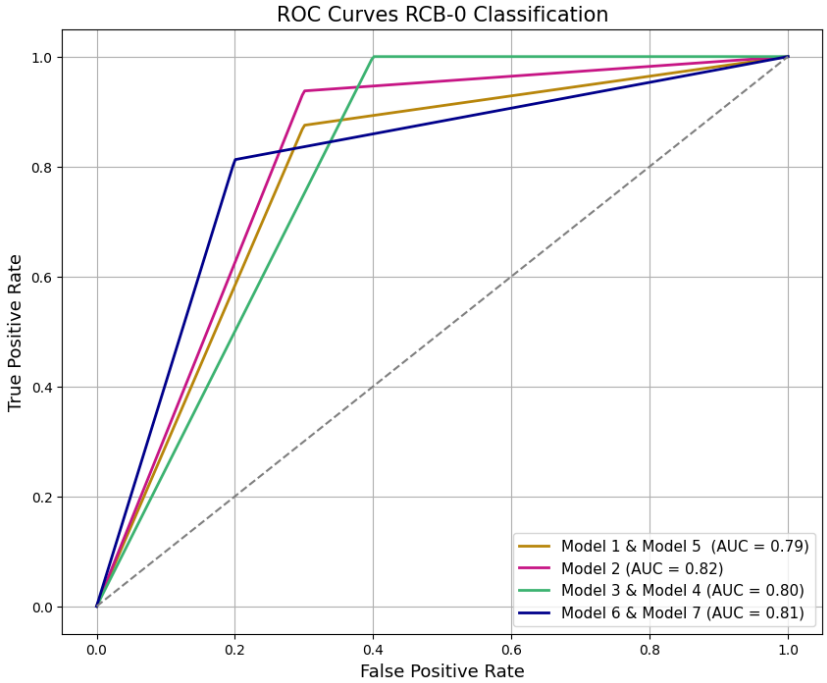

Figure 19 *Best ROC curves (clinical, radiological, or clinical and radiological data) for RCB-0 classification.*

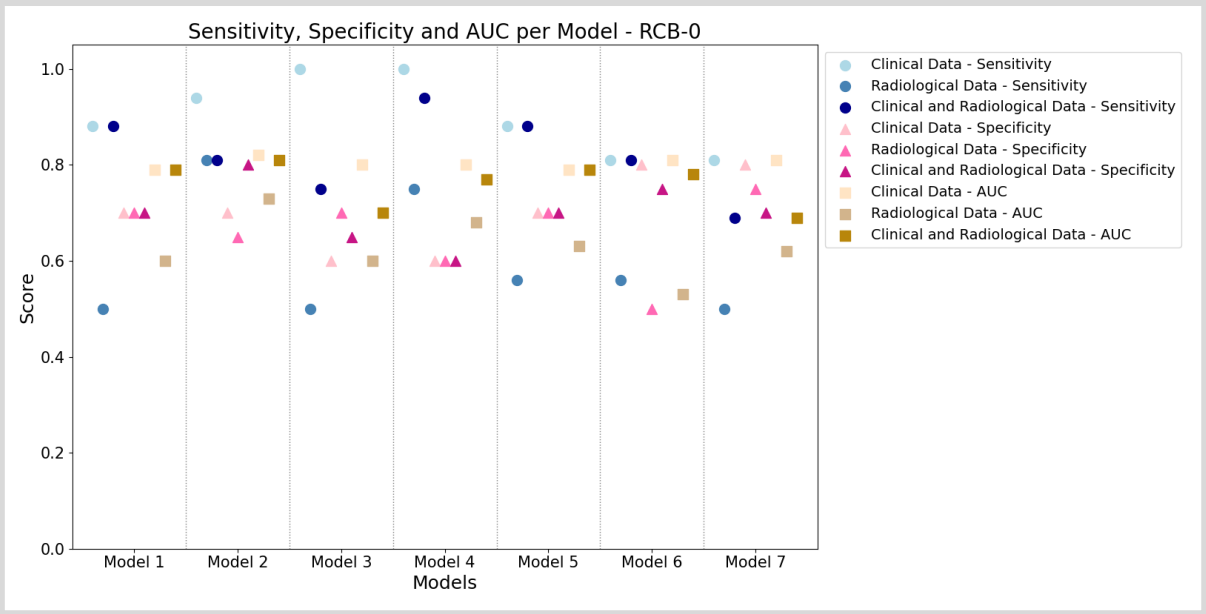

Figure 20 *Sensitivity, specificity, and AUC of the multi class classification model for RCB-0, categorized by clinical data (circle), radiological data (triangle), and the combination of clinical and radiological data (square).*

4.4 Discussion

4.4.1 Radiological Features

In this study, the focus was specifically on using only 3D shape features, calculated with Pyradiomics. Shape features refer to the geometric properties of a structure, such as size and shape. These characteristics are intrinsic to the structure itself and are not dependent on the pixel intensity values in the MRI. This approach is advantageous because the intensity of MRIs can vary across different acquisition times due to several factors, including MRI settings. By primarily concentrating on shape features, the analysis included stable and reproducible characteristics that are less sensitive to intensity fluctuations. This enables the tracking of biological changes, such as tumor growth, rather than technical variations. Furthermore, shape features can be linked to clinical relevance; for instance, patients who show a reduction in tumor volume during neoadjuvant chemotherapy often have a higher likelihood of achieving a pathological complete response (pCR). Additionally, a consistent decrease in tumor size is associated with a greater likelihood of pCR, particularly in the HR+ subgroup [46].

However, one intensity feature was included in this analysis: the change in intensity over time. When a tumor responds to neoadjuvant chemotherapy, necrosis can occur, leading to a decrease in the number of tumor cells and a reduction in the blood supply to the tumor. This response may be observed as a decrease in the intensity of the images. However, MRI intensities are not standardized, meaning that tissue can exhibit different intensity values across different scans. To ensure a reliable analysis using MRI intensity feature, tumor segmentation was normalized against a reference tissue, in this case, the sternum. The sternum is a tissue that absorbs little to no contrast agent, ensuring that this structure remains consistent in intensity across different scans. By dividing the average tumor intensity by the average sternum intensity, the tumor intensity was standardized. This normalization method minimizes variations in intensity between different scans.

Besides this, the radiological features were calculated based on segmentations that are not entirely perfect. The segmentations made with the MAMA-MIA network achieved a DSC of 0.69, indicating a discrepancy between the automated segmentation and the actual tumor size and shape. For this reason, a manual review was performed by a technician physician in training. The technician physician in training had no prior experience in evaluating breast tumor segmentations, which raises concerns about the quality of the tumor segmentation corrections. As a result, the calculated radiological features may differ from the actual tumor characteristics. Furthermore, the time intervals between the MRIs were not consistent across patients due to the varying chemotherapy regimens used for different subtypes of breast cancer. This variability in time intervals could also impact the radiological features, as tumor responses may change over time.

4.4.2 Statistical analysis

The clinical data showed a correlation between doxorubicin and trastuzumab (SCC 0.94). Besides this, doxorubicin and cyclophosphamide showed the same correlation (SCC 0.94). Additionally, there was a correlation between the presence of the Her2+ receptor and trastuzumab (SCC 0.93). The combination of the medications doxorubicin and cyclophosphamide can be used to treat all subtypes of breast cancer. Doxorubicin helps reduce tumor mass by damaging the DNA within tumor cells, preventing them from dividing and ultimately causing cell death. Cyclophosphamide works by adding an alkyl group to the DNA of the tumor cells, disrupting cell division and leading to cell death. On the other hand, trastuzumab is used specifically for Her2+ breast cancer, as it binds to the Her2 receptors on the tumor cells. This binding inhibits the signals that promote cell division, thereby slowing tumor growth.

When conducting multiple statistical tests on the same dataset, the probability of obtaining at least one false positive result increases. For this reason, corrections were made for multiple testing. The Bonferroni correction is a stringent method that divides the established significance level (a) by the number of tests performed (b). As a result, each p-value must be smaller than (a/b) to be considered significant. While this approach reduces the risk of false positives, it increases the likelihood of false negatives, especially when a large number of tests are conducted. An alternative approach for the correction of multiple testing is the Benjamini-Hochberg correction, which is based on the false discovery rate (FDR). Rather than applying a strict threshold for each test, this method ranks all p-values from lowest to highest and adjusts the threshold incrementally based on their rank. This allows for the detection of true effects even when a larger number of tests are performed, without a significant increase in false positives [79]. Given that this study did not involve a large number of tests, the Bonferroni method was chosen for correction for multiple tests.

In the multivariate analysis, the odds ratio is one of the outcome parameters. The odds ratio is a measure of the association between an independent variable and pathologic response while controlling for other variables. An odds ratio of one indicates no association between the variable and the outcome; in this case, the probability of the outcome does not change with a variation in the independent variable. Among the statistically significant clinical variables, it is notable that the Her2+ receptor and tumor grade exhibit particularly high positive associations with odds ratios of 19.8 and 2.5, respectively. This is consistent with existing literature, which shows that the Her2+ receptor subgroup is more likely to achieve a pCR following neoadjuvant chemotherapy and tumors with a higher grade generally exhibit a more favorable response to neoadjuvant chemotherapy, further enhancing the chances of achieving pCR [80]. The statistically significant radiological variables, elongation and surface-to-volume ratio, show an

odds ratios of 8.7 and 0.072, respectively. This suggests a greater likelihood of achieving pCR when the tumor assumes a rounded shape. On the other hand, the analysis indicates that a higher surface-to-volume ratio decreases the probability of a pathologic response. A higher surface-to-volume ratio can imply that the tumor has a more irregular or convoluted shape, or that it consists of a non-homogeneous mass divided into multiple parts. The findings are consistent with the literature, which states that round and oval tumors have a higher pCR rate compared to irregular tumors [81].

In both univariate and multivariate analyses, the variables PR and tumor type exhibit confidence intervals that include one. A confidence interval that spans one indicates uncertainty about the direction and strength of the relationship between the variable and the outcome, making it difficult to draw reliable conclusions in these cases. Furthermore, the wide range in the confidence interval, observed in certain variables, such as elongation, implied considerable variability in the data. This variability complicated the identification of a consistent association, as the observed effects may differ significantly across individual cases. To achieve a more robust and reliable OR, it would be necessary to increase the dataset.

4.4.3 Machine learning prediction models

In this study, stratification was applied to pathologic response and the date of the first MRI. This approach ensures that the proportional class distribution of pathologic response and the date of the first MRI scan is maintained across the training and test sets. Stratification by pathologic response was essential because it was the primary outcome measure in this study. Ensuring that the proportional distribution of pathologic classes is preserved across the training and test sets prevents the model from being biased toward the majority class. Without this proportionality, the model might fail to generalize effectively, leading to unreliable performance metrics, particularly for the minority class. Besides this, the dates of the MRIs are from the period 2005-2023. Due to technological advancements, such as improvements in resolution and image processing algorithms, the quality of MRIs has increased during this period. Additionally, treatment methods themselves and their effectiveness have also evolved over the years. Stratifying by MRI date helps ensure that advancements in imaging technology and changes in clinical practices are evenly represented in the training and test sets

The goal of this study was to predict the pathologic response, with the aim of providing personalized breast cancer care. Personalized care in this context might involve avoiding surgical removal of the breast tissue if the model can accurately predict the pCR after neoadjuvant chemotherapy. The evaluation of the models in this study took both sensitivity and specificity into account. A high sensitivity indicates that the model is effective in correctly identifying all patients with a pathological response, which align with the goal of this study. However, it is also important that the model exhibits high specificity, as this reflects its ability to correctly identify patients without a pCR. If the model fails to adequately predict non-pCR patients, it may falsely predict a pCR, potentially resulting in missing necessary follow-up treatment. For this reason, the optimal model performance involves a trade-off between sensitivity and specificity.

The various analyses in this study indicated that the models with the best trade-off between sensitivity and specificity for a pathologic prediction used both clinical-pathological and radiological data, which is consistent with findings in the existing literature. For example, Jung et al. used clinical and demographic variables to develop a machine learning model for the prediction of pCR following neoadjuvant chemotherapy. The study included a total of 1003 patients. The clinical-pathological variables used in the machine learning model included age,

BMI, T-stage, N-stage before the start of chemotherapy, serum carbohydrate antigen 15-3, Ki-67, and receptor status for ER, PR, and Her2+. A total of five different machine learning models were trained: gradient boosting machine, support vector machine, random forest, decision tree, and neural network. External validation of the gradient boosting machine classifier demonstrated a sensitivity of 72.8% and a specificity of 77.7% [82]. On the other hand, Fan et al. made a prediction model for pCR with radiomic features and included 57 patients. 158 radiomics features were calculated from DCE-MRI prior to the start of neoadjuvant chemotherapy. Feature selection was performed using the wrapper subset evaluator. Subsequently, a logistic regression classifier was trained and tested using leave-one-out cross-validation. This study achieved an AUC of 0.703 [83]. Additionally, Zeng et al. included 117 patients to predict pCR based on both clinical and radiological characteristics. Several clinical-pathological features were selected, such as receptor status and the Ki-67 index. Feature selection for these variables was performed using logistic regression. A total of 851 radiological features were calculated using radiomics, both prior to the initiation of neoadjuvant chemotherapy and after 2-4 cycles. Delta radiomics features were also computed. Feature selection for the radiological variables was conducted using the intraclass correlation coefficient, Pearson- or Spearman tests. The selected features were used in a logistic regression model to predict the likelihood of pCR. The combination of clinical-pathological and radiomics features resulted in a sensitivity of 0.875 and a specificity of 0.850 [57]. There are no studies that utilized clinical or radiological data to predict the RCB score. Beside this, it remains in this study uncertain whether all RCB outcomes align with reality. The conversion of a binary response to an RCB score was done using the method from Appendix A. However, this conversion method includes certain assumptions, making it unclear if all conversions are reliable.

In this study, a multi-class prediction model was trained and tested with the RCB score as the outcome variable. A multi-class prediction model was chosen because accurately predicting each RCB class can be beneficial for the prognostic assessment. However, a binary approach was also applied in this study to evaluate the RCB-0 score. This approach used the one-vs-all method, comparing RCB-0 against RCB-1, RCB-2, and RCB-3. The specific focus on RCB-0 was selected because RCB-0 is the most important class to predict in this clinical problem, as it corresponds to pCR. Predicting RCB-0 could indicate that surgical removal of breast tissue may not be necessary. In this specific clinical context, predicting the degree of residual disease (RCB-1, RCB-2, or RCB-3) is less relevant, as all patients without a pathological complete response will still require surgical treatment, regardless of which remaining RCB class they belong to. It is recommended to train and test a binary model using the one-vs-all (RCB-0 vs RCB-1, RCB-2, and RCB-3) approach in addition to the multi-class prediction model, as the results may differ from the approach taken in this study.

It was initially hypothesized that normalizing the radiological features would have a positive effect on the performance of the machine learning models. Normalization ensures standardization of values within a consistent range, which was expected to enhance the model's ability to learn and recognize patterns. However, this hypothesis did not hold across for the binary and RCB-0 predictions (Model 5 Figure 17, 18, 19, 20). This could be due to the fact that TPOT also utilized machine learning models that are less affected by the scale of the data, such as DT and RF, which reduces the impact of normalization.
In the prediction models with the binary outcome specificity was generally higher than sensitivity. This is likely due to the predominance of the no-pCR class in these models, which leads the model to be more effectively trained to accurately predict this class. However, for the RCB-0 class, there was a better balance between specificity and sensitivity, which is likely due

to the larger number of RCB-0 cases in the dataset compared to the other RCB classes. For this reason, balancing the outcome variable was applied with over- and undersampling. It was decided not to apply undersampling for balancing the RCB prediction models, due to the rarity of the RCB-1 class. In case of the binary pCR classification, it was expected that balancing the outcome variable would enhance the sensitivity, as the pCR class represented the minority class. For the predictive models with a binary response (pCR vs no pCR) an increase in sensitivity was observed when employing both balancing techniques on the outcome variable (Figure 18, Model 6 and 8). Furthermore, it was expected that balancing the dataset would lead to a convergence of sensitivity and specificity, as each class would be equally represented. This effect was observed with undersampling (Figure 17, models 8 and 9), where sensitivity and specificity became more similar. This happens because undersampling reduces the number of instances from the majority class, which shifts the model's focus toward the minority class. As a result, the model tends to achieve a more balanced performance between sensitivity and specificity. However, this effect was less pronounced with oversampling (Figure 17, models 6 and 7). Oversampling with SMOTE generates synthetic data points by interpolating between existing instances of the minority class. While this helps balance the dataset, it can lead to overfitting, where the model becomes too specialized in recognizing these synthetic instances. Since synthetic data points may lack the variability of real-world examples, this can impair the model's ability to generalize to new, unseen data.

When a patient responds to neoadjuvant chemotherapy, the difference in tumor size is greatest between the MRI at timepoint 1 and the MRI at timepoint 3. For this reason, the delta radiomic features were calculated based on the MRI at timepoint 1 and timepoint 3. However, this study shows that predicting pCR provides the best trade-off between sensitivity and specificity when only the MRI from timepoint 1 and timepoint 2 are used for the radiological features, both in the binary predictions and RCB-0 classifications. This suggest that the initial chemotherapy response, for some reason, had more effect on the tumor than the later chemotherapy response. So, the MRI scans taken at time points 1 and 2 provide more meaningful information for predicting pCR compared to the MRI scans at time point 3. In these cases, it is expected that later tumor responses (observed between MRI timepoint 2 and MRI timepoint 3) are more consistently present, for example due to chemotherapy resistance. This uniformity in later response may limit its ability to distinguish between outcomes in pathologic response. To further investigate this result, it may be valuable in future research to retrain the other models using only the MRI from timepoint 1 and timepoint 2 in case of binary and RCB-0 predictions.

An analysis was also conducted on the Her2+ and TNBC subgroups. However, the test population for these subgroups consisted of a relatively small number of patients, raising concerns regarding the reliability of the results. Besides this, it is expected that better results could be achieved by training separate models for each specific breast cancer subgroup. If more patients of each subgroup are included in future studies, the concordance between the model and the pathologist can be assessed. Currently, the concordance between rCR and pCR for the Her2+ subtype is estimated at 50%, while for the HR+ subtype and TNBC, the concordance is reported to be lower, around 30% [18]. To demonstrate the clinical value of the model, the concordance between the model and the pathologist must exceed the currently noted concordance between rCR and pCR. This would indicate that the model provides more reliable and accurate predictions compared to traditional radiological assessments, highlighting its potential in clinical practice. This is especially important in HR+ and TNBC subtypes, where the concordance between rCR and pCR is relatively low (30%).

Furthermore, for the clinical implementation of the machine learning predictive model for the pathological response in breast cancer patients, it is essential to collect more data and perform external validation. The current dataset is not sufficient to achieve the level of accuracy required for reliable clinical use. A larger dataset enables the model to learn patterns more effectively, improving its ability to identify complex relationships and make accurate predictions. Besides this, external validation is necessary to test the model's performance outside the context of the initial data. This involves applying the model to data from different institutions or populations to assess its generalizability. Without external validation, it is difficult to determine whether the model's predictions can be trusted across different clinical environments.

4.5 Conclusion

In this study, 18 clinical-pathological variables were combined with 14 different 3D shape features and one intensity feature for the prediction of pCR. Besides this, the radiological delta features were also computed. This study suggests that, before the initiation of neoadjuvant chemotherapy, six statistically significant features can be identified: ER receptor, PR receptor, Her2 receptor, tumor type, tumor grade, elongation, and the surface-to-volume ratio. TPOT was used to train and test various dataset configurations. For both binary pCR prediction and RCB-0 classification, the optimal trade-off between sensitivity and specificity was achieved using both clinical and radiological data. For the binary pCR and RCB-0 outcomes, the model that included MRI data from timepoints 1 and 2 (model 2) showed sensitivities of 0.75 and 0.81, respectively. The specificities for these outcomes were 0.83 and 0.80. For clinical implementation of the prediction models, it is recommended to train and test on a larger dataset and perform external validation.

5 SUMMARY

5.1 General Summary

This research aimed to investigate the potential of AI in predicting pCR in patients diagnosed with stage I-III breast cancer who are receiving neoadjuvant chemotherapy. Specifically, we investigated how clinical-pathological and radiological information can be leveraged to enhance predictive accuracy regarding treatment outcomes. For this study, a total of 18 clinical-pathological variables have been included, which are important for understanding patient health status and tumor characteristics. These variables include for example receptor status (ER, PR and Her2+), tumor type, tumor grade, ASA score, and the number of chemotherapy cycles.

258 patients underwent a DCE-MRI before, during, and after the neoadjuvant chemotherapy, with a total of three scans per patient (MRI1, MRI2 and MRI3). This imaging modality is used for assessing tumor characteristics and response to therapy. Tumor segmentation was performed using two deep learning networks: the Zhang network and the MAMA -MIA network. During the evaluation of these networks, it was found that the MAMA-MIA network outperformed the Zhang network in terms of segmentation accuracy. The initial segmentation results yielded a DSC of 0.69 for the MAMA-MIA network, indicating a quite good level of agreement between the predicted and actual tumor boundaries. In contrast, the Zhang network showed in visual inspections that it often failed to segment the tumor accurately, which limited further application and evaluation.
Following a manual review and necessary adjustments of the automatically segmented tumors by the MAMA-MIA network, the DSC improved significantly to 0.84.

From the segmented tumors, a set of 14 3D shape features were calculated. Additionally, one intensity feature was calculated, which involved normalization against a reference tissue, the sternum. The statistical analysis was conducted on the features available at the start of neoadjuvant chemotherapy, identifying six features that were statistically significant in predicting pathologic responses: ER receptor, PR receptor, Her2 receptor, tumor type, tumor grade, elongation, and the surface-to-volume ratio.

Various machine learning models were trained using TPOT, incorporating clinical, radiological, or both clinical and radiological data. In this process, a total of nine dataset configurations were used. The outcome variables were either binary (pCR vs no-pCR) or multi class (RCB-0, RCB-1, RCB-2 or RCB-3). In case of the multi class models, a one-vs-all (RCB-0 vs RCB-1, RCB-2, RCB-3) evaluation was also performed. For the binary outcome, pCR predictions appeared to have the best trade-off between sensitivity and specificity when employing combined data, with only radiological features from MRI1 and MRI2. The sensitivity and specificity achieved in this case were respectively 0.75 and 0.83. The multi-class prediction model for RCB showed the highest accuracy of 0.69 and Cohen's kappa of 0.51 using clinical and radiological data from MRI1 and MRI2. The best trade-off between sensitivity and specificity for RCB-0 was also with combined data, where only radiological features from MRI1 and MRI2 were included. For this case, the sensitivity and specificity equals respectively 0.81 and 0.80. While the initial findings are encouraging, it is recommended to expand the dataset to facilitate more accurate predictions. Additionally, external validation is crucial for the clinical implementation of these predictive models.

6 FUTURE PERSPECTIVES

6.1 Literature Review

This study was the first study within Deventer Hospital focusing on predicting pCR in breast cancer patients undergoing neoadjuvant chemotherapy. For this reason, the decision was made to first explore the data using hand-crafted features extracted from segmented breast tumors, along with a machine learning-based prediction for pCR. Despite the fact that a machine learning-based prediction with TPOT also has a black-box nature, efforts were made to understand and explain as many aspects as possible in the prediction process. However, the results of this study showed that no optimal reliability can be achieved for predicting pCR with the method; sensitivity and specificity generally fluctuate in a range of 0.60–0.80. There are several ways to further improve the reliability of the pCR prediction. For example, more radiological features, such as texture features, could be added to the machine learning based prediction model. Additionally, deep learning models appear to have a positive effect on the reliability of pCR prediction in breast cancer patients:

- Yungsong et al. set up a study comparing the performance of deep learning features and radiomics features in predicting pCR in breast cancer patients undergoing neoadjuvant chemotherapy [84]. These features were based on pretreatment DCE-MRI. For the deep learning analysis, a rectangular box of 128x128x3 was used to select three consecutive slices of the tumor that showed the largest cross-sectional area. Additionally, the data was normalized between 0-1, and data augmentation using rotation and flipping was applied. The features were extracted using a pre-trained CNN ResNeXt50. This network consisted of three fully connected layers, with the output being the probability of pCR. Kinetic and molecular information was added to the first fully connected layers. In this method, heatmap features can be extracted, allowing for better determination of the extent to which a deep learning feature influences the prediction of pCR, thereby enhancing the interpretability of the model. This study shows that the deep learning model, using a combination of image, kinetic, and molecular information, provides the best performance with an accuracy of 0.77.

- Honygi et al. used a CNN to predict pCR, RCB, and the progression free survival in breast cancer patients undergoing neoadjuvant chemotherapy [54]. These predictions are made using longitudinal multiparametric MRI, demographic information, and molecular subtypes as input. In this study, the breast tumor is not segmented before the image is used as input into the deep learning model. Additionally, three different deep learning configurations are used for extracting the radiological features:

  a) Stacking method. This method overlays two MRI images from different time points. These stacked images are fed into a ResNet-based CNN model. This approach combines the images for feature extraction.
  b) Concatenation method. In this method, images from two time points are processed separately through the first layers of the neural network. The outputs of these layers (feature maps) are combined into one larger input. This larger input is processed further by the next layers.
  c) Integrated method. The two MRI images from different time points were fed through two separate convolutional branches of the network. Each branch processed one image independently. The result of this were two different feature maps. A pixel-wise operation was performed on the two feature maps: pixel-wise addition and pixel-wise subtraction. The results were combined and passed through the rest of the network. This way, the

network learned not only to process images from different time points but also how the images evolve in relation to each other, preserving temporal information completely. In each method, non-imaging features were first processed through three fully connected layers, and then concatenated with the image features extracted from the MRI images. Two fully connected layers processed the combined image and non-imaging features to predict the final clinical outcome.

The integrated method shows the highest accuracy in this study, with a score of 0.81.

- El Adoui et al. set up a study in which multiple CNN models were trained and tested for predicting pCR in breast cancer patients [53]. In each model, the volume of interest was cropped. When tumor segmentation was applied, the tumor within the volume of interest was segmented using a U-Net deep learning architecture. For combining pre- and post-treatment MRI images, concatenation was performed, followed by fully connected layers.

  The following models were developed:
  1. Using only pretreatment examination with segmentation (single-input CNN)
  2. Using only posttreatment examination with segmentation (single-input CNN)
  3. Using only pretreatment examination without segmentation (single-input CNN)
  4. Using only posttreatment examination without segmentation (single-input CNN)
  5. Using both pretreatment and posttreatment examination with segmentation (multi-input CNN)
  6. Using both pretreatment and posttreatment examination without segmentation (multi-input CNN)

  This study shows that the highest accuracy, 0.91, was achieved with model 6.

- Gao et al. developed a Multi-model Response Prediction (MRP) system capable of predicting the response to neoadjuvant chemotherapy in breast cancer patients using real-world data [85]. This system was trained on pre-NAT mammogram images and longitudinal MRI scans. In addition to using imaging data, the system incorporates radiological, histopathological, personal and clinical (RHPC) information, which includes a variety of patient data types: radiological assessments, histopathological evaluations, personal patient information, and clinical data.

  The system utilized two separate models:
1. iMGrhpc – this model used pre-NAT mammogram images combined with RHPC data.
2. iMRrhpc – this model used longitudinal MRI scans and RHPC data. For the training of the MRI-based model, weights from a previously trained model on a large medical dataset were loaded. This technique, known as transfer learning, involves retraining (or fine-tuning) a model that has already been trained on similar medical imaging data with the specific MRI data for this project.

  Each model itself comprises two modules:
1. A knowledge-learning module: this module attempts to predict RHPC information using only image features. This process is also known as cross-model knowledge learning. The aim of this module is to learn how RHPC features are associated with the images. This strengthens image feature extraction, allowing the model to better identify subtle details and patterns in the imaging data that are relevant to clinical characteristics.

2. A response prediction module: this module uses both RHPC and imaging information to predict the treatment response.
   The individual models, iMGrhpc and iMRrhpc, are combined into an MRP system. This integrated system enables a combined prediction by leveraging both RHPC data and imaging modalities (mammogram and MRI). The MRP system is trained on a cohort of 2682 patients from the Netherlands Cancer Institute. Model performance is evaluated on an internal test set with 120 patients and three external test sets with diverse patient groups: 288 patients from Duke University, 85 patients from Fujian Provincial Hospital (FJPH), and 508 patients from the I-SPY2 study.

   The results of this study showed that the MRP system provides more accurate predictions of the response to neoadjuvant chemotherapy in the pre-NAT, mid-NAT, and post-NAT phases compared to the single-modality iMGrhpc and iMRrhpc models. In the external FJPH dataset, the MRP system showed a sensitivity of 0.75 and a specificity of 0.80 at the pre-NAT stage. For the post-NAT phase, the sensitivity and specificity of the MRP system was 0.75 and 0.73, respectively.

The deep learning networks used for classification often have many parameters that are essential for the model to learn and extract features from the data. A substantial amount of data is typically needed to effectively train these parameters. In peripheral hospitals, this volume of data is often unavailable, so a pretrained network is typically used to set up a deep learning model [86]. This pretrained network can be fine-tuned with the data from the peripheral hospital—in this case, Deventer Hospital. However, for deploying a deep learning network to predict pathologic response in breast cancer patients following neoadjuvant chemotherapy, no pretrained network is currently publicly available.

For the future perspective of this study involving deep learning networks with data from Deventer Hospital, the following steps are recommended:

- *Incorporating Clinical and Radiological Data:* It is recommended to use both clinical and radiological data in predicting pCR in breast cancer patients undergoing neoadjuvant chemotherapy. Clinical data can be concatenated with radiological data prior to the fully connected layers in a deep learning network. Additionally, it is advised to include longitudinal MRI images. To combine the feature maps from multiple MRI time points, for example the integrated method of El Adoui et al. can be used.
- *Avoiding Tumor Segmentation in MRI Input Images:* It is not recommended to segment the tumor in the MRI input images. The study by El Adoui et al. demonstrated that tumor segmentation does not add value to the prediction of pCR in breast cancer patients using deep learning imaging features. Omitting tumor segmentation not only saves considerable time but also reduces uncertainty in feature calculations, given that tumor segmentation does not have a perfect reliability. A rectangular box around the volume of interest could be used for deep learning feature extraction, as done by both Yungsong et al. and El Adoui et al.
- *Using a Pretrained Network:* It is recommended to use a pretrained network so that the model can be fine-tuned on Deventer Hospital data. However, no pretrained network for this research question is currently available online. It is therefore advisable to reach out to other centers in the Netherlands that have a pretrained network for similar purposes and inquire about the availability of these models. Additionally, for clinical applicability, it will be important to consider the explainability of the model, for example, by using a saliency map.

- *External Validation of the Network:* In addition to internal validation with a test set, it is recommended to perform external validation. For a grant application, contact has already been made with Medisch Spectrum Twente and Ziekenhuisgroep Twente regarding the inclusion of more patients.

References

[1] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast Off J Eur Soc Mastology 2022;66:15. https://doi.org/10.1016/J.BREAST.2022.08.010.

[2] NKR Cijfers | Incidentie - Grafiek n.d. https://nkr-cijfers.iknl.nl/viewer/incidentie-per-jaar?language=nl_NL&viewerId=dc53fb13-d636-4548-8bf7-818cc335326e (accessed October 14, 2024).

[3] Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA. Subtypes of Breast Cancer. Breast Cancer 2022:31–42. https://doi.org/10.36255/EXON-PUBLICATIONS-BREAST-CANCER-SUBTYPES.

[4] Mohammed AA. The clinical behavior of different molecular subtypes of breast cancer. Cancer Treat Res Commun 2021;29. https://doi.org/10.1016/j.ctarc.2021.100469.

[5] Oncoguide n.d. https://oncoguide.nl/#!/projects/7/tree/10543/101 (accessed April 5, 2024).

[6] Bhushan A, Gonsalves A, Menon JU. Current State of Breast Cancer Diagnosis, Treatment, and Theranostics. Pharmaceutics 2021;13. https://doi.org/10.3390/pharmaceutics13050723.

[7] Mann RM, Kuhl CK, Kinkel K, Boetes C. Breast MRI: guidelines from the European Society of Breast Imaging. Eur Radiol 2008;18:1307–18. https://doi.org/10.1007/s00330-008-0863-7.

[8] Yankeelov TE, Gore JC. Dynamic Contrast Enhanced Magnetic Resonance Imaging in Oncology: Theory, Data Acquisition, Analysis, and Examples. Curr Med Imaging Rev 2009;3:91–107. https://doi.org/10.2174/157340507780619179.

[9] Hack CC, Voiβ P, Lange S, Paul AE, Conrad S, Dobos GJ, et al. Local and Systemic Therapies for Breast Cancer Patients: Reducing Short-term Symptoms with the Methods of Integrative Medicine. Geburtshilfe Frauenheilkd 2015;75:675. https://doi.org/10.1055/S-0035-1557748.

[10] Definition of localized therapy - NCI Dictionary of Cancer Terms - NCI n.d. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/localized-therapy (accessed July 31, 2024).

[11] Palumbo MO, Kavan P, Miller WH, Panasci L, Assouline S, Johnson N, et al. Systemic cancer therapy: achievements and challenges that lie ahead. Front Pharmacol 2013;4. https://doi.org/10.3389/FPHAR.2013.00057.

[12] Board PATE. Breast Cancer Treatment (PDQ®). PDQ Cancer Inf Summ 2024:1–5.

[13] Burguin A, Diorio C, Durocher F. Breast Cancer Treatments: Updates and New Challenges. J Pers Med 2021;11:808. https://doi.org/10.3390/JPM11080808.

[14] Korde LA, Somerfield MR, Carey LA, Crews JR, Denduluri N, Hwang ES, et al. Neoadjuvant Chemotherapy, Endocrine Therapy, and Targeted Therapy for Breast Cancer: ASCO Guideline. J Clin Oncol Off J Am Soc Clin Oncol 2021;39:1485–505. https://doi.org/10.1200/JCO.20.03399.

[15] Fazal F, Bashir MN, Adil ML, Tanveer U, Ahmed M, Chaudhry TZ, et al. Pathologic Complete Response Achieved in Early-Stage HER2-Positive Breast Cancer After Neoadjuvant Therapy With Trastuzumab and Chemotherapy vs. Trastuzumab, Chemotherapy, and Pertuzumab: A Systematic Review and Meta-Analysis of Clinical Trials. Cureus 2023;15:e39780. https://doi.org/10.7759/cureus.39780.

[16] Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet (London, England) 2014;384:164–72. https://doi.org/10.1016/S0140-6736(13)62422-8.

[17] Li X, Dai D, Chen B, Tang H, Wei W. Oncological outcome of complete response after neoadjuvant chemotherapy for breast conserving surgery: a systematic review and meta-analysis. World J Surg Oncol 2017;15:210. https://doi.org/10.1186/s12957-017-1273-6.
[18] Gampenrieder SP, Peer A, Weismann C, Meissnitzer M, Rinnerthaler G, Webhofer J, et al. Radiologic complete response (rCR) in contrast-enhanced magnetic resonance imaging (CE-MRI) after neoadjuvant chemotherapy for early breast cancer predicts recurrence-free survival but not pathologic complete response (pCR). Breast Cancer Res 2019;21:19. https://doi.org/10.1186/s13058-018-1091-y.
[19] Woo J, Ryu JM, Jung SM, Choi HJ, Lee SK, Yu J, et al. Breast radiologic complete response is associated with favorable survival outcomes after neoadjuvant chemotherapy in breast cancer. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol 2021;47:232–9. https://doi.org/10.1016/j.ejso.2020.08.023.
[20] van der Voort A, Louis FM, van Ramshorst MS, Kessels R, Mandjes IA, Kemper I, et al. MRI-guided optimisation of neoadjuvant chemotherapy duration in stage II–III HER2-positive breast cancer (TRAIN-3): a multicentre, single-arm, phase 2 study. Lancet Oncol 2024;0. https://doi.org/10.1016/S1470-2045(24)00104-9.
[21] Chen X, He C, Han D, Zhou M, Wang Q, Tian J, et al. The predictive value of Ki-67 before neoadjuvant chemotherapy for breast cancer: a systematic review and meta-analysis. Future Oncol 2017;13:843–57. https://doi.org/10.2217/FON-2016-0420.
[22] Zhang Z, Zhang H, Yu J, Xu L, Pang X, Xiang Q, et al. miRNAs as therapeutic predictors and prognostic biomarkers of neoadjuvant chemotherapy in breast cancer: a systematic review and meta-analysis. Breast Cancer Res Treat 2022;194:483–505. https://doi.org/10.1007/s10549-022-06642-z.
[23] Mao Y, Qu Q, Zhang Y, Liu J, Chen X, Shen K. The value of tumor infiltrating lymphocytes (TILs) for predicting response to neoadjuvant chemotherapy in breast cancer: a systematic review and meta-analysis. PLoS One 2014;9. https://doi.org/10.1371/JOURNAL.PONE.0115103.
[24] Feng K, Jia Z, Liu G, Xing Z, Li J, Li J, et al. A review of studies on omitting surgery after neoadjuvant chemotherapy in breast cancer. Am J Cancer Res 2022;12:3512–31.
[25] Alberts, Bray, Hopkin, Johnson, Lewis, Raff, Roberts W. Essential Cell Biology. 4th ed. Garland Science; 2013.
[26] Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov 2022;12:31–46. https://doi.org/10.1158/2159-8290.CD-21-1059.
[27] Startpagina - Borstkanker - Richtlijn - Richtlijnendatabase n.d. https://richtlijnendatabase.nl/richtlijn/borstkanker/startpagina_-_borstkanker.html (accessed November 25, 2024).
[28] American College of Radiology ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. ACR, Am Coll Radiol 2013. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads.
[29] Alamoodi M. Factors Affecting Pathological Complete Response in Locally Advanced Breast Cancer Cases Receiving Neoadjuvant Therapy: A Comprehensive Literature Review. Eur J Breast Heal 2024;20:8–14. https://doi.org/10.4274/ejbh.galenos.2023.2023-11-2.
[30] Susini T, Biglia N, Bounous VE. Prognostic Factors Research in Breast Cancer Patients: New Paths. Cancers (Basel) 2022;14. https://doi.org/10.3390/cancers14040971.
[31] Gündoğdu A, Uluşahin M, Çekiç AB, Kazaz SN, Güner A. Pathological complete response and associated factors in breast cancer after neoadjuvant chemotherapy: A retrospective study. Turkish J Surg 2024;40:073–81. https://doi.org/10.47717/TURKJSURG.2024.6308.
[32] Huong PT, Nguyen LT, Nguyen X-B, Lee SK, Bach D-H. The Role of Platelets in the

Tumor-Microenvironment and the Drug Resistance of Cancer Cells. Cancers (Basel) 2019;11. https://doi.org/10.3390/cancers11020240.
[33] Chen Z, Han F, Du Y, Shi H, Zhou W. Hypoxic microenvironment in cancer: molecular mechanisms and therapeutic interventions. Signal Transduct Target Ther 2023 81 2023;8:1–23. https://doi.org/10.1038/s41392-023-01332-8.
[34] Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. Ki67 in breast cancer: prognostic and predictive potential. Lancet Oncol 2010;11:174–83. https://doi.org/10.1016/S1470-2045(09)70262-1.
[35] Pan D, Wei K, Ling Y, Su S, Zhu M, Chen G. The Prognostic Role of Ki-67/MIB-1 in Cervical Cancer: A Systematic Review with Meta-Analysis. Med Sci Monit 2015;21:882. https://doi.org/10.12659/MSM.892807.
[36] Abdel-Fatah TM, Powe DG, Ball G, Lopez-Garcia MA, Habashy HO, Green AR, et al. Proposal for a modified grading system based on mitotic index and Bcl2 provides objective determination of clinical outcome for patients with breast cancer. J Pathol 2010;222:388–99. https://doi.org/10.1002/PATH.2775.
[37] Al-Janabi S, Van Slooten HJ, Visser M, Van Der Ploeg T, Van Diest PJ, Jiwa M. Evaluation of Mitotic Activity Index in Breast Cancer Using Whole Slide Digital Images. PLoS One 2013;8. https://doi.org/10.1371/JOURNAL.PONE.0082576.
[38] Ranganathan K, Sivasankar V. MicroRNAs - Biology and clinical applications. J Oral Maxillofac Pathol 2014;18:229. https://doi.org/10.4103/0973-029X.140762.
[39] Eren T, Karacin C, Ucar G, Ergun Y, Yazici O, İmamoglu GI, et al. Correlation between peripheral blood inflammatory indicators and pathologic complete response to neoadjuvant chemotherapy in locally advanced breast cancer patients. Med (United States) 2020;99:E20346. https://doi.org/10.1097/MD.0000000000020346.
[40] Yang G, Liu P, Zheng L, Zeng J. Novel peripheral blood parameters as predictors of neoadjuvant chemotherapy response in breast cancer. Front Surg 2022;9. https://doi.org/10.3389/FSURG.2022.1004687.
[41] Pesapane F, De Marco P, Rapino A, Lombardo E, Nicosia L, Tantrige P, et al. How Radiomics Can Improve Breast Cancer Diagnosis and Treatment. J Clin Med 2023;12. https://doi.org/10.3390/jcm12041372.
[42] Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to Radiomics. J Nucl Med 2020;61:488. https://doi.org/10.2967/JNUMED.118.222893.
[43] Pesapane F, Agazzi GM, Rotili A, Ferrari F, Cardillo A, Penco S, et al. Prediction of the Pathological Response to Neoadjuvant Chemotherapy in Breast Cancer Patients With MRI-Radiomics: A Systematic Review and Meta-analysis. Curr Probl Cancer 2022;46:100883. https://doi.org/10.1016/J.CURRPROBLCANCER.2022.100883.
[44] Nardone V, Reginelli A, Grassi R, Boldrini L, Vacca G, D'Ippolito E, et al. Delta radiomics: a systematic review. Radiol Med 2021;126:1571–83. https://doi.org/10.1007/S11547-021-01436-7.
[45] Guo L, Du S, Gao S, Zhao R, Huang G, Jin F, et al. Delta-Radiomics Based on Dynamic Contrast-Enhanced MRI Predicts Pathologic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Chemotherapy. Cancers (Basel) 2022;14:3515. https://doi.org/10.3390/CANCERS14143515/S1.
[46] Wang M, Du S, Gao S, Zhao R, Liu S, Jiang W, et al. MRI-based tumor shrinkage patterns after early neoadjuvant therapy in breast cancer: correlation with molecular subtypes and pathological response after therapy. Breast Cancer Res 2024;26:1–15. https://doi.org/10.1186/S13058-024-01781-1/FIGURES/6.
[47] Lakshmanan V, Robinson S, Munn M. Machine Learning Design Patterns Solutions to Common Challenges in Data Preparation, Model Building, and MLOps 2021.

[48] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci 2021;2:160. https://doi.org/10.1007/s42979-021-00592-x.
[49] Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 2019;36:250–6. https://doi.org/10.1093/bioinformatics/btz470.
[50] Katoch S, Chauhan SS, Kumar V. A review on genetic algorithm: past, present, and future. Multimed Tools Appl 2021;80:8091–126. https://doi.org/10.1007/S11042-020-10139-6/FIGURES/8.
[51] Khan N, Adam R, Huang P, Maldjian T, Duong TQ. Deep Learning Prediction of Pathologic Complete Response in Breast Cancer Using MRI and Other Clinical Data: A Systematic Review. Tomogr (Ann Arbor, Mich) 2022;8:2784–95. https://doi.org/10.3390/tomography8060232.
[52] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9:611. https://doi.org/10.1007/S13244-018-0639-9.
[53] El Adoui M, Drisis S, Benjelloun M. Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images. Int J Comput Assist Radiol Surg 2020;15:1491–500. https://doi.org/10.1007/s11548-020-02209-9.
[54] Dammu H, Ren T, Duong TQ. Deep learning prediction of pathological complete response, residual cancer burden, and progression-free survival in breast cancer patients. PLoS One 2023;18:e0280148. https://doi.org/10.1371/journal.pone.0280148.
[55] van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022;79:102470. https://doi.org/https://doi.org/10.1016/j.media.2022.102470.
[56] Wang AQ, Karaman BK, Kim H, Rosenthal J, Saluja R, Young SI, et al. A Framework for Interpretability in Machine Learning for Medical Imaging n.d.
[57] Zeng Q, Ke M, Zhong L, Zhou Y, Zhu X, He C, et al. Radiomics Based on Dynamic Contrast-Enhanced MRI to Early Predict Pathologic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Therapy. Acad Radiol 2023;30:1638–47. https://doi.org/https://doi.org/10.1016/j.acra.2022.11.006.
[58] Huang YH, Zhu T, Zhang XL, Li W, Zheng XX, Cheng MY, et al. Longitudinal MRI-based fusion novel model predicts pathological complete response in breast cancer treated with neoadjuvant chemotherapy: a multicenter, retrospective study. EClinicalMedicine 2023;58. https://doi.org/10.1016/J.ECLINM.2023.101899.
[59] Qin JB, Liu Z, Zhang H, Shen C, Wang XC, Tan Y, et al. Grading of Gliomas by Using Radiomic Features on Multiple Magnetic Resonance Imaging (MRI) Sequences. Med Sci Monit 2017;23:2168. https://doi.org/10.12659/MSM.901270.
[60] Zhang J, Saha A, Zhu Z, Mazurowski MA. Hierarchical Convolutional Neural Networks for Segmentation of Breast Tumors in MRI With Application to Radiogenomics. IEEE Trans Med Imaging 2019;38:435–47. https://doi.org/10.1109/TMI.2018.2865671.
[61] Trimpl MJ, Primakov S, Lambin P, Stride EPJ, Vallis KA, Gooding MJ. Beyond automatic medical image segmentation—the spectrum between fully manual and fully automatic delineation. Phys Med Biol 2022;67:12TR01. https://doi.org/10.1088/1361-6560/AC6D9C.
[62] Wang S, Sun K, Wang L, Qu L, Yan F, Wang Q, et al. Breast Tumor Segmentation in DCE-MRI with Tumor Sensitive Synthesis. IEEE Trans Neural Networks Learn Syst 2023;34:4990–5001. https://doi.org/10.1109/TNNLS.2021.3129781.
[63] Khaled R, Vidal J, Vilanova JC, Martí R. A U-Net Ensemble for breast lesion segmentation in DCE MRI. Comput Biol Med 2022;140:105093.

https://doi.org/10.1016/J.COMPBIOMED.2021.105093.
[64] Yue W, Zhang H, Zhou J, Li G, Tang Z, Sun Z, et al. Deep learning-based automatic segmentation for size and volumetric measurement of breast cancer on magnetic resonance imaging. Front Oncol 2022;12. https://doi.org/10.3389/FONC.2022.984626.
[65] Saarinen T, Savukoski SM, Pesonen P, Vaaramo E, Laitinen J, Varanka-Ruuska T, et al. Climacteric status at age 46 is associated with poorer work ability, lower 2-year participation in working life, and a higher 7-year disability retirement rate: a Northern Finland Birth Cohort 1966 study. Menopause 2024;31:275–81. https://doi.org/10.1097/GME.0000000000002327.
[66] Residual Cancer Burden Calculator n.d. https://www3.mdanderson.org/app/medcalc/index.cfm?pagename=jsconvert3 (accessed November 21, 2024).
[67] Zhang J, Cui Z, Shi Z, Li Z, Liu Z, Shen D. A robust and efficient AI assistant for breast tumor segmentation from DCE-MRI via a spatial-temporal framework. Patterns 2023;4:100826. https://doi.org/10.1016/j.patter.2023.100826.
[68] Garrucho L, Reidel C-A, Kushibar K, Joshi S, Osuala R, Tsirikoglou A, et al. MAMA-MIA: A Large-Scale Multi-Center Breast Cancer DCE-MRI Benchmark Dataset with Expert Segmentations 2024.
[69] Zapaishchykova A, Tak D, Boyd A, Ye Z, Aerts HJWL, Kann BH. SegmentationReview: A Slicer3D extension for fast review of AI-generated segmentations. Softw Impacts 2023;17:100536. https://doi.org/10.1016/j.simpa.2023.100536.
[70] 3D Slicer image computing platform | 3D Slicer n.d. https://www.slicer.org/ (accessed November 21, 2024).
[71] Hatamikia S, George G, Schwarzhans F, Mahbod A, Woitek R. Breast MRI radiomics and machine learning-based predictions of response to neoadjuvant chemotherapy – How are they affected by variations in tumor delineation? Comput Struct Biotechnol J 2024;23:52. https://doi.org/10.1016/J.CSBJ.2023.11.016.
[72] Carvalho ED, Veloso Silva RR, Mathew MJ, Duarte Araujo FH, De Carvalho Filho AO. Tumor Segmentation in Breast DCE- MRI Slice Using Deep Learning Methods. 2021 IEEE Symp. Comput. Commun., 2021, p. 1–6. https://doi.org/10.1109/ISCC53001.2021.9631444.
[73] Zhao X, Liao Y, Xie J, He X, Zhang S, Wang G, et al. BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification. Comput Biol Med 2023;164:107255. https://doi.org/10.1016/J.COMPBIOMED.2023.107255.
[74] Zhou Y, Aryal S, Bouadjenek MR. A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms 2024.
[75] Kim M, Hwang K-B. An empirical evaluation of sampling methods for the classification of imbalanced data. PLoS One 2022;17:e0271260. https://doi.org/10.1371/journal.pone.0271260.
[76] Miot HA. Correlation analysis in clinical and experimental studies. J Vasc Bras 2018;17:275. https://doi.org/10.1590/1677-5449.174118.
[77] Qian B, Yang J, Zhou J, Hu L, Zhang S, Ren M, et al. Individualized model for predicting pathological complete response to neoadjuvant chemotherapy in patients with breast cancer: A multicenter study. Front Endocrinol (Lausanne) 2022;13:955250. https://doi.org/10.3389/fendo.2022.955250.
[78] Kim H-Y. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restor Dent Endod 2017;42:152–5. https://doi.org/10.5395/rde.2017.42.2.152.
[79] Jafari M, Ansari-Pour N. Why, When and How to Adjust Your P Values? Cell J 2019;20:604–7. https://doi.org/10.22074/cellj.2019.5992.

[80] Masood S. Neoadjuvant chemotherapy in breast cancers. Womens Health (Lond Engl) 2016;12:480–91. https://doi.org/10.1177/1745505716677139.
[81] Michishita S, Kim SJ, Shimazu K, Sota Y, Naoi Y, Maruyama N, et al. Prediction of pathological complete response to neoadjuvant chemotherapy by magnetic resonance imaging in breast cancer patients. The Breast 2015;24:159–65. https://doi.org/10.1016/J.BREAST.2015.01.001.
[82] Jung JJ, Kim EK, Kang E, Kim JH, Kim SH, Suh KJ, et al. Development and External Validation of a Machine Learning Model to Predict Pathological Complete Response After Neoadjuvant Chemotherapy in Breast Cancer. J Breast Cancer 2023;26:353. https://doi.org/10.4048/JBC.2023.26.E14.
[83] Fan M, Wu G, Cheng H, Zhang J, Shao G, Li L. Radiomic analysis of DCE-MRI for prediction of response to neoadjuvant chemotherapy in breast cancer patients. Eur J Radiol 2017;94:140–7. https://doi.org/https://doi.org/10.1016/j.ejrad.2017.06.019.
[84] Peng Y, Cheng Z, Gong C, Zheng C, Zhang X, Wu Z, et al. Pretreatment DCE-MRI-Based Deep Learning Outperforms Radiomics Analysis in Predicting Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer. Front Oncol 2022;12:846775. https://doi.org/10.3389/FONC.2022.846775/FULL.
[85] Gao Y, Ventura-Diaz S, Wang X, He M, Xu Z, Weir A, et al. An explainable longitudinal multi-modal fusion model for predicting neoadjuvant therapy response in women with breast cancer. Nat Commun 2024;15:9613. https://doi.org/10.1038/s41467-024-53450-8.
[86] Demircioğlu A. Deep Features from Pretrained Networks Do Not Outperform Hand-Crafted Features in Radiomics. Diagnostics 2023;13. https://doi.org/10.3390/diagnostics13203266.
[87] Li JH, Guo SX, Ma RL, He J, Zhang XH, Rui DS, et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Med Res Methodol 2024;24:41. https://doi.org/10.1186/S12874-024-02173-X.

Appendix A – From binair pathological respons to RCB score: 'The RCB translation method'

For calculating the RCB score, the following parameters are required :

1. Primary tumor bed area
2. Overall cancer cellularity (percentage of area)
3. Percentage of cancer that is in situ
4. Number of positive lymph nodes
5. Diameter of the largest metastasis

All pathology reports for the included patients of this study were reviewed to identify these necessary parameters, as they are sometimes mentioned in the report but without an RCB score calculation. In this way, it was possible to retrospectively establish an RCB score for a subset of patients. If a parameter is reported as a range (e.g., 10-50%), the RCB score is calculated using both the lower bound (in this case, 10%) and the upper bound (in this case, 50%) of the range. This results in two possible RCB scores for a single patient. If these scores are the same, the RCB score can be included in the database. If the scores differ, the pathologist must review the patient's case to determine the accurate score.

To validate the accuracy of this RCB translation method, tissue from 12 patients were collected so the pathologist could evaluate them using the standard scoring method. Simultaneously, the new RCB translation method was applied to these same patients. The results were as follows:

- For 9 of the 12 patients, the same RCB score was obtained with both the standard method and the new RCB translation method.
- The parameter 'percentage of cancer that is in situ' was found to be the most frequently ambiguous in the pathology reports. Among these 12 patients, the average value for this parameter was 15%. Consequently, it was decided to use a range of 1-30% when this parameter is unclear in the pathology report.

Appendix B - Radiomics 3D shape features

Table 10 *3D shape features, calculated with pyradiomics*

Feature	Description
Elongation	Calculated as the ratio of the largest to the second largest principal component axes. This feature helps in identifying long, stretched shapes versus more compact ones
Flatness	Calculated as the ratio of the smallest principal axis to the largest principal axis of the object. It is particularly useful for understanding the relative thickness or thinness of an object.
Minor axis length	The second largest axis length and is calculated using the second largest principal component.
Least axis length	The smallest axis length and is calculated using the smallest principal component. This feature essentially measures the minimum spread or extent of the region in any dimension.
Major axis length	The largest axis length and is calculated using the largest principal component. This axis is in the direction in which the object extends the most.
Maximum 2D diameter column	Defined as the largest distance in the row-slice plane
Maximum 2D diameter row	Defined as the largest distance in the column-slice plane
Maximum 2D diameter slice	Defined as the largest distance in the row-column plane
Maximum 3D diameter	The longest straight-line distance between any two points of a 3D object.
Mesh volume	The 3D space enclosed by a surface mesh
Sphericity	A measure of the roundness of the shape of the segmentation relative to a sphere. The value of 1 indicates a perfect sphere.
Surface area	Total area of the outer boundary of a segmentation represented by its surface mesh
Surface volume ratio	Calculated by dividing the total surface area of the segmentation by the volume of the segmentation
Voxel volume	Volume of a 3D pixel

Appendix C – Imputation missing radiological features

Introduction

There were patients with one missing MRI (n=33), resulting in missing radiological features for these patients. Several strategies are available for dealing with missing values [74], including:

1. Deletion of patients: removing patients with missing values can lead to a smaller dataset, but it preserves the integrity of the data.
2. Imputation with simple statistical measures: a common approach is to impute missing values using simple statistical measures, such as the mean or median. This method assumes that the imputed value is representative of the central tendency of the data.
3. Machine learning techniques: another approach is the use of machine learning techniques to predict missing values. This method offers the potential to identify patterns and relationships within the data, potentially leading to more accurate imputations compared to traditional statistical imputations.

Materials and method

During this study, six imputations methods were tested. First, patients with an actual missing MRI were removed from the dataset (n = 33). The remaining dataset (n= 258) was split into a training set (60%), a validation set (20%) and a test set (20%). In both the validation and test sets, 20% of the data was randomly removed. The original pattern of missing data was preserved: either all variables from MRI time point 1 were missing, or all variables from MRI time point 2, or all variables from MRI time point 3. Subsequently, several methods were tested to estimate these created missing values [87] :

1. Statistical parameters: mean, median, and most frequent
2. Machine learning models: KNN, RF, and Bayesian. For hyperparameter optimization, the best set of hyperparameters was selected based on evaluation using the validation set. The following parameters were used:

- KNN: number of neighboring data points (K) to take into account for the prediction equals 1, 2, 3, 5, 7, 10.
- RF: the number of decision trees equals 10, 20, and 30; the tree depths equals 10, 20, and 30.
- Bayesian: the limit of the number of iterations for optimal settings equals 10, 20, and 30.

For each imputed missing value, the absolute error was calculated. This absolute error is divided by the mean of the corresponding radiological feature. The mean percentage error was determined for each radiological feature.

Results

In Figure 21, the mean percentage errors from the six imputation methods for each radiological feature are visible. There were three DCE-MRI time points per patient, each with 14 radiological features. The results indicate that for the majority of radiological features, the mean percentage error exceeds 10% across all imputation methods. However, there were exceptions: the median, most frequent, and RF imputation methods each have one radiological feature with a mean percentage error below 10%. Specifically, for the median and RF methods, this feature was the delta elongation feature, while for the most frequent method, it was the delta feature sphericity.

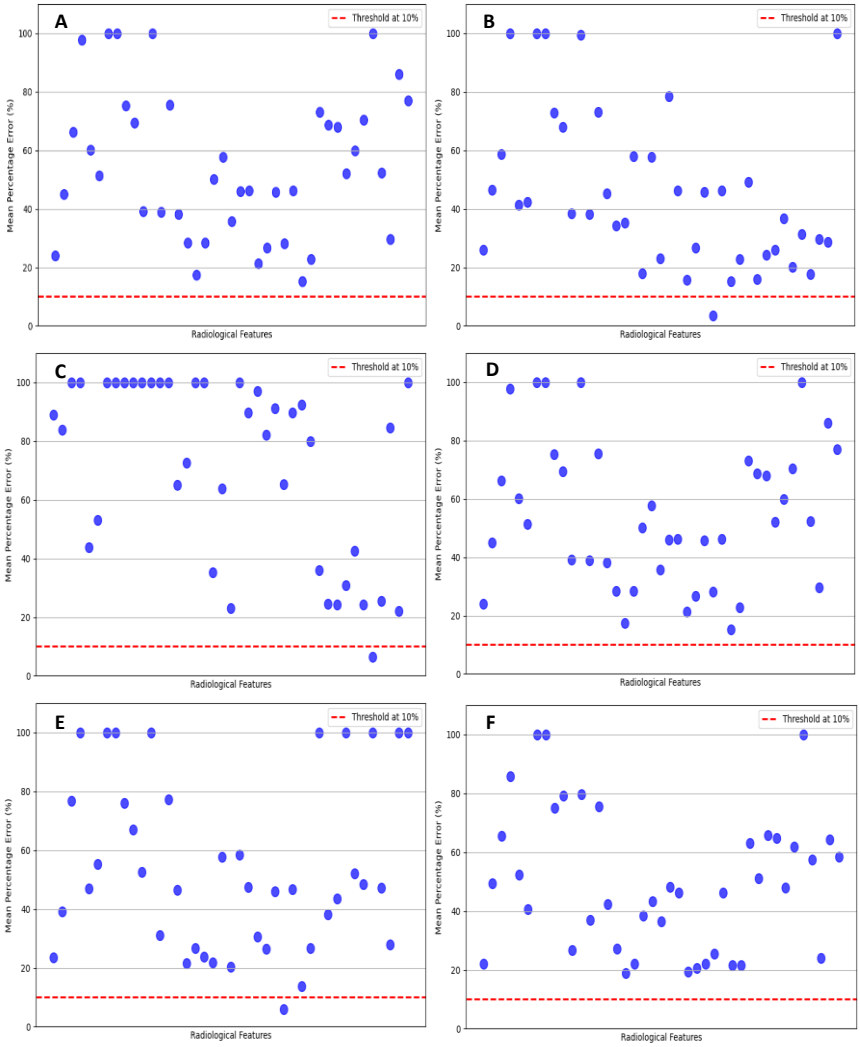

Figure 21 *Mean percentage error scatterplots for each radiological feature. There were three DCE-MRI time points per patient, each with 14 radiological features. The red line is a threshold line of 10%. A: imputation with mean, B: imputation with median, C: imputation with most frequent, D: imputation with KNN prediction, E: imputation with RF prediction, F: imputation with Bayesian prediction.*

Discussion

In this study, the dataset was too heterogeneous to justify reliable imputations using mean, median or most frequent values. The central tendency can be significantly influenced by outliers or extreme values, leading to imputed values that do not reflect the actual variability within the dataset. In this case, using imputation with a statistical measure could result in distortions, leading to incorrect conclusions. Additionally, attempts were made to predict the missing values using various machine learning models. These results were insufficient, likely due to the lack of sufficient data to enable the models to learn the correct patterns. It is expected that expanding the dataset, further optimization of hyperparameters and experimentation with different algorithms are necessary to improve the performance of the imputation of missing values with machine learning models.

Conclusion

Since no imputation method was able to estimate the majority of the radiological variables with a margin of error of less than 10%, it was decided to remove patients with missing radiological variables.

Appendix D - Spearman correlation coefficient clinical data

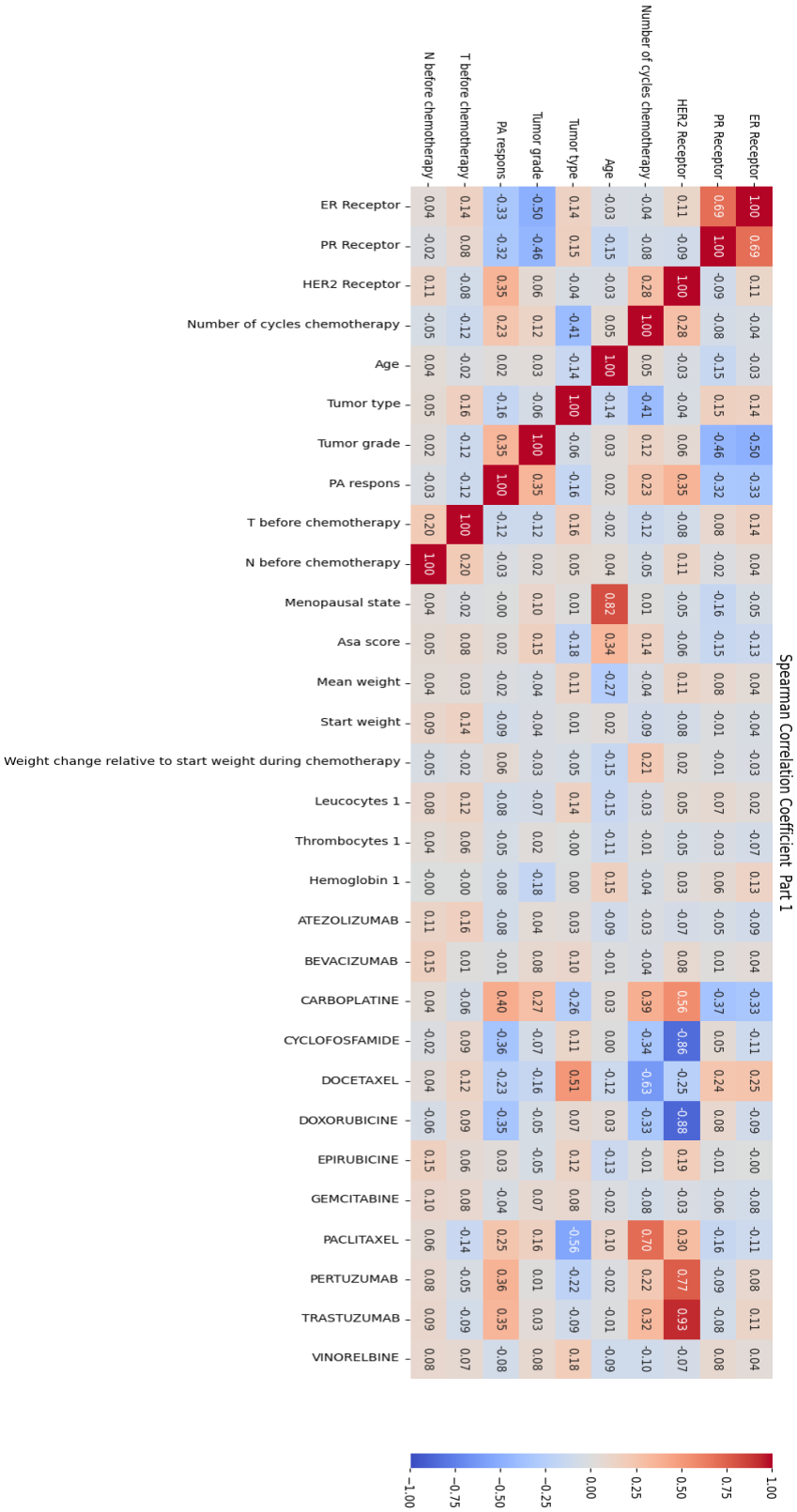

Figure 22 *Heatmap SCC Clinical Data, Part I.*

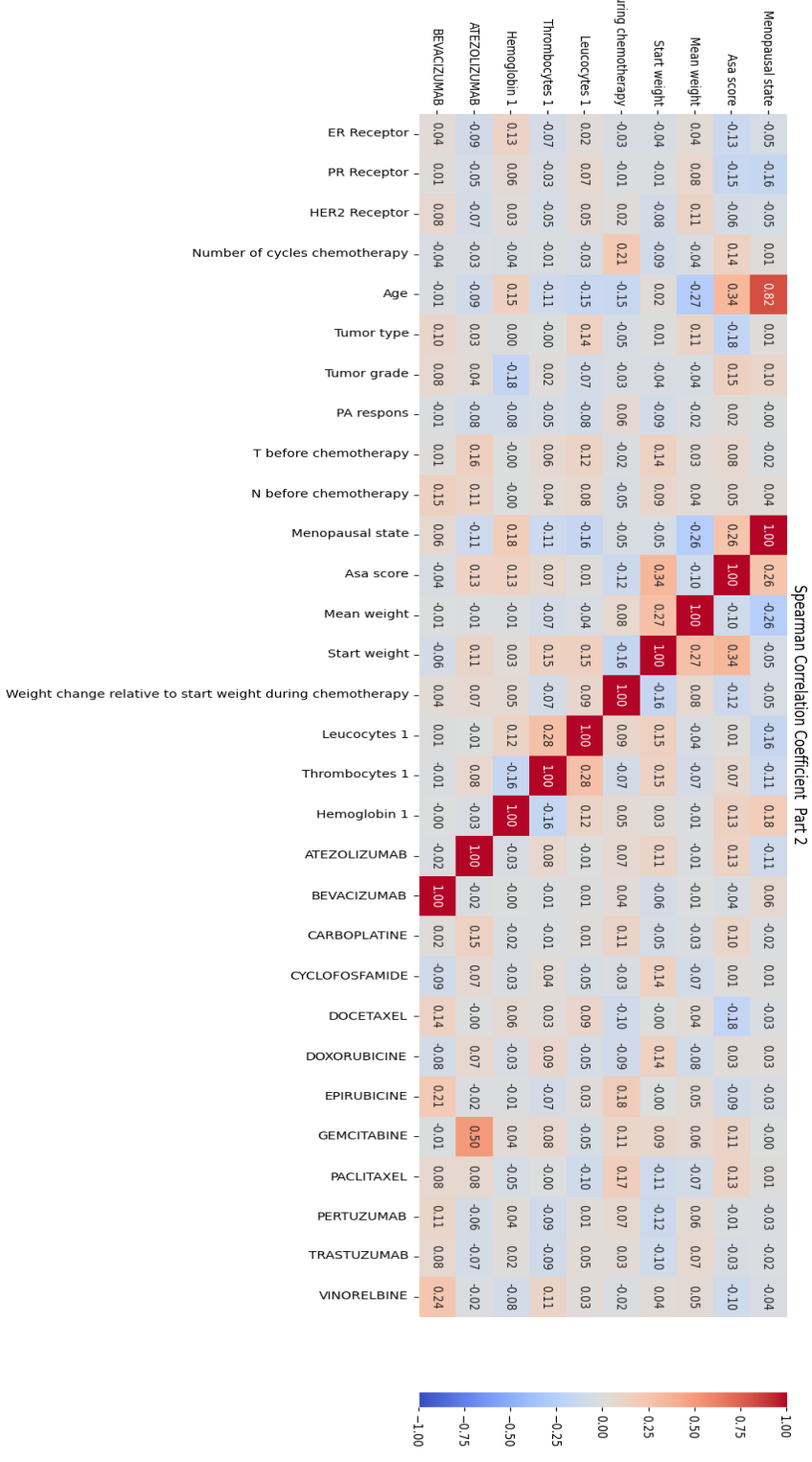

Figure 23 *Heatmap SCC Clinical Data, Part II.*

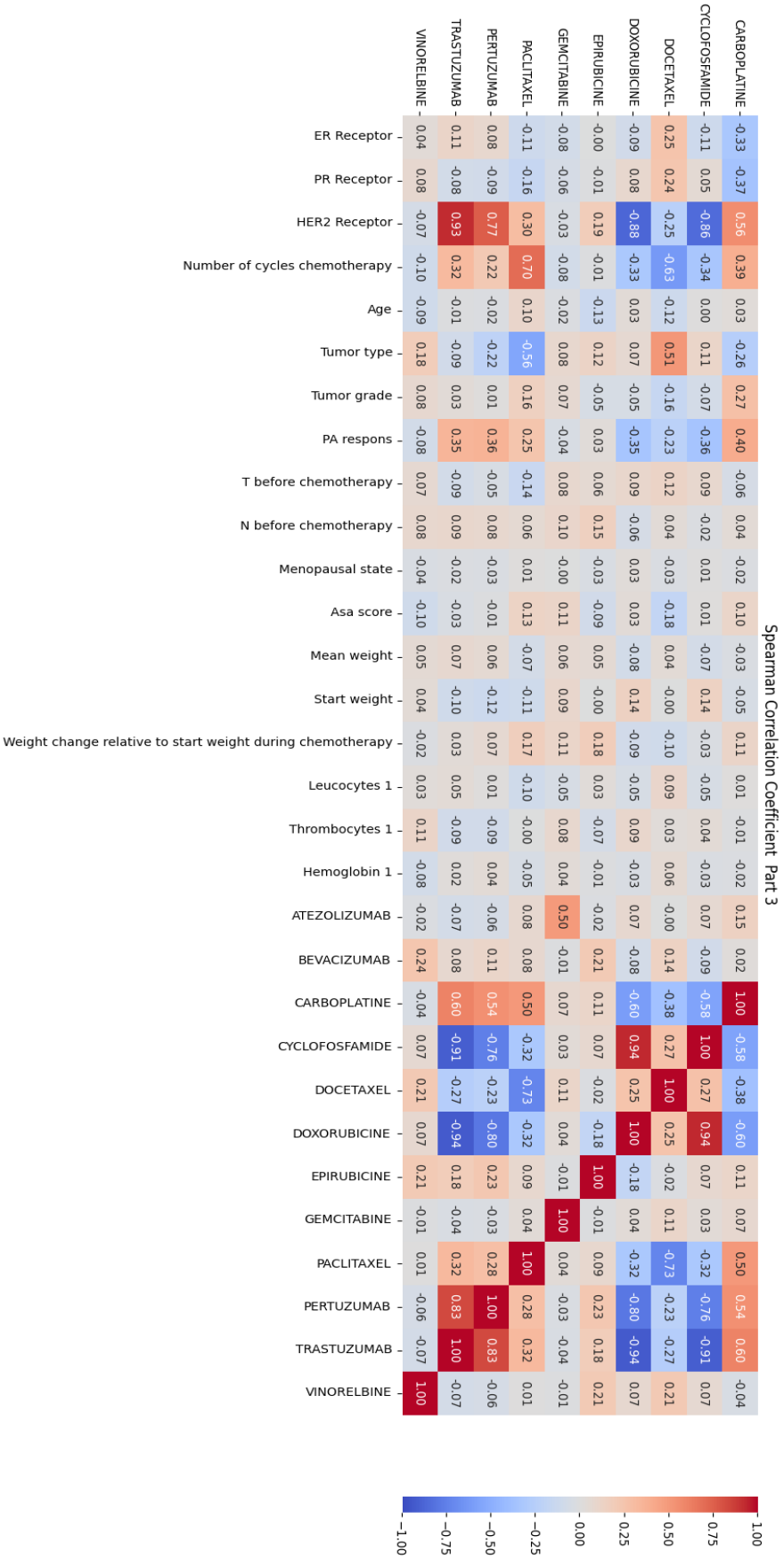

Figure 24 *Heatmap SCC Clinical Data, Part III.*

Appendix E - Spearman correlation coefficient radiological data

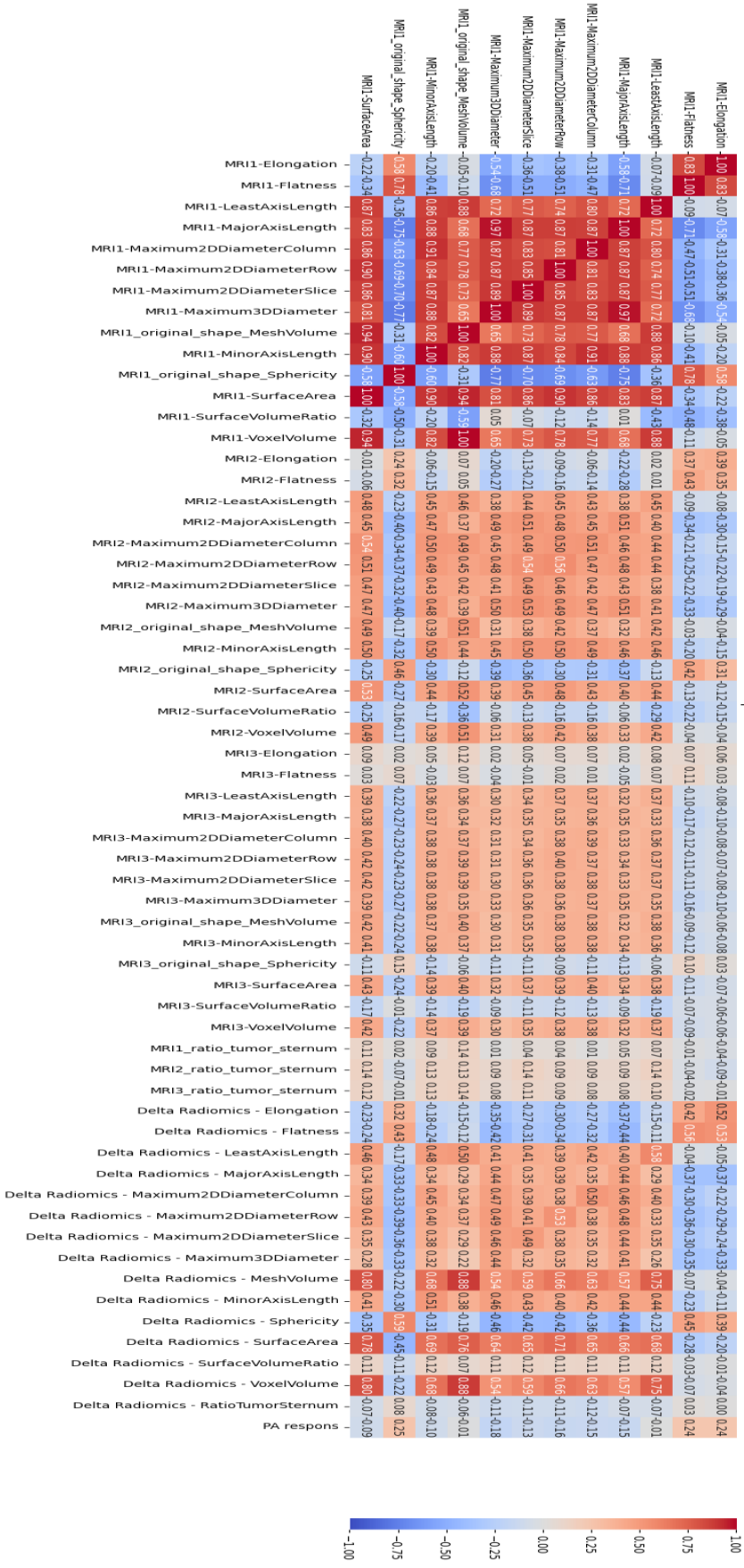

Figure 25 *Heatmap SCC Radiological Data, Part I.*

Figure 26 *Heatmap SCC Radiological Data, Part II.*

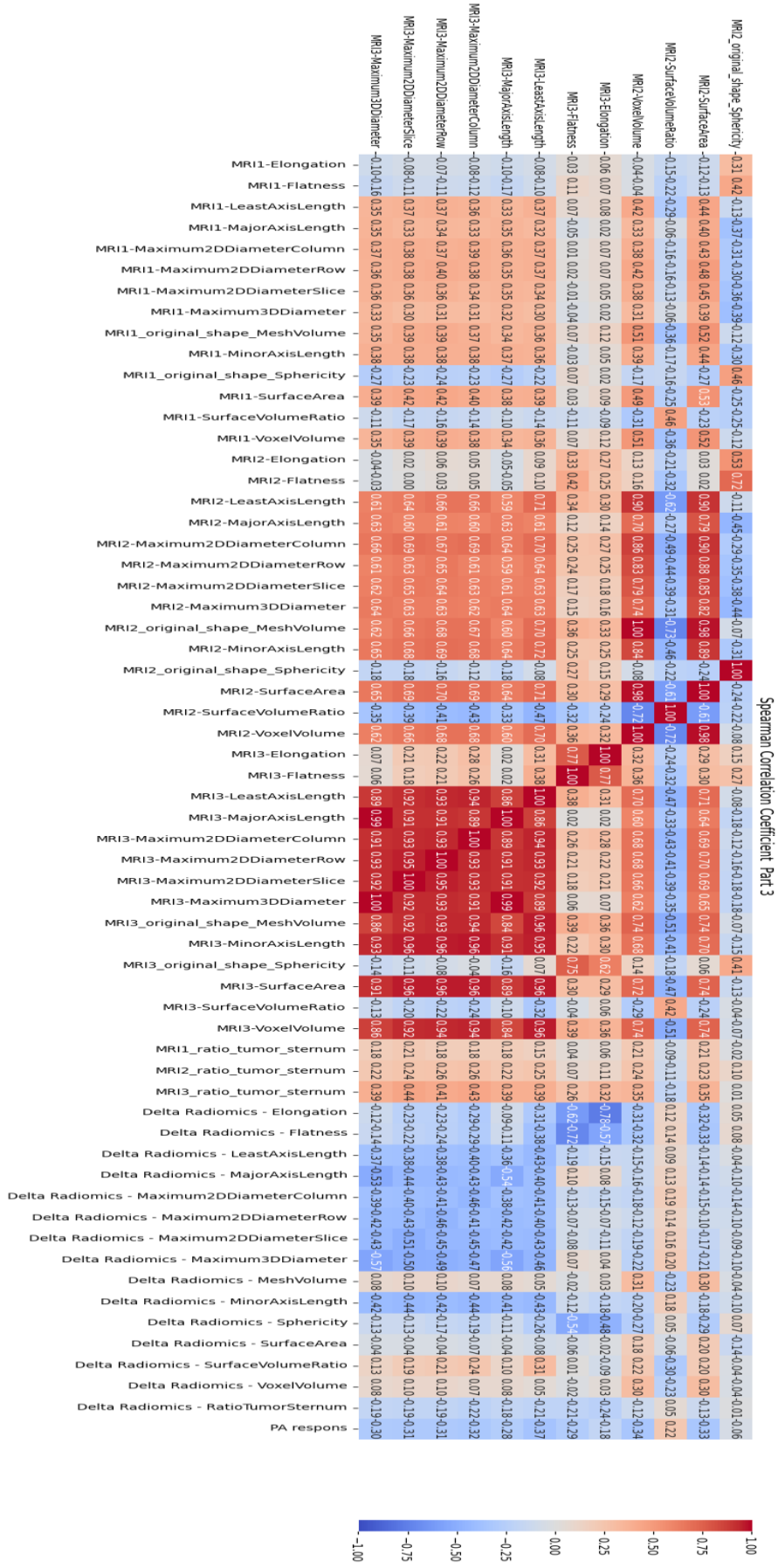

Figure 27 *Heatmap SCC Radiological Data, Part III.*

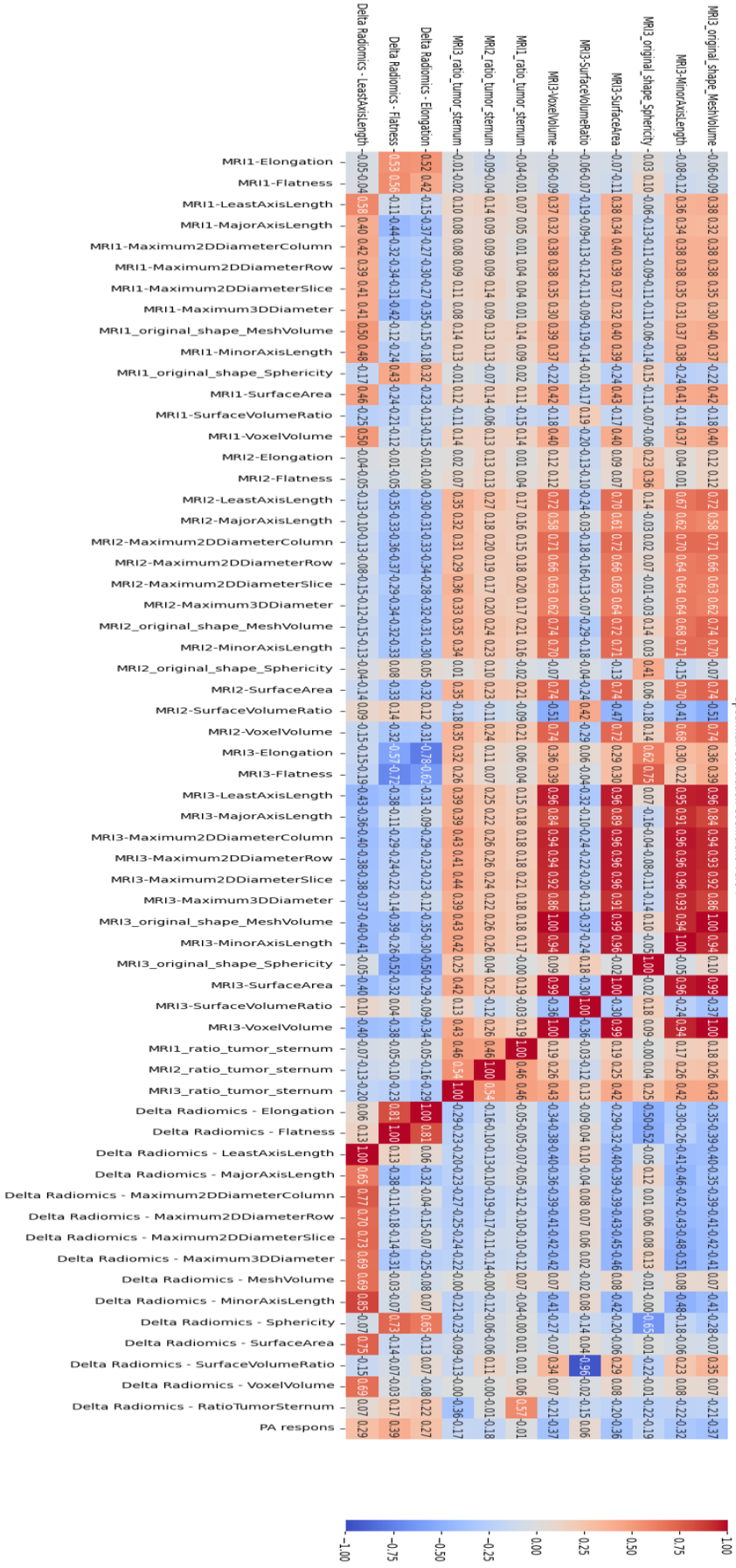

Figure 28 *Heatmap SCC Radiological Data, Part IV.*

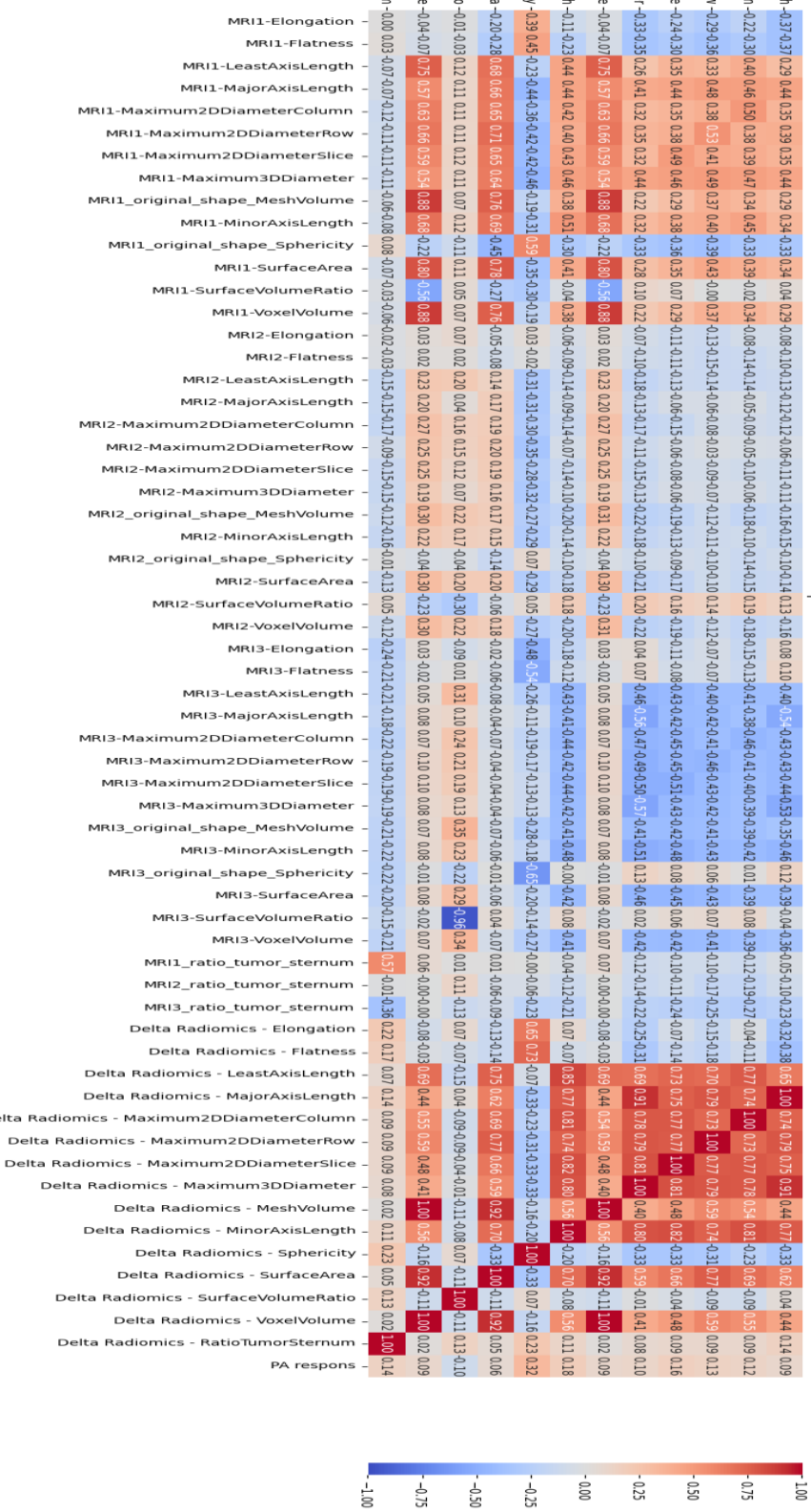

Figure 29 *Heatmap SCC Radiological Data, Part V.*

Appendix F - Spearman correlation coefficient clinical data combined with radiological data

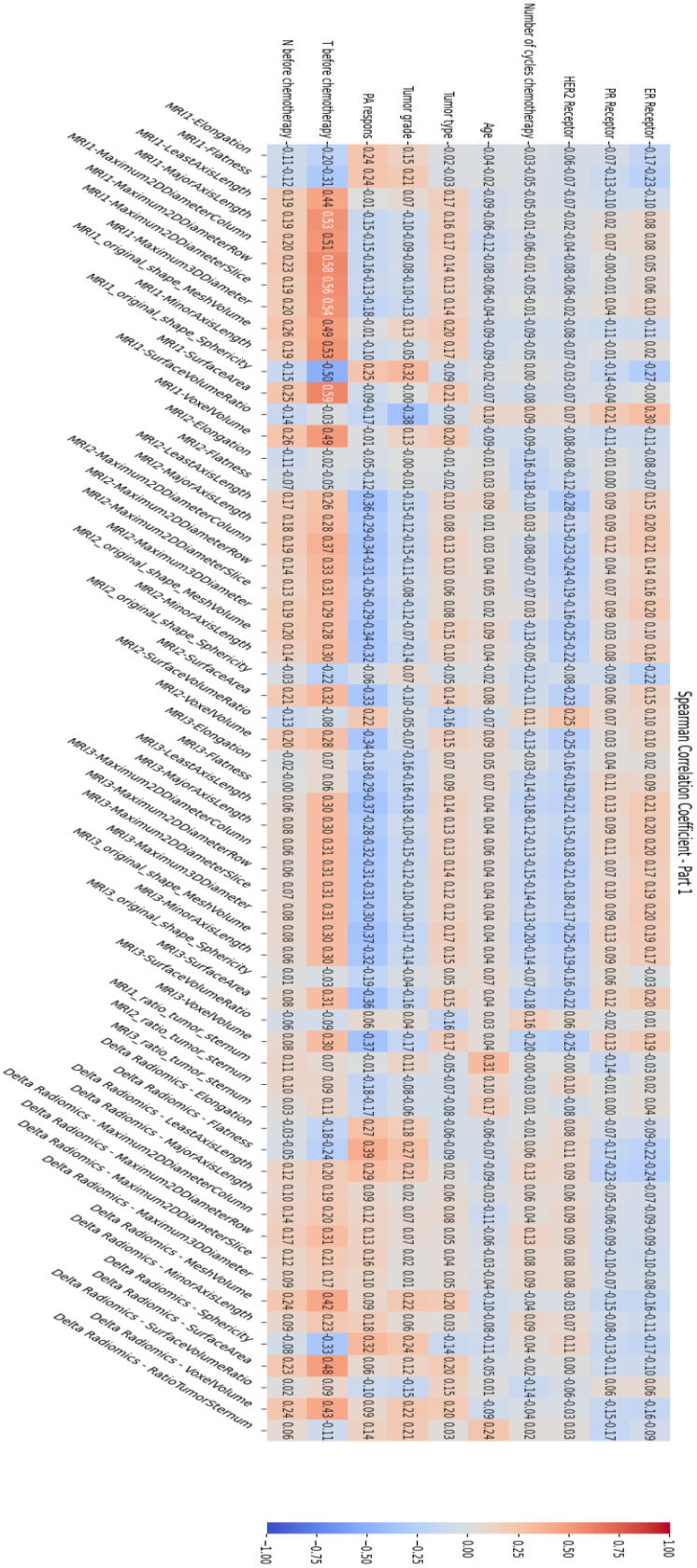

Figure 30 *Heatmap SCC Clinical and Radiological Data, Part I.*

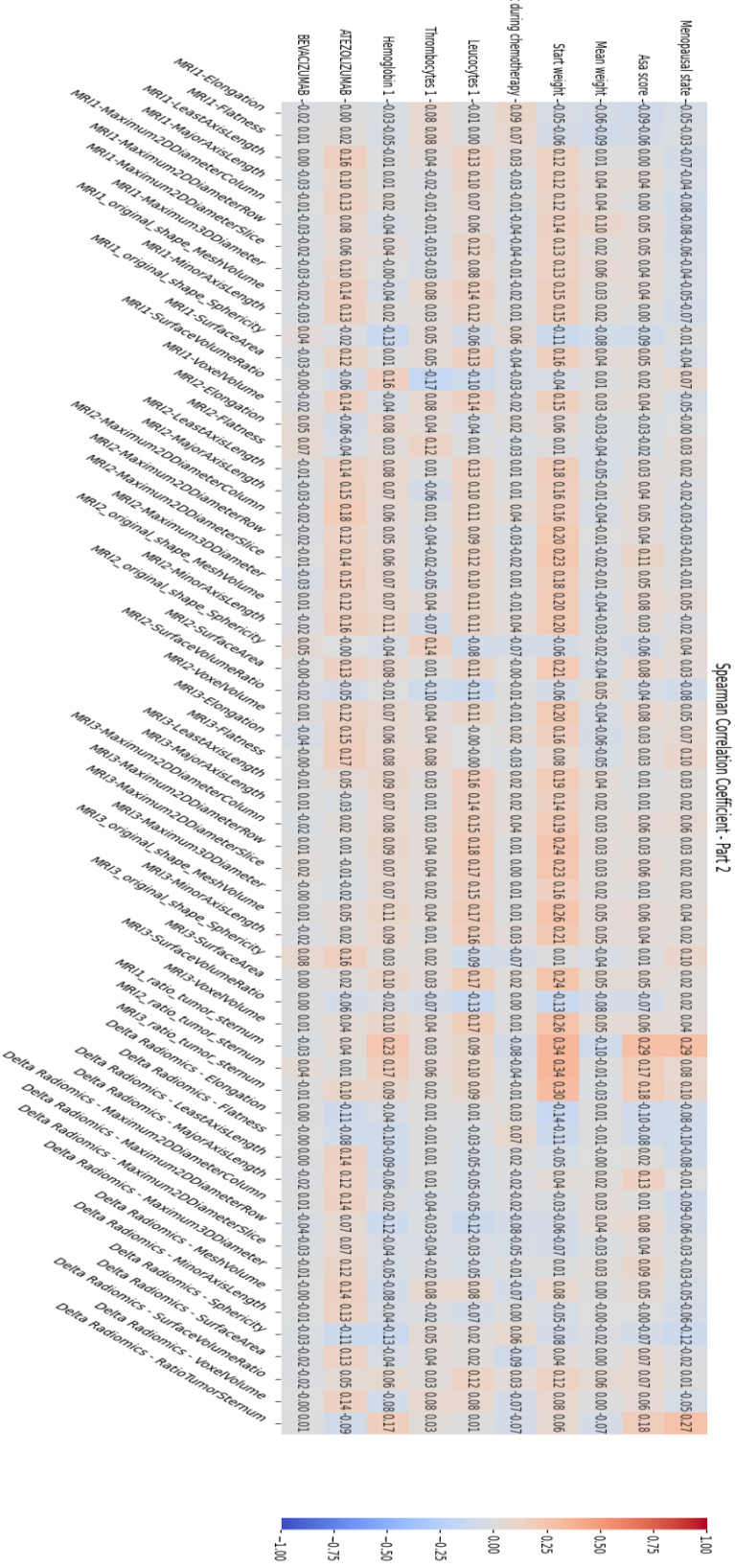

Figure 31 *Heatmap SCC Clinical and Radiological Data, Part II.*

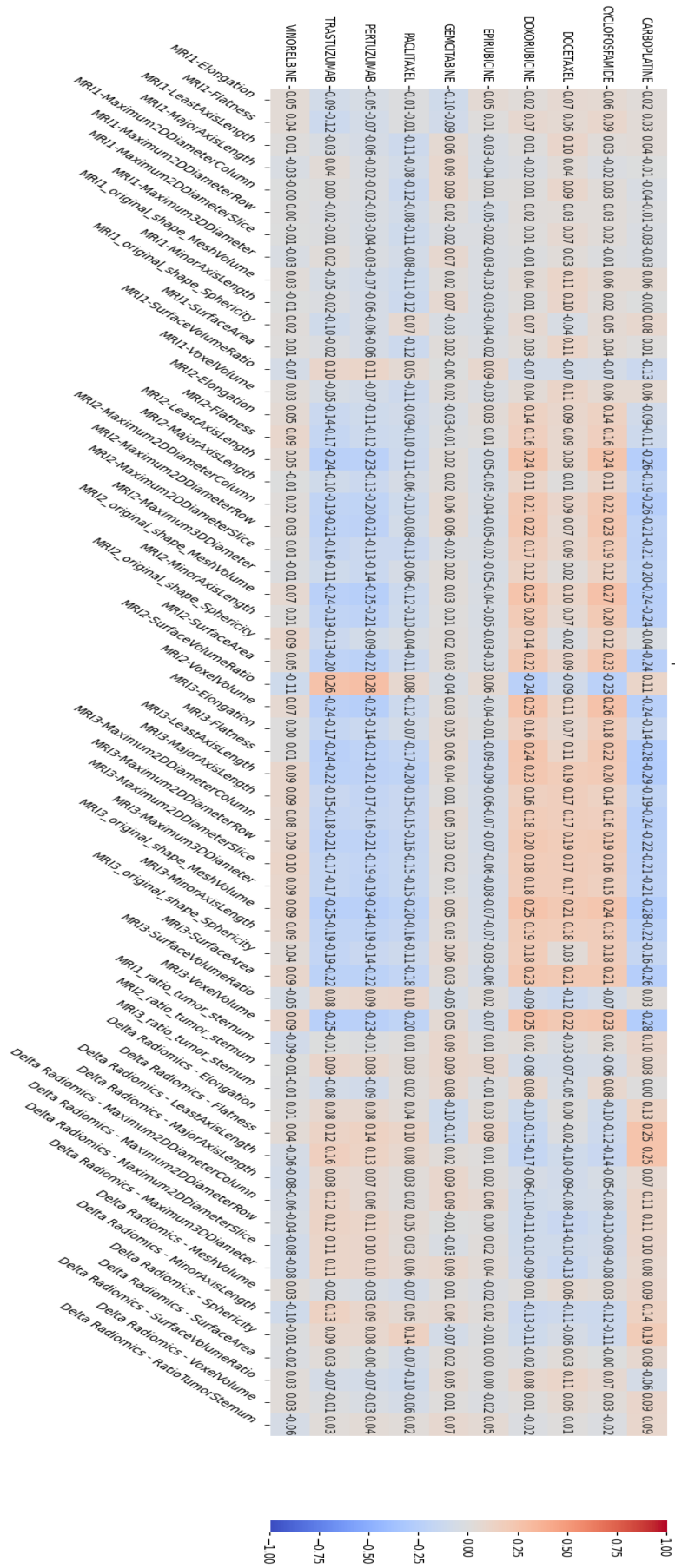

Figure 32 *Heatmap SCC Clinical and Radiological Data, Part III.*

Appendix G – Results Binary Classification Model

Table 11 *Results of binary classification model*

Model	Data	Sensitivity	Specificity	AUC
1	Clinical	0.31	0.83	0.57
1	Radiological	0.44	0.92	0.68
1	Clinical and Radiological	0.44	0.83	0.64
2	Clinical	0.25	0.89	0.57
2	Radiological	0.31	0.92	0.61
2	Clinical and Radiological	0.75	0.83	0.79
3	Clinical	0.31	0.83	0.57
3	Radiological	0.50	0.86	0.68
3	Clinical and Radiological	0.69	0.72	0.70
4	Clinical	0.31	0.83	0.57
4	Radiological	0.25	1.00	0.62
4	Clinical and Radiological	0.38	0.83	0.60
5	Clinical	0.31	0.83	0.57
5	Radiological	0.44	0.89	0.66
5	Clinical and Radiological	0.56	0.83	0.70
6	Clinical	0.44	0.75	0.59
6	Radiological	0.44	0.89	0.66
6	Clinical and Radiological	0.56	0.78	0.67
7	Clinical	0.44	0.75	0.59
7	Radiological	0.50	0.78	0.64
7	Clinical and Radiological	0.69	0.81	0.75
8	Clinical	0.75	0.75	0.75
8	Radiological	0.75	0.72	0.74
8	Clinical and Radiological	0.69	0.76	0.68
9	Clinical	0.75	0.75	0.75
9	Radiological	0.75	0.75	0.75
9	Clinical and Radiological	0.75	0.69	0.72

Appendix H – Results Binary Classification - Breast Cancer Subtypes

Table 12 *Results of binary classification model – Her2+ subgroup*

Model	Data	Sensitivity	Specificity	AUC
1	Clinical	1.0	0.25	0.62
1	Radiological	0.75	0.75	0.75
1	Clinical and Radiological	1.0	0.50	0.75
2	Clinical	1.0	0.25	0.62
2	Radiological	0.50	1.00	0.75
2	Clinical and Radiological	1.0	0.50	0.75
3	Clinical	1.0	0.25	0.62
3	Radiological	0.75	0.50	0.62
3	Clinical and Radiological	1.0	0.25	0.62
4	Clinical	1.0	0.25	0.62
4	Radiological	0.50	1.00	0.75
4	Clinical and Radiological	0.75	0.25	0.50
5	Clinical	1.0	0.25	0.62
5	Radiological	0.75	0.50	0.62
5	Clinical and Radiological	1.0	0.50	0.75
6	Clinical	0.75	0.50	0.62
6	Radiological	0.25	1.00	0.62
6	Clinical and Radiological	1.0	0.75	0.88
7	Clinical	0.75	0.50	0.62
7	Radiological	0.50	0.75	0.62
7	Clinical and Radiological	1.0	0.75	0.88
8	Clinical	1.0	0.50	0.75
8	Radiological	0.75	0.50	0.62
8	Clinical and Radiological	1.0	0.25	0.62
9	Clinical	1.0	0.50	0.75
9	Radiological	0.75	0.50	0.62
9	Clinical and Radiological	1.0	0.50	0.50

Table 13 *Results of binary classification model – TNBC subgroup*

Model	Data	Sensitivity	Specificity	AUC
1	Clinical	0.12	0.80	0.46
1	Radiological	0.25	0.80	0.53
1	Clinical and Radiological	0.38	0.70	0.54
2	Clinical	0	1.0	0.50
2	Radiological	0.25	1.0	0.62
2	Clinical and Radiological	0.88	0.80	0.84
3	Clinical	0.12	0.80	0.46
3	Radiological	0.25	0.80	0.53
3	Clinical and Radiological	0.62	0.50	0.56
4	Clinical	0.12	0.80	0.46
4	Radiological	0.12	1.0	0.56
4	Clinical and Radiological	0.38	0.80	0.59
5	Clinical	0.12	0.80	0.46
5	Radiological	0.25	0.80	0.53
5	Clinical and Radiological	0.38	0.70	0.54
6	Clinical	0.50	0.40	0.45
6	Radiological	0.62	0.70	0.66
6	Clinical and Radiological	0.50	0.50	0.50
7	Clinical	0.50	0.40	0.45
7	Radiological	0.62	0.50	0.56
7	Clinical and Radiological	0.75	0.60	0.68
8	Clinical	1.00	0.30	0.65
8	Radiological	0.75	0.50	0.62
8	Clinical and Radiological	1.00	0.25	0.62
9	Clinical	1.00	0.30	0.65
9	Radiological	0.75	0.50	0.62
9	Clinical and Radiological	0.75	0.40	0.57

Appendix I – Results RCB Classification

Table 14 *Results of RCB classification*

Model	Data	Accuracy	Cohen's kappa
1	Clinical	0.56	0.30
1	Radiological	0.47	0.13
1	Clinical and Radiological	0.61	0.38
2	Clinical	0.58	0.34
2	Radiological	0.64	0.41
2	Clinical and Radiological	0.69	0.51
3	Clinical	0.56	0.29
3	Radiological	0.44	0.14
3	Clinical and Radiological	0.53	0.26
4	Clinical	0.56	0.29
4	Radiological	0.58	0.32
4	Clinical and Radiological	0.56	0.29
5	Clinical	0.56	0.30
5	Radiological	0.56	0.29
5	Clinical and Radiological	0.61	0.38
6	Clinical	0.53	0.29
6	Radiological	0.44	0.10
6	Clinical and Radiological	0.67	0.49
7	Clinical	0.53	0.29
7	Radiological	0.50	0.24
7	Clinical and Radiological	0.56	0.32

Appendix J – Results RCB-0 Classification

Table 15 *Results of RCB classification – RCB-0*

Model	Data	Sensitivity	Specificity	AUC
1	Clinical	0.88	0.70	0.79
1	Radiological	0.50	0.70	0.60
1	Clinical and Radiological	0.88	0.70	0.79
2	Clinical	0.94	0.70	0.82
2	Radiological	0.81	0.65	0.73
2	Clinical and Radiological	0.81	0.80	0.81
3	Clinical	1.00	0.60	0.80
3	Radiological	0.50	0.70	0.60
3	Clinical and Radiological	0.75	0.65	0.70
4	Clinical	1.00	0.60	0.80
4	Radiological	0.75	0.60	0.68
4	Clinical and Radiological	0.94	0.60	0.77
5	Clinical	0.88	0.70	0.79
5	Radiological	0.56	0.70	0.63
5	Clinical and Radiological	0.88	0.70	0.79
6	Clinical	0.81	0.80	0.81
6	Radiological	0.56	0.50	0.53
6	Clinical and Radiological	0.81	0.75	0.78
7	Clinical	0.81	0.80	0.81
7	Radiological	0.50	0.75	0.62
7	Clinical and Radiological	0.69	0.70	0.69

Appendix K – Confusion Matrix RCB Classification

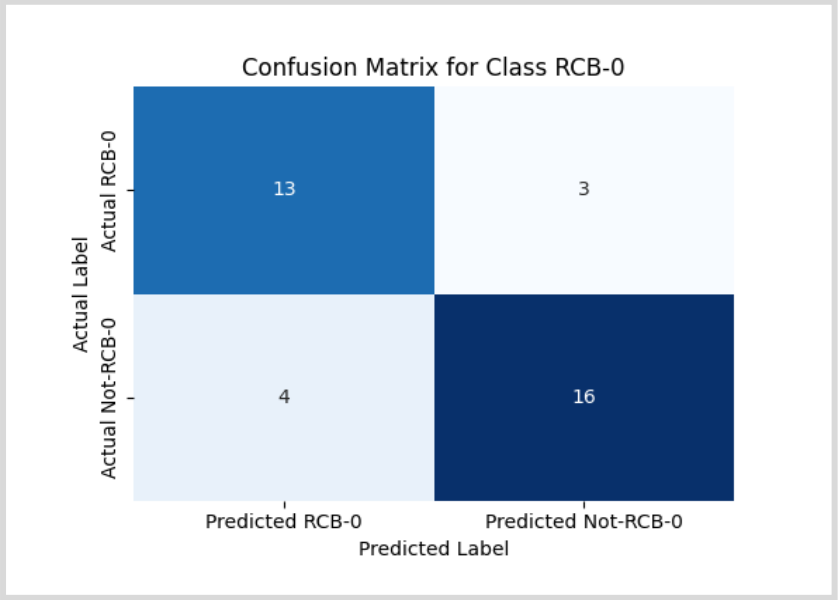

Figure 33 *Confusion matrix for RCB-0.*