applied sciences

MDPI

Review

Unsupervised Learning Techniques for Breast Lesion Segmentation on MRI Images: Are We Ready for Automation?

Marina Fedon Vocaturo [1,2], Luisa Altabella [1,*], Giuseppe Cardano [3], Stefania Montemezzi [3] and Carlo Cavedon [1]

[1] Medical Physics Unit, Department of Pathology and Diagnostics, Azienda Ospedaliera Universitaria Integrata, P.le Stefani 1, 37126 Verona, Italy
[2] Specialization School in Medical Physics, University of Padova, Via Marzolo 8, 35131 Padova, Italy
[3] Department of Pathology and Diagnostics, Radiology Unit, Azienda Ospedaliera Universitaria Integrata, P.le Stefani 1, 37126 Verona, Italy
* Correspondence: luisa.altabella@aovr.veneto.it

Abstract: In the era of precision medicine, increasing importance is given to machine learning (ML) applications. In breast cancer, advanced analyses, such as the radiomic process, characterise tumours and predict therapy responses. Breast magnetic resonance imaging (MRI) plays a key role in screening, staging, and treatment monitoring. Lesion segmentation on MRI is essential both to assess tumour growth and as a baseline for radiomic feature extraction. Manual segmentation is time-consuming and prone to inter-operator variability, limiting access to large labelled datasets and robust analyses. The use of ML for breast lesion segmentation on MRI has been investigated through a systematic review of PubMed, exploring studies published over the last 10 years. Results are compared in terms of performance, primarily using the Dice score. Early unsupervised methods achieved a mean Dice score of ∼0.75, surpassing traditional supervised methods (∼0.70). In contrast, deep learning (DL) approaches based on U-Net achieved higher average scores of 0.79. Further customised supervised DL approaches reached a mean Dice score of ∼0.83. However, there is still a gap in research on unsupervised DL techniques, which could help reduce bias and human variability. Future work may also explore multiparametric and multitechnique data, integrating more representative samples, including non-mass lesions.

Keywords: machine learning; deep learning; segmentation; breast lesion segmentation; breast magnetic resonance imaging; breast cancer

check for updates

Academic Editor: Cosimo Nardi

Received: 27 December 2024
Revised: 19 February 2025
Accepted: 21 February 2025
Published: 24 February 2025

Citation: Fedon Vocaturo, M.; Altabella, L.; Cardano, G.; Montemezzi, S.; Cavedon, C. Unsupervised Learning Techniques for Breast Lesion Segmentation on MRI Images: Are We Ready for Automation? *Appl. Sci.* **2025**, *15*, 2401. https://doi.org/10.3390/app15052401

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Every year, the global burden of cancer approaches nearly 20 million new cases, resulting in more than 9 million fatalities, according to the Global Cancer Statistics (GLOBOCAN 2022, https://gco.iarc.who.int/today, accessed on 2 November 2024). Among all cancer types, breast cancer is one of the most prevalent, ranking second in incidence after lung cancer, with almost 2.3 million cases annually. In terms of mortality, it ranks fourth, following lung, colorectal, and liver cancer, causing over 666,000 deaths each year [1]. Breast cancer exhibits high variability in genotype, phenotype, and treatment responses. Investigations into genetic expression have enabled the categorisation of breast cancer into a heterogeneous group of diseases, each associated with distinct genetic aberrations and characterised by unique histological features and clinical behaviours [2].

1.1. Applications of Breast MRI

To reduce the number of fatalities, it is essential to enhance both screening and treatment methods. The intrinsic diversity of breast cancer necessitates increasingly tailored

therapies to effectively lower mortality rates and improve patients' quality of life. Mammography is the most common screening procedure worldwide and has been shown to reduce mortality by approximately 19% [3]. To enhance accuracy and reduce false positives, additional imaging techniques such as MRI and ultrasound are frequently used. The role of MRI has become increasingly significant in screening, staging, and treatment monitoring, especially in cases where mammography is less effective, such as in individuals with dense breast tissue and younger women carrying BRCA gene mutations [4]. MRI is a valuable alternative because it is independent of tissue density, offers higher sensitivity, and does not involve ionising radiation. Breast MRI has emerged as a crucial tool for high-risk screening, assessing unknown primary cancers, identifying bilaterality and multicentricity, distinguishing scar tissue from local recurrences, checking implant integrity, monitoring the response to neoadjuvant chemotherapy (NAC), and determining tumour size and extent, thereby aiding in surgical planning [5]. In particular, contrast-enhanced MRI can capture detailed tumour characteristics by highlighting the tumour's neo-angiogenesis, blood vessel formation with increased permeability, which is responsible for the uptake and washout patterns of gadolinium-based contrast agents. MRI thus has interesting predictive capabilities regarding biological behaviour, as early changes in intracellular metabolism can provide valuable insights for estimating treatment response [4]. Dynamic contrast-enhanced (DCE) MRI has demonstrated potential in extracting functional parameters to assess and optimise treatment strategies for neoadjuvant chemotherapy [6]. Specifically, changes in tumour size and the apparent diffusion coefficient (ADC) have proven to be reliable predictors of pathological complete response (pCR) in NAC, which is used to reduce tumour size before surgery [7].

The basic protocol for breast MRI involves the dynamic T1-weighted (T1w) contrast-enhanced sequence. However, increasingly, a multiparametric approach is being adopted in routine practice, incorporating T2-weighted (T2w) and diffusion-weighted imaging (DWI) alongside the standard sequence. Image acquisition typically occurs in the axial plane, providing comprehensive visualisation of both breasts [8]. T1w images are taken both before and after the administration of a contrast agent. Lesion detection generally occurs 60–90 s after contrast injection, as most lesions show enhancement during this period. When fat suppression is not used, subtraction of pre- and post-contrast images is necessary to accurately identify cancerous sites. For detection purposes, two time points are usually sufficient. Differentiating between malignant and benign lesions involves analysing time-intensity curves, obtained by acquiring multiple T1w images over 5–7 min post-contrast injection. Benign lesions typically show a persistent increase in signal intensity, whereas malignant lesions often exhibit a washout or plateau pattern. T2w imaging, both with and without fat suppression, allows for the identification of cysts and detailed lesion morphology, enhancing the differentiation between benign and malignant lesions based on signal intensity. Complementing this, DWI measures the movement of water molecules and uses the ADC to further differentiate breast lesions. Malignant lesions generally exhibit low ADC values due to restricted water diffusion, which correlates with lower signal intensity [8].

Apart from malignant and benign lesions, breast lesions can be discriminated in terms of morphological features, with mass, focus/foci, and non-mass enhancements. A mass lesion is a 3D region larger than 5 mm, characterised by its shape (round, oval, lobulated, irregular), margins (smooth, irregular, spiculated), and internal enhancement (homogeneous, heterogeneous, etc.). In contrast, enhancements of size < 5 mm are considered foci and cannot be characterised morphologically. Conversely, non-mass-like enhancements have no space-occupying effects, covering an area that is neither a mass nor blood vessels. In this case, the lesion is characterised by internal enhancement, symmetry, and overall

enhancement distribution [9]. Mass lesions can often be discriminated as malignant or benign simply by considering the shape and margins, with benign lesions having a regular appearance. On the other hand, for non-mass enhancement, there is an overlap in the appearance, making the differentiation particularly challenging [10].

In the era of precision medicine, new tools capable of enhancing clinical practice are being developed. One area of growing interest is the radiomic process, which involves extracting quantitative features from medical images, converting these images into mineable data that can support decision-making [11]. Numerous studies have combined radiomics with machine learning techniques to classify lesions, improving screening procedures by distinguishing benign from malignant lesions. These studies have also focused on classifying different molecular subtypes associated with varying risk factors, treatment responses, and disease progression, as well as investigating the classification of receptor status and proliferation markers [12]. Moreover, the use of radiomics has also been investigated for predicting the response to NAC, demonstrating improvements in predictive models when integrated with radiomic features as reported by [13–15].

The radiomic process involves a series of steps: image acquisition and reconstruction, identification and segmentation of the region of interest, image preprocessing, extraction of radiomic features, dimensionality reduction, and model development. Of all these steps, segmentation is the most critical and challenging, as all subsequent features are generated based on the delineated region and are highly dependent on its accuracy [11]. This poses a significant challenge, as many tumours have indistinct borders, and variations in segmentation can undermine the reproducibility of the extracted features and subsequent analyses. In general, segmentation of malignant lesions with irregular shapes and non-mass enhancement is particularly challenging and time-consuming, even for experienced operators.

Traditionally, segmentation is performed by highly experienced radiologists who have worked in the field for many years. However, radiomic features are significantly affected by variations in operator segmentations, with only a small fraction of features remaining robust under inter-observer manual segmentation variability [16,17]. Such differences can undermine the reproducibility and robustness of the obtained results, leading to poorly generalisable models that are difficult to share and implement in clinical practice.

The need for advanced diagnostic and prognostic tools is becoming increasingly important, particularly in countries where healthcare is less accessible, as interest grows alongside the development of the artifical intelligence (AI) era. Furthermore, the use of more complex ML models demands larger datasets to effectively tune parameters and avoid overfitting. However, manual segmentation is highly time-consuming. To address both reproducibility issues and the substantial time investment required for segmentation, as well as to facilitate the creation of larger datasets, research into automated methods for segmentation has been intensifying.

The growing interest in ML techniques has already been examined in various studies [12,18–21], which have highlighted the diversity of applications for machine learning and deep learning in breast cancer, ranging from diagnostic tools to predictive models. Since all these approaches rely heavily on the accurate delineation of the region of interest (ROI), it is crucial to encourage more research focused on the standardisation of ROI depiction.

1.2. Background on Segmentation

After image acquisition, images are analysed by expert radiologists who identify suspect tumorous tissues, which can be either localised or involve multiple sites. It is crucial to define and delineate the ROI of a lesion to measure tumour size and extent accurately.

Monitoring changes in size can be instrumental in determining prognosis and assessing the response to therapy. Additionally, the delineation of the ROI forms the foundation of the radiomic process, enabling the development of customised ML models for various purposes. Segmentation can be used for classification to distinguish between malignant and benign lesions, differentiate various molecular alterations, assess and predict disease progression, and estimate potential treatment outcomes. In this way, it aids in selecting the most appropriate therapy and personalising medical procedures to enhance patient care. Depending on the complexity of the task, different segmentation methods are employed. Traditionally, manual segmentation is regarded as the gold standard, but it is highly time-consuming and limits the creation of large datasets, which are essential for training complex models. To address the challenges of time efficiency and reproducibility, various semi-automatic and fully automatic procedures have been developed and implemented over time. For simpler tasks, such as segmenting whole organs, traditional techniques that do not require ML are often used. These methods include image thresholding, atlas-based segmentation, active contour models, and region growing techniques.

When the object of interest is a small lesion or a portion of tissue with an irregular shape, multiple fragments, and intensity variations, more complex methods involving ML have proven to be more effective, as in the case of breast lesions. ML, closely related to data science and statistics, focuses on extracting valuable information from data and making predictions. It can be considered a subset of AI that develops algorithms capable of learning automatically from data. Traditionally, ML is divided into two frameworks: supervised learning, where both the input and the desired output are used to train the model, and unsupervised learning, where algorithms learn patterns from data without any label information. In the context of segmentation, the main advancements have been made with unsupervised clustering techniques and more sophisticated supervised and unsupervised deep learning algorithms.

Clustering is one of the most used unsupervised ML methods capable of uncovering structures in unlabelled datasets. Pixels are associated with different clusters based on customised similarity or distance measures. Clustering comes in different variations. A simple but effective approach is K-means, in which data is partitioned into K clusters using K initial centroids. The algorithm works as a minimisation problem, optimising a given cost function, for instance, based on the sum of the squared differences between cluster means and pixel intensity values. Alternative distance definitions, other than the Euclidean distance can also be used. The choice of K and centroids is made at the beginning, using, for instance, random values, and points are associated with the closer centroid. The procedure then iteratively refines the clusters using expectation maximisation until the chosen convergence criterion is met [22]. Despite being efficient and always ensuring convergence to a local minimum [22], k-means also has some drawbacks because the results strongly depend on the initialisation of centroids. Moreover, computational complexity is important to consider as it depends on the number of data elements, clusters and iterations [23]. K-means is regarded as a hard clustering technique, in which each point belongs to one and only cluster. However to handle more complex tasks as medical image segmentation, different approaches have been introduced, such as Fuzzy C-means (FCM), allowing data points to partially belong to different clusters, which can improve segmentation performances [24]. These kinds of models can be trained easily without any knowledge of ground truth. However, this also means that the accuracy may be lower with respect to supervised approaches, as no feedback from the gold standard is used.

Remarkable performance improvements have been achieved using deep learning models hold higher accuracy in different fields including medical imaging [25]. Deep neural networks (DNN) rely on the basic neuron unit, which performs a weighted sum

of the input features and after a non-linear transformation (activation) produces a scalar output. DNN usually consist of multiple layers made of several neurons; the output of one layer is used as input for the subsequent one and so on for all the layers until the final layer producing the ultimate output is reached. The training procedure for a DNN involves the optimisation of all the parameters exploiting the gradient descent technique [22]. The number of parameters increases with the depth of the net, requiring a larger amount of training data to learn all the parameters. In particular, the architecture of convolutional neural networks (CNN) emerged for segmentation tasks due to its smaller number of parameters that come from sharing weights among neurons of the same layer [25]. CNN rely on different layer types: convolutional layers, in which a kernel of weights is convolved with the image producing a feature map; non-linear layers that apply the activation function to the produced feature map and lastly pooling layers that allow progressive downsampling of the initial image increasing the receptive field. These layers can be stacked and used to obtain features from progressively larger receptive fields, enabling the production of accurate segmentations [22]. Most of the commonly used models rely on fully convolutional networks using only convolutional layers and encoder-decoder networks. The latter consists of the encoder part exploiting convolutional and pooling layers to capture context and the decoder part for deconvolution and localization, finally producing pixelwise prediction on the image providing the segmentation mask [25]. In the last years with the improvements in computational capabilities especially of GPUs training complex models has become more and more feasible, marking the flourishing of DNN [22]. An overall perspective on DL approaches for general medical image segmentation can be found in [26].

To compare different models it is essential to define proper metrics, depending on the segmentation task different choices can be made, considering volume based, overlap and distance based metrics. Considering the segmentation of medical volumes, model performances are usually estimated using the Dice-Sørensen coefficient (DSC) to estimate the overlap between automatic segmentation and ground truth, but also to measure reproducibility [27]. Given two sets, S_a and S_t being respectively the automatic and the ground truth segmentation, the dice score is defined by twice the cardinality of the intersection divided by the sum of the cardinalities of each set $DSC = \frac{2|S_a \cap S_t|}{|S_a| + |S_t|}$ [27].

The scope of this review is to explore the main results in terms of breast lesion segmentation on MR images with a particular accent on the evolution of DL, exploring results obtained with the most promising models and investigating their limits to guide new improvements of the models. The goal is to highlight the current state of the art in segmentation techniques, providing insights that could aid in the development of more effective models for automatic lesion delineation. The review addresses the advantages, limitations, and potential pitfalls of existing approaches. Since the first promising results of radiomics studies, increasing efforts have been made in terms of standardisation, such as the Imaging Biomarker Standardisation Initiative (IBSI) [28], which mainly focuses on how features must be computed. However, there is still a significant gap in the standardisation of segmentation methods, which highly impacts the reproducibility and reliability of studies. As evidenced by [29], segmentation methods influence radiomic features, particularly affecting shape features. Thus, further investigation into existing segmentation methods, along with their potential and pitfalls, might be crucial to achieving standardisation.

2. Materials and Methods

A systematic literature review was conducted to explore the state-of-the-art in breast lesion segmentation. This research focused primarily on studies from the past decade to highlight the recent proliferation of DL techniques in this field. Before limiting the search time window to the last decade a rapid search was made for the same query including

a larger time window. In Figure 1 the number of papers included every year shows an increasing trend, with a first peak in 2013 and other peaks approaching the twenties. For this reason and to focus on recent advancements only the window around 2013–2024 was considered.

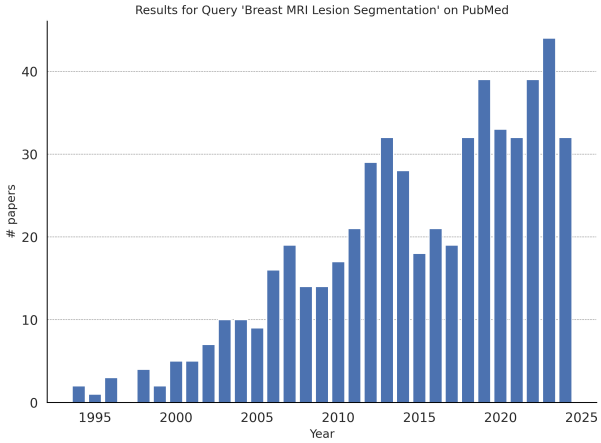


Figure 1. Trend in the number of indexed studies on PubMed related to "breast MRI lesion segmentation" over the years. The graph illustrates the growing interest and research activity in this field.

Studies published between January 2013 and July 2024 were searched in the PubMed database (US National Library of Medicine, http://ncbi.nlm.nih.gov/pubmed, accessed on 31 July 2024). The research query included the terms "breast MRI lesion segmentation" to identify studies involving any type of breast lesion, benign or malignant, detectable by any MR imaging modality. Each record (title and abstract) retrieved was manually screened by a reviewer, and ineligible records were removed. The selection procedure is illustrated in the flowchart in Figure 2, following the PRISMA 2020 guidelines [30]. The initial search yielded 359 records for the specified keywords and time interval. Before screening, records were checked for duplicates (0 cases). Papers not in English were removed to avoid language barriers and potential misunderstandings from poor translation (6 cases), and papers without accessible full texts were also excluded (21 cases). This left 332 records to be screened based on titles and abstracts. Records deemed off-topic, such as those involving clinical trials, case reports, general reviews, or studies related to different organs and imaging modalities, were discarded (157 cases). Additionally, articles that only included manual segmentation were excluded (78 cases), as well as studies focused on segmenting the whole breast tissue rather than the lesion itself (4 cases). This process resulted in 93 papers assessed for eligibility. After a full-text review, records that lacked specificity regarding segmentation, either using it only as an intermediate step without reporting segmentation performance or failing to provide proper validation, were excluded (61 cases). Moreover studies that did not use any machine learning techniques were excluded (3 cases). Studies employing automatic segmentation techniques with very small datasets ($\leq$20 patients) were also removed, as small sample sizes can lead to unreliable and non-generalizable results (7 cases). Ultimately, 21 articles were included in the review. All the included studies were approved by the respective local institutional review boards. Written informed consent was obtained for most cases, ensuring that no personal information could be identified. In other cases, informed consent was not required as the data were either fully anonymised or publicly available. The articles were analysed by examining the aim of the segmentation task, the type of model used, the MRI sequences employed with consideration for multiparametric models, and the sample size and data quality, with attention to the generalizability and flexibility of the proposed pipelines. The results and discussion are presented in the subsequent sections.

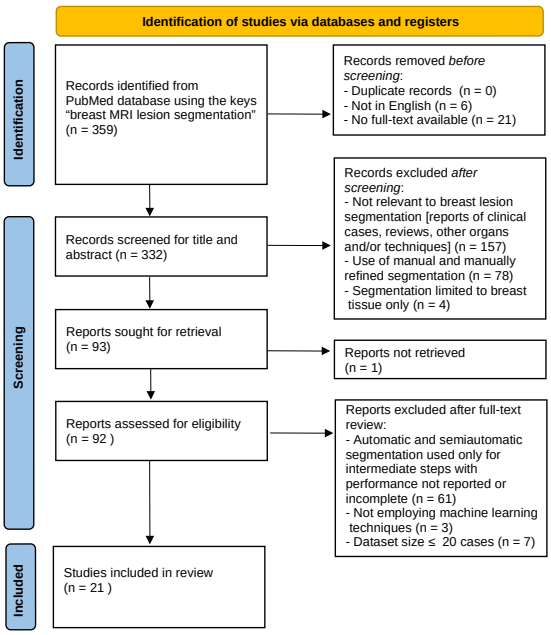


Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram showing the different phases of the review.

3. Results

All 21 studies that met the eligibility criteria were analysed. The main findings are summarised in Table 1, including the magnetic resonance (MR) technique employed, the magnetic field strength of acquisition, the number of patients involved, the primary research questions addressed, the algorithm used, its validation process, and the main performance metrics of the proposed models. The studies can be classified into three distinct clusters based on the type of ML algorithms employed, showing a correlation with advancements in specialised hardware. In the earlier years of the investigated period, studies mainly focused on unsupervised clustering approaches (3 cases), followed by an exploration of traditional supervised ML algorithms (2 cases). In contrast, in the second half of the examined decade, there was an increasing interest in deep learning techniques (16 cases), especially those based on CNN, reflecting improvements in GPU technology, which have gradually made these computationally intensive methods feasible.

3.1. Early Unsupervised Models

Early-stage ML approaches to automatic segmentation mainly rely on the application and refinement of basic clustering techniques, such as FCM, K-means, and mean shift clustering, investigated respectively by Jayender et al. [31], McClymont et al. [32], and Hu et al. [33]. In general, information from the whole DCE-MRI and its time-intensity curve (TIC) is used to generate the segmentation mask of the lesion; however, complementary information from anatomical T1w and T2w images has also been used, for instance, by McClymont et al. [32]. Different preprocessing and postprocessing procedures have been employed, reflecting the use of clustering both as a final step or as a baseline requiring further operations.

Table 1. Review of breast lesion segmentation studies. Model performances are presented using mean Dice scores (DSC) or equivalent metrics.

Paper	Magnetic Field	MR Technique	Patient Number	Question	Algorithm	Supervised Approach	Validation	Performance
McClymont et al. 2014 [32]	1.5 T	DCE, T1w ,T2w	128	detection and segmentation of suspicious tissue in breast	mean-shift clustering and graph-cuts on a region adjacency graph	No	train: 35 cases test 1: 93 cases test 2: 9 cases	median DSC test set 1: 0.76 (IQR 0.17) test set 2: 0.75 (IQR 0.16)
Jayender et al. 2014 [31]	3.0 T	DCE	24	segmentation of invasive ductal carcinomas	time series analysis based on linear dynamic system and FCM	No	test: 24 cases	mean DSC vs. radiologist: 0.77 vs. CADstream output: 0.72
Hu et al. 2015 [33]	1.5 T, 3.0 T	DCE	36	segmentation of breast lesions	image manifold revealing and k-means clustering	No	test: 36 cases	mean accuracy: 0.97 ± 0.01 mean specificity: 0.98 ± 0.01 mean sensitivity: 0.90 ± 0.04
Vogl et al. 2019 [34]	3.0 T	DCE, DWI + other mod. ^{18}F-FDG-PET.	34	segmentation of breast lesions and characterization (ben. vs. mal.)	voxel wise random forest classifier based on local textural, intensity-based, and kinetic features	Yes	train: 1000 voxels per class and per case valid.: LOOCV on a new case.	DSC mean: 0.665 ± 0.236 median: 0.757 (IQR 0.189)
Min et al. 2020 [35]	1.5 T	DCE, T1w ,T2w	117	detection, segmentation and characterization of breast lesions (ben. vs. mal.)	lesion candidates generated by 3D multiscale morphological sifting, RUSboost to classify into lesion or normal tissues	Yes	train: 58 cases test: 59 cases	DSC mean: 0.72 ± 0.15 median: 0.74
Spuhler et al. 2019 [36]	1.5 T	DCE	317	segmentation of breast lesions, prediction of sentinel lymph node status	U-Net	Yes	train: 263 cases test: 54 cases	mean DSC vs. Radiologist 1: 0.71 ± 0.16 vs. Radiologist 2: 0.61 ± 0.17 vs. Radiologist 3: 0.67 ± 0.18
Galli et al. [37]	1.5 T	DCE	33	segmentation of breast lesions	three-time-points framework, modified 2D U-Net	Yes	10-fold CV train: 4320 slices valid.: 480 slices test: 480 slices	mean DSC: 0.70
Vidal et al. 2022 [38]	1.5 T	DCE	46	breast lesion segmentation	ensemble of 3 3D patch-based modified U-Net	Yes	5-fold CV	mean DSC all lesions: 0.68 main lesion only: 0.80
Rahimpour et al. 2023 [39]	-	DCE, T1C	141	segmentation of breast lesions	visual ensemble of CNN	Yes	train: 111 cases test: 30 cases	mean DSC: 0.79 ± 0.16

Table 1. *Cont.*

Paper	Magnetic Field	MR Technique	Patient Number	Question	Algorithm	Supervised Approach	Validation	Performance
Qin et al. 2022 [40]	1.5 T	DCE, T2w	160	segmentation of breast lesions	recurrent attention U-Net	Yes	train: 128 cases test: 32 cases	mean DSC: 0.78
Zhou et al. 2023 [41]	3.0 T	DCE	1760	segmentation of breast lesions and axillary lymph node metastasis discrimination	Improved U-Net incorporating attention module and edge feature extraction modules	Yes	train: 1110 cases internal test: 476 cases, external test: 174 cases	mean DSC internal test: 0.83 external test: 0.82
Zhu et al. 2022 [42]	3.0 T	DCE, DWI	3303	segmentation and characterization of breast lesions (ben. vs. mal.)	2D V-Net (for DCE) attention U-Net (for DWI)	Yes	train: 2339 cases valid.: 724 cases internal test: 469 cases, external test: 123 cases	mean DSC V-Net: 0.860 att. U-Net: 0.829
Liu et al. 2023 [43]	1.5 T	DCE	82	segmentation and characterization of breast lesions (molecular subtype alteration)	2D-Dense U-Net based on the Dense Net-121	Yes	train: 66 cases test: 16 cases	mean DSC: 0.72 (on all dataset)
Zhang et al. 2023 [44]	3.0 T	DCE	222	segmentation and characterization of breast lesions (ben. vs. mal.)	nnU-Net	Yes	train: 196 cases test: 50 cases	mean DSC: 0.831
Liu et al. 2023 [45]	1.5 T for training, 3.0 T for testing	DCE	361	segmentation and pCR prediction	two-step lesion segmentation method based on nnUNet	Yes	train: 246 test: 115	mean DSC: 0.83
Douglas et al. [46]	1.5 T, 3.0 T	DCE	689	segmentation of breast lesions	2D U-Net, 3D U-Net	Yes	5-fold CV	median DSC 2D U-Net: 0.780 (0.774, 0.787) 3D U-Net: 0.721 (0.710, 0.732)
Zhang et al. 2022 [47]	3.0 T for training, 1.5 T for testing	DCE	339	segmentation of abnormal breast lesions	mask region-based CNN, with ResNet-101 as backbone	Yes	10-fold CV train: 241 cases test: 98 cases	mean DSC: 0.79
Peng et al. [48]	1.5 T	T1C	590	segmentation of breast lesions	lesion morphology aware network	Yes	train: 413 cases valid.: 59 cases test: 118 cases	mean DSC: 0.87 $\pm$ 0.14

Table 1. *Cont.*

Paper	Magnetic Field	MR Technique	Patient Number	Question	Algorithm	Supervised Approach	Validation	Performance
Sun et al. 2023 [49]	1.5 T, 3.0 T	-	248	segmentation and characterization of BI-RADS category 4 breast lesions (ben. vs. mal.)	three-stream decoder (boundary, location and segmentation) based on ResNet-34 and U-Net	Yes	train: 70% test: 30%	mean DSC: 0.70
Ru et al. 2023 [50]	3.0 T	DCE	150 private dataset, 10 public dataset	segmentation of breast lesions	attention U-Node 3D	Yes	3-fold CV	mean DSC private dataset: 0.7027 public dataset: 0.6970
Chen et al. 2024 [51]	-	DCE, ADC, T2w	600, 439	multi-modality tumour segmentation, pCR prediction	Modality-Specific Information Disentanglement framework, implementing Modality Specific Attention module and the Modality Trusty Gating module, Deep Supervision module	Yes	4-fold CV train 1: 360 cases valid. 1: 120 cases test 1: 120 cases train 2: 263 cases valid. 2: 88 cases test 2: 88 cases	mean DSC dataset 1: 0.8903 $\pm$ 0.0195 dataset 2: 0.8445 $\pm$ 0.0127

CNN = Convolutional Neural Networks; CV = Cross-Validation; DCE = Dynamic Contrast Enhanced; DSC = Dice-Sørensen Coefficient, DWI = Diffusion-Weighted Imaging; FCM = Fuzzy C-Means, LOOCV = Leave-One-Out Cross-Validation; pCR = Pathological Complete Response; RUS = Random Under-Sampling; T1C = T1-Weighted with Contrast; T1w = T1-Weighted; T2w = T2-Weighted; TIC = Time-Intensity Curve.

In particular, the most significant work in terms of sample size was performed by McClymont et al. [32]. Their work aimed to segment both benignant and malignant breast lesions, utilising multimodal clinical breast data (DCE-MRI, T1w, T2w) and implementing mean-shift clustering followed by the creation of a region adjacency graph on which graph cuts are applied. The dataset consisted of a total of 128 subjects, that were divided into a training set (35 cases) and two test sets with data from different scanners (respectively with 93 and 9 lesions), reserving the training set for tuning model parameters. The input image consisted of vector-valued voxels obtained by stacking values from each modality. The need to minimise false positives in the chest cavity and reduce computational times prompted the inclusion of a preliminary step for segmenting the breast tissue. Mean-shift clustering is then applied on the vector-valued voxels within the restricted breast ROI, grouping similar voxels in the multidimensional feature space. Once the clusters have been obtained, a region adjacency graph over them is calculated, blood vessels are identified, and only the suspicious tissues are segmented by applying graph cuts employing a dissimilarity measure evaluating the difference of pre- and post-contrast signals. Finally, lesion candidates with too small a size are removed, and the remaining lesions are ranked based on the relative enhancement. Performances were assessed against the manual segmentation of an experienced radiographer and obtained a median DSC of 0.76 (IQR 0.17) and 0.75 (IQR 0.16), respectively, on the first and second test sets. Overall, their pipeline achieved a fully automatic segmentation, exploiting the potential of multimodal data and reaching comparably robust results on data from different scanners and vendors.

Another approach to segment breast lesions, in particular for invasive ductal carcinomas, was proposed by Jayender et al. [31]. They proposed to exploit the full DCE-MRI information by using time-series analysis and FCM clustering. Relevant features can be extracted from the TIC, creating an observation vector that was modelled with a linear dynamic system. The clustering was set with two classes, allowing discrimination between tumour and healthy voxels on a voxel-by-voxel basis, producing a final map with tumour delineation. The procedure was tested on a dataset with 24 patients, comparing automatic segmentation with the manual segmentation of both a radiologist and preexisting commercial software, obtaining, respectively, a mean dice score of 0.77 and 0.72.

In addition, Hu et al. [33] investigated another clustering method exploiting the TIC to segment generic (both malignant and benign) breast lesions from DCE-MRI. Their line of action involved the use of all the available time-points T to create high-dimensional images of size $n \times T$, where n is the number of pixels. To reduce computational costs while preserving information, the image dimensionality was subsequently reduced to a 3-dimensional space by using the non-linear method of Laplacian Eigenmaps, preserving TIC information and manifold structure. Finally, k-means clustering was applied to the 3D enhanced dataset, where the lesion becomes more evident, resulting in the final segmentation. The whole procedure was tested on 36 clinical cases, comparing automatic segmentation with the radiologist's gold standard segmentation, achieving overall high specificity (0.98 ± 0.01) and sensitivity (0.90 ± 0.04).

3.2. Traditional Supervised Models

Different ways of addressing the segmentation task emerged with the growing interest in supervised learning for classification. Intensity and dynamic features can be associated with a single voxel or a group of voxels of a candidate lesion, allowing the training of binary classifiers able to identify the lesion, as respectively explored by Vogl et al. [34] and Min et al. [35].

In particular, Vogl et al. [34] adopted a multimodal approach exploiting imaging data, including DCE-MRI, DWI, and also ^{18}F-fluorodeoxyglucose-PET, to detect, segment,

and classify both malignant and benign breast lesions. Their procedure involved the extraction of local, intensity-based, and kinetic features, and the training of a Random Forest (RF) classifier able to discriminate lesion and normal tissue voxel-wise. Further preprocessing stages involved the segmentation of the whole breast area based on a region growing algorithm. Intensity-related features were extracted from each voxel and each modality and used to train a RF classifier. For each patient, 1000 randomly selected samples per class were used in the training phase. Feature selection was performed using two methods, Gini importance and minimum redundancy maximum relevance, with the latter obtaining better results and less correlated features. Performances were assessed using leave-one-out cross-validation (LOOCV) on a dataset of 34 subjects, comparing the automatic segmentation with the manual one, obtaining a mean dice of about 0.67 using all the modalities, a dice of 0.618 using DWI and DCE, and a dice of 0.58 using only the DCE, suggesting enhanced performances for multimodal approaches. The multiparametric approach actually improved the segmentation, reducing false-positive cases, identified as blood vessels and parenchymal enhancement.

Another notable approach by Min et al. [35] addressed the issue of segmentation of benign and malignant lesions from breast T1w, T2w, and DCE-MRI data. They first generate candidate lesions through 3D morphological sifting and then classify them as normal tissue or lesion using the random under-sampling (RUSBoost) classifier. The preprocessing stage involves the segmentation of the whole breast region on the T1w image using Otsu thresholding and Hayton's algorithm and splitting into left and right breasts, allowing the processing of one breast at a time. Next, from the subtraction image (DCE pre-contrast and first DCE post-contrast), lesion candidates are generated using morphological filters with linear structuring elements at three different scales (1, 1/2, and 1/4). Scaled images also undergo multilevel Otsu thresholding and size thresholding. Finally, features of four categories—morphological, intensity, texture, and enhancement kinetics—are extracted from all the available images (T1w, T2w, and DCE-MRI). The extracted features are used to determine if the candidate is a real lesion or not. The model was trained on 58 subjects, leaving 59 cases for testing. Automatic segmentation is compared to radiologists' manual segmentation, achieving an average dice score of 0.72 ± 0.15 with a median of 0.74.

3.3. Deep Learning Models

Among more recent studies, there has been increasing interest in deep learning, particularly in exploring CNN based on the U-Net architecture first proposed by Ronneberger et al. [52] for general biomedical applications (11 cases). Apart from the U-Net encoder-decoder structure, other multi-stage neural network architectures have also been presented (5 cases).

3.3.1. U-Net and Derived Models

U-Net's typical symmetrical encoder-decoder architecture dominates the landscape of deep learning segmentation for medical imaging. Initial studies, such as the one conducted by Spuhler et al. [36], utilised the original 2D model, including partial information from the DCE-series, using only two input channels. However, the intrinsic challenge of segmenting breast lesions, especially malignant ones, calls for increasing customisation of the network architecture, including generalisation to 3D datasets to capture small lesions and the introduction of customised attention modules. Moreover, several preprocessing steps (generation of a breast ROI, cropping, normalisation, etc.) were commonly used to reduce misleading contributions, together with data augmentation techniques to prevent overfitting and the use of modified loss functions (e.g., introduction of a Dice contribution to binary cross-entropy) to improve the model's performance. Attempts to use more infor-

mative time points exploiting the full TIC information have been made by Galli et al. [37], who used three time points, and by Vidal et al. [38] and Rahimpour et al. [39], who employed ensembles of 3D U-Net models, each with diverse combinations of time points as input. Qin et al. [40] introduced a recurrent attention module, while Zhou et al. [41] incorporated an edge feature extraction module to better capture edge details and small tumours. Attention modules were also employed by Zhu et al. [42]. Further modifications to the architecture, such as DenseUNet and nnU-Net, were effectively explored by Liu et al. [43], Zhang et al. [44], and again by Liu et al. [45]. Finally, comparisons between general 2D and 3D U-Net approaches and the use of specific time points were explored by Douglas et al. [46], reflecting the difficult evaluation and selection of the ideal model.

One of the first applications of the U-Net architecture for segmenting malignant and benign breast lesions was conducted by Spuhler et al. [36]. They used a U-Net architecture with two separate input channels for the pre-contrast and the first post-contrast images of a DCE-MRI series acquired in the sagittal plane. This choice reflected the approach typically used by radiologists when performing manual segmentation. The network was fed with individual 2D slices containing lesions and was trained using a binary cross-entropy (BCE) loss function. To improve model performance, data augmentation techniques, including small random rotations, were also applied. The model was trained on 263 subjects and tested on 54 subjects, using reference segmentations from three different radiologists. When the first radiologist's segmentations were used to train the model, a mean Dice score of 0.71 ± 0.16 was obtained with respect to their own segmentations, while Dice scores of 0.61 ± 0.17 and 0.67 ± 0.18 were achieved when compared with the ground truth from two other radiologists. The procedure was repeated using segmentations from a second radiologist for training, yielding slightly lower Dice scores of 0.67 ± 0.20, 0.64 ± 0.22, and 0.61 ± 0.18 for comparisons with different operators. These results were comparable to the mean inter-operator variability, which ranged between 0.61 and 0.68. The effectiveness of the obtained segmentations was also assessed in terms of predictive capability. The segmentations produced by the model were used as a baseline for feature extraction to predict lymph node metastases, achieving area under the curve (AUC) values comparable to those obtained using manually generated segmentations. This suggests that the U-Net captures relevant information and exhibits predictive power comparable to radiologists for this specific task.

An improvement of the two-input channel U-Net comes from the work of Galli et al. [37], who proposed using a three-time-point framework to segment breast lesions from DCE-MRI. The method is based on the use of three specific time points to capture the contrast agent's kinetic information, specifically the pre-contrast image, the post-contrast image after 2 min, and the post-contrast image after 6 min. Their model works on 2D slices and requires several preprocessing steps: the generation of a breast mask using a multiplanar 2D U-Net to remove misleading tissues and the application of motion correction. From the masked image, a three-channel image is created by stacking slices from the three different time points. Additionally, the left and right breasts are separated for the training procedure. A further difference from the original U-Net is the use of a Dice loss function. The model was trained and tested using 2D slices from a set of 33 subjects. Performance was estimated through 10-fold cross-validation (CV), obtaining a median Dice score of about 0.70, considering both lesion and healthy slices. Moreover, the three-time-point framework was compared with the basic U-Net and U-Net++, achieving higher performance compared to the original models, which had median Dice scores of about 0.67 and 0.65, respectively.

Another approach for the segmentation of malignant lesions using multiple time points of the DCE-MRI series was presented by Vidal et al. [38]. Starting from the idea of

using the entire sequence information, they proposed an ensemble of three U-Nets, each using different time points as input channels. They modified the basic U-Net architecture to handle 3D data by replacing standard convolutional blocks with residual blocks while maintaining the usual BCE loss. Several preprocessing steps were necessary to train the network, including the generation of a breast ROI using landmark detection to remove other enhanced organs, normalisation, and balanced extraction of smaller 3D patches from the ROI. The performance of the ensemble was assessed using 5-fold CV on a dataset with 46 cases, comparing automatic segmentation with radiologists' ground truth. The final segmentations were obtained through both majority voting and union of the three U-Nets, achieving mean Dice scores of 0.679 ± 0.258 and 0.674 ± 0.224, respectively. However, some false positives corresponding to axillary lymphadenopathy and vessels were reported. When considering only the main lesion, mean Dice scores of 0.728 ± 0.243 and 0.790 ± 0.172 were achieved, respectively. They also observed that maintaining the original image resolution yielded better results compared to setting a forced isotropic voxel size. The U-Net ensemble was further compared to FCM segmentations, consistently showing better results on the same dataset. Thus, the study suggests the possibility of using fully automatic methods that do not require user selection of the ROI.

An alternative multiple time point ensemble solution was proposed in the study by Rahimpour et al. [39], which focuses on the use of 3D U-Net to segment malignant breast tumours in DCE-MRI. In this case, three separate 3D U-Nets were trained using the following inputs: only the first post-contrast image, the first post-contrast image and the first subtracted image with a fusion strategy at the image level, and the first post-contrast image and the first subtracted image with a fusion strategy at the feature level. A weighted combined loss was used, consisting of a BCE and a Dice term. Preprocessing steps were required, including correction for bias field gain, resampling, cropping to a fixed-size bounding box to cover the whole breast area, and intensity normalisation. The model was developed on a dataset of 141 cases, with 111 cases used for training and 30 for testing. The data were collected from various institutes using diverse scanners. Ground truth was provided by two experienced radiologists, with an inter-operator variability in terms of Dice score of 0.78 ± 0.10. The performances of the individual components of the ensemble reached mean Dice scores of 0.73 ± 0.23, 0.75 ± 0.20, and 0.70 ± 0.26, respectively. The visual choice from the ensemble allowed for a higher score of 0.79 ± 0.16, while strategies such as majority voting or averaging yielded lower performances (approximately 0.75 ± 0.22), indicating that human intervention is still required to achieve optimal segmentation.

Another frontier involves the enhancement of the U-Net architecture proposed by Qin et al. [40], combining a recurrent attention module with a dense residual module based on dilated convolution, to segment malignant breast lesions in T2-weighted DCE-MRI. In this case, a two-step approach is used, with a preprocessing stage where a 2D U-Net is trained to segment the whole breast area, followed by a second stage in which lesion segmentation is performed only within the ROI. To effectively handle imbalanced datasets, a Tversky loss function was used. Moreover, to improve model accuracy, exhaustive data augmentation was applied, including scaling, mirroring, and elastic deformations. The network processes individual 2D slices, and a final 3D reconstruction is obtained by merging all the 2D output slices. The model was trained on 128 subjects and tested on 32 cases, obtaining a mean Dice score of 0.78. The enhanced model outperformed other pre-existing models, including U-Net, attention U-Net, residual U-Net, dense U-Net, and V-Net, which all scored no higher than 0.76 on the same dataset.

In addition, Zhou et al. [41] also investigated breast lesion segmentation on DCE data, using an improved version of U-Net that incorporates not only an attention module but also an edge feature extraction module to accurately capture edge details and small

lesions. In this case, preprocessing consisted of cropping the lesion layer, normalisation, and applying data augmentation techniques such as rotation, scaling, flipping, and shifting. The model uses a single input channel, taking the most significant phase of the DCE-MRI to identify the lesion as suggested by radiologists. The dataset consisted of 1110 training cases, 476 internal testing cases, and another group of 174 cases from other centres for external testing. A comparison between the model's results and manual segmentation achieved a mean Dice score of about 0.83 and 0.82, respectively, in the internal and external test sets. The improved U-Net also outperformed other state-of-the-art methods, such as SegNet, U-Net++, U-Net, Att-U-Net, CS-Net, and Multistage Hierarchical Learning, which achieved Dice scores in the range of 0.69 to 0.77. Moreover, the segmentations were used to extract features for discriminating axillary lymph node status, obtaining promising results that reinforced the reliability of the model.

Other types of enhanced U-Net have been used by Zhu et al. [42], who aimed to segment BIRADS 4 breast lesions from DCE-MRI and DWI. In particular, to segment the DCE-MRI, a 2D V-Net was used, employing data from three phases: the pre-contrast, the initial uptake, and the phase after an 8–10 min delay. To segment the DWI, an Attention U-Net was preferred. In both cases, the same combined loss was used, consisting of a sum of BCE and Dice loss. Both models were trained, tested, and validated on a dataset of 2823 subjects. Performance was assessed by comparing automatic segmentation with manual radiologist ground truth, obtaining a mean Dice score of 0.86 for the V-Net on the DCE-MRI and 0.829 for the Attention U-Net on the DWI. Moreover, both automatic segmentation tools were used to provide a baseline for lesion classification, reaching high AUC values, suggesting that the models produced reliable segmentations.

Despite focusing more on the classification of molecular subtypes rather than on segmentation, the work of Liu et al. [43] also provides some results in terms of segmentation performance using derivations of U-Net. In particular, they aim to segment malignant breast lesions from DCE-MRI using a 2D-DenseU-Net based on the DenseNet-121 architecture. The segmentation operates on 2D slices. To improve training and alleviate overfitting, data augmentation including scaling and random mirroring was performed. The dataset consists of 82 cases, of which 66 are used for training and 16 for testing, achieving a mean Dice score of approximately 0.720 compared to the radiologist's manual ground truth.

Similarly, Zhang et al. [44] used the nnU-Net model to segment both malignant and benign breast lesions. They used a single image as input to the network, specifically the first post-contrast image of the DCE series. Preprocessing involved two steps: splitting the dataset into right and left breasts to augment data, reduce overfitting, and increase the relative size of the lesion for easier detection; and a second stage of intensity normalisation. The model was trained and tested on 196 and 50 cases, respectively. Performance was evaluated with respect to manual ground truth provided by radiologists, achieving a mean Dice score of approximately 0.831, which is very close to the inter-reader variability assessed at 0.830, demonstrating overall good performance. Moreover, the segmentation served as a baseline for radiomic feature extraction and the development of a classifier with good predictive values, suggesting the efficacy of the segmentation.

Liu et al. [45] also used nnU-Net to perform malignant breast lesion segmentation from 3D DCE-MRI data. In particular, they proposed a two-stage approach, with the first stage focused on mammary gland segmentation and the second stage on segmenting the lesion within the mammary ROI. Both stages exploit the same nnU-Net architecture, capable of handling 3D data. The first stage is essential to reduce irrelevant information with similar enhancement, such as the heart, effectively reducing false positives. To include multicentric data, a histogram matching technique was employed, which served both to increase the model's transferability to other datasets and to provide data augmentation,

preventing overfitting. To capture both global and local information, the loss function combined cross-entropy and Dice terms. To address computational costs, a patch-based approach with overlapping patches was employed, where edge pixels were assigned different weights, resulting in a fused final segmentation. As a postprocessing step, only the largest connected contour was retained, effectively reducing the impact of false positives by discarding smaller fragments. The dataset consisted of 361 breast cancer samples from two hospitals equipped with different scanners, using 246 samples from the first institution for training and 115 samples from the second for testing. The model's performance was evaluated in two scenarios: without histogram matching, yielding a mean Dice score of approximately 0.78, and with histogram matching, which improved the score to 0.83. This demonstrated the model's strong generalisation capabilities. Moreover, the automatic segmentations served as a reference to extract radiomic features and build a predictive model for neoadjuvant response of axillary lymph nodes, showing promising results and further enhancing the model's reliability.

Douglas et al. [46] focused on comparing the performance of 2D and 3D U-Net models for general malignant mass and non-mass breast lesions on DCE-MRI, with respect to 3D FCM and 2D manual segmentations produced by radiologists. To obtain 3D segmentations from the 2D U-Net, individual 2D slices were stacked to create a 3D volume. The dataset consisted of 689 patients, with a total of 994 unique lesions of varying sizes and types. Model performance was evaluated using 5-fold CV. An initial comparison between FCM and U-Net resulted in a median Dice score of 0.78 (IQR: 0.774, 0.787) for the 2D U-Net and 0.721 (IQR: 0.710, 0.732) for the 3D U-Net. A second comparison utilised a subset of 71 2D segmentations provided by radiologists, corresponding to the central slice of each lesion. In this case, the 2D U-Net achieved a median Dice score of 0.864 (IQR: 0.845, 0.889), while the 3D U-Net reached a score of 0.802 (IQR: 0.766, 0.834). Overall, the 2D version of U-Net was shown to outperform its 3D variant with statistical significance. Further investigation revealed that using the second subtracted image of the DCE series yielded significantly higher performance compared to using the first subtracted image. Finally, other significant differences were discovered in terms of lesion type, as mass lesions generally achieved better segmentation compared to non-mass lesions. Overall, this work underlines the potential of using a 2D approach for effective 3D segmentation, offering a simpler and more efficient alternative to complex 3D techniques.

3.3.2. Alternative Approaches

In spite of the predominant use of U-Net, other deep learning approaches have also been explored. For instance, bounding box-based detection methods, which focus on identifying regions of interest before segmentation and primarily utilise ResNet as a backbone, were proposed by Zhang et al. [47] and Peng et al. [48]. A subsequent approach investigated by Sun et al. [49] utilised multiple decoder streams to identify lesion locations, refine boundaries, and produce complete segmentations. In addition, there are also multi-encoder architectures capable of handling multimodal data, such as the one developed by Chen et al. [51]. Moving away from traditional CNN, the use of neural ordinary differential equations has also been investigated, as demonstrated by Ru et al. [50], who proposed an attention-based U-Node.

In particular, the work of Zhang et al. [47] aimed to segment malignant mass lesions using DCE-MRI. The model was based on a custom derivation of a mask region-based CNN, employing the pretrained ResNet-101 as its backbone. The main features of the system involve the parallel processes of region proposal, lesion detection, and segmentation. The architecture has three input channels: one for the pre-contrast image, useful for identifying the chest area, and two for subtraction images, respectively representing the diseased breast

and the contralateral normal one. On each 2D slice, several predefined bounding boxes are placed to identify abnormalities, which are then ranked and classified as containing lesions or not. Finally, a segmentation network produces the lesion segmentation. The model was trained with 241 cases of fat-saturated images, while testing was performed on 98 subjects with non-fat-saturated images. Automatic segmentation was compared with FCM ground truth. Performance on the training set was estimated via 10-fold CV, obtaining a mean Dice score of 0.82, ranging from 0.64 to 0.97. On the test set, a mean Dice score of 0.79 was reported, with values ranging from 0.31 to 0.97. Despite good performance in some cases, small lesions were not detected, and strong parenchymal enhancement led to false positives. Nevertheless, the model can be regarded as robust, performing well on both fat-saturated and non-fat-saturated images.

Peng et al. [48] applied an alternative hierarchical approach to 2D breast tumour segmentation on T1w with contrast (T1C) MRI. In the first stage, the breast region is segmented using a simple U-Net, and within this breast area, the second stage of lesion segmentation is applied. They developed a lesion morphology-aware network based on a new hybrid loss function (revised binary sigmoid + Dice) that maximises inter-class distance, reducing both false positives and false negatives. The customised network utilises adaptive deformable convolutions that interact with spatial-wise and channel-wise attention modules, where the receptive field is adjusted based on lesion morphology through a learnable dilation parameter. The model is based on the structure of DeepLab V3+, with ResNet-101 serving as the backbone for the encoder. A total of 413 subjects were used for training, 59 for validation, and 118 for testing. The network was also able to detect tumours missed by radiologists and produce smoother boundaries. Performance assessed against manually delineated tumours achieved a mean Dice score of 0.87 ± 0.14. Other models, including U-Net, DeepLab V3+, and hierarchical CNN, were also compared, with lower Dice scores (mean Dice scores ranging from 0.83 to 0.85).

Sun et al. [49] proposed a three-stream subnet for BIRADS 4 lesion segmentation on breast MRI. The segmentation network includes three decoders that simultaneously work to locate the lesion, identify boundaries, and produce the final segmentation. Unlike previous approaches, ResNet-34 is used as a building block, with a location stream focusing on extracting features only from the lesion area, thereby preventing the loss of small lesions. Additionally, an upsampling layer aids in refining boundaries. Finally, the segmentation decoder, based on U-Net, incorporates multiscale receptive fields to adapt to lesions of different sizes. The dataset consists of 248 subjects, with 70% used for training and the remaining 30% reserved for testing. Automatic segmentation was compared to radiologists' ground truth, achieving a mean Dice score of 0.70 on the test data. Performance was also assessed for other models, which yielded lower results, such as CE-Net (Dice 0.67), Shape Attentive U-Net (Dice 0.66), U-Net+ with data augmentation (Dice 0.65), U-Net+ (Dice 0.63), AttU-Net (Dice 0.63), U-Net (Dice 0.61), and W-Net (Dice 0.60). The importance of using a three-stream approach was also investigated, showing slightly lower performance when stages were removed (mean Dice of 0.69 when only the segmentation stream was used). Moreover, the segmentations were used as the basis for a further classification subnet, enabling discrimination between benign and malignant lesions.

The challenge of malignant tumour segmentation was addressed by Chen et al. [51], who employed multiparametric imaging using DCE-MRI as the primary modality, complemented by T2w and ADC data. A first stage was reserved for the generation of synthetic T2w and ADC data from MRI using a residual U-Net with L_1 loss. Real and synthetic images were compared and used to generate modality-specific attention maps. Segmentations were produced using a multiencoder architecture with latent feature fusion, enhanced by modality-specific attention modules, modality trust gating modules, and deep supervision

modules with dice losses. Performances were assessed with 4-fold CV on two separate datasets, comprising 600 and 493 subjects, respectively. In each case, 60% of the data was used for training, 20% for validation, and 20% for testing. The model achieved a Dice score of 0.8903 ± 0.0195 on the first dataset and 0.8445 ± 0.0127 on the second. The proposed model significantly outperformed other multimodality methods in both cases, demonstrating overall efficacy. In addition, the automatic segmentation of breast lesions was used to train a model for predicting NAC response, achieving a predictive accuracy comparable to that of manually segmented tumours, underscoring the reliability of the proposed framework.

A different approach based on neural ordinary differential equations (ODE) was investigated by Ru et al. [50]. Their work relies on the fact that deep neural networks can represent a discretisation of a continuous dynamical system, thus deep learning problems can also be formulated in terms of differential equations that can be solved using traditional techniques. They combined the concepts of neural ODE with the U-Net encoder-decoder structure to segment malignant and benign breast lesions in both MRI and ultrasound (US) data. The use of ODE blocks in place of standard convolutional blocks allows for a reduction in the number of parameters, thereby mitigating the risk of overfitting. To improve performance, the Convolutional Block Attention Module (CBAM) was implemented, incorporating channel and spatial attention submodules. Additionally, a combined loss function, including cross-entropy and Dice terms, was employed. The model demonstrated generalisation for both 2D and 3D data. Performance on 2D DCE-MRI was assessed using a private dataset of 150 patients with 3-fold CV, yielding a mean Dice score of 0.70. Similar results were obtained on a 3D dataset of 10 patients. The Attention-U-Node was compared with other 2D and 3D models, consistently achieving superior performance on the given datasets. Furthermore, the same network was successfully trained and tested on multiple US datasets, demonstrating good adaptability to different imaging modalities.

A direct comparison of results from different models and approaches is not straightforward, as no common public dataset was used to evaluate performance on a standard basis without introducing bias. Most of the studies are based on private, local datasets related to the local population, each having its own limitations due to small dataset size and variability, the use of different 2D vs. 3D approaches, the inclusion of lesion-only slices or complete slices, and differing validation procedures. Despite these many differences, an attempt at raw quantitative comparison across different model categories has been made to highlight the main trends. Only studies reporting mean or median Dice scores (assuming a negligible deviation from a Gaussian distribution) have been considered. For the main approach categories, a comprehensive Dice score has been estimated by averaging scores using the sample size of the test set as a weight, prioritising studies with larger and more inclusive test sets. The main results are summarised in Table 2.

Table 2. Mean DSC and standard error across the main approach categories.

Early Unsupervised	Traditional Supervised	U-Net and Derived	Alternative DL
0.75 ± 0.01	0.70 ± 0.03	0.79 ± 0.03	0.83 ± 0.04

Early unsupervised methods achieved a mean Dice score of 0.75 ± 0.01, surpassing traditional supervised methods, which achieved 0.70 ± 0.03. However, the introduction of deep learning methods further improved segmentation performance, achieving a higher overall Dice score of 0.79 ± 0.03 for U-Net-based architectures, and an even higher score for alternative deep learning approaches, with a mean Dice score of 0.83 ± 0.04. These results suggest the potential for further advancements in model architectures and training strategies.

4. Discussion

Across the last years, diverse ML models were used to deal with the challenge of breast lesion segmentation on MRI data. The differences in sample size, inclusion criteria, and methodology make it difficult to compare results and performances across the diverse proposed methods.

A first essential criticality regards the sample size, that in the majority of cases is reduced. The paucity of publicly available datasets with labels restricts very often the analysis to private local samples from a single centre, using a single scanner and only a limited variety of lesions. Differences in scanners and acquisition parameters could be challenging, and it is difficult to forecast the effects on model performances without having external validation sets. In some fortunate cases where multicentric and multiscanner datasets were available, solutions such as data augmentation based on histogram matching to data from other centres proved to be highly effective, as pointed out by the study of Liu et al. [45]. Another possible issue is the fact that most of the studies with larger samples use subjects of Asian populations exclusively, thus there is the need to include more heterogeneous data also from other populations, as suggested by Chen et al. [51]. Moreover, in general, reduced sample sizes fail to represent different parenchymal compositions, often leading to false positive detections. Private datasets tend to be smaller in proportion to the local population around the centre; moreover, data labelling is often a burden for radiologists, who might handle more urgent tasks, dilating times to collect large enough samples. Moreover one has to consider that the use of ML in medical imaging plays a critical role in improving diagnostic and prognostic stages. However, the risk of bias must always be taken into account. The potential for mitigating cognitive bias in the interpretation of images might be compromised by unbalanced training datasets, suboptimal architecture design, and incorrect application of models [53]. The use of imbalanced training datasets has been shown to lead to poor performance when applied to heterogeneous populations, whereas models trained on balanced demographic groups achieve better performance. It is, therefore, essential to investigate further and work toward building commonly shared, balanced datasets for training, to create equitable models and reduce potential sources of bias. To achieve unbiased models, improvements in terms of publicly available heterogeneous datasets are also mandatory. Before designing a study, it could be useful to have common training or testing datasets based on a global population to reduce bias. For this purpose, some datasets are already publicly available, such as the BreastDM database from [54], or data provided by The Cancer Imaging Archive, which includes various imaging datasets like Breast MRI QIN-BREAST [55] and QIN Breast DCE-MRI [56]. In the former years, unsupervised clustering techniques and simpler supervised models were explored, thus the need for large datasets was not evident. However, the preference for highly performant deep learning approaches heavily demands larger and larger datasets in order to train effectively the more complex models without overfitting. The predominant approach is, in fact, the U-Net, that like all supervised methods requires labels. To reduce both labelling time and bias, and to address the lack of local annotations provided by experts, more attention should be directed toward the integration of weakly supervised models utilising minimally labelled slices in 3D lesions, such as deep multiple instance learning [57]. Alternatively, unsupervised networks, such as Generative Adversarial Networks (GANs) [58], or generalisations of the U-Net architecture designed to work in a completely unsupervised manner, like the W-Net [59], could be explored. Additionally, hybrid frameworks that combine the advantages of both supervised and unsupervised approaches present another promising avenue for research.

Another difference between the studies is the use of different types of lesions, thus segmentation performances on a specific type of lesion might not be generalis-

able to all the other types, with some cases that are highly challenging. In particular, McClymont et al. [32], who dealt with both malignant and benign cases, reported that the detection of ductal carcinoma in situ is particularly difficult, and this kind of lesion had a lower detection rate in their framework. Vogl et al. [34] reported that benign lesions with low contrast uptake were sometimes undetected. In general, few studies addressed non-mass lesions, as they are difficult to discriminate from the background, non-mass lesions having more irregular shapes and being possibly segmental, as reported by Min et al. [35]. In general, segmentation of mass lesions is more accurate with respect to non-mass lesions, as confirmed by the work of Douglas et al. [46]. Thus, investigation on this front, including datasets with a larger variety of lesions, is still necessary.

In terms of data type, the most common approach is to use the standard DCE-MRI series; however, there is no standard choice, as the selected time points vary across the analysed studies. With larger datasets and performant hardware, the use of the full sequence encoding the complete enhancement kinetic information is possible; however, this is not always the case. Early approaches, such as the one proposed by Hu et al. [33], tried to capture all available information while reducing data dimensionality. In most other cases, the choice instead involves using a pre-contrast and a post-contrast image or a single-phase subtracted image. Galli et al. [37] suggested, as a compromise, the use of three time points corresponding to the pre-contrast image, and the images at 2 min and 6 min after contrast injection. Douglas et al. [46] found that using the second post-contrast image yielded higher performances compared to the first post-contrast image on their dataset. However, the decision to exploit the full kinetic information might also depend on the lesion type to be segmented, as segmentation pipelines for generic lesion types might benefit from more information, whereas models focused on specific lesion types could reduce computational burden by using limited information from the sequence. In general, few studies have investigated the impact and improvement of using multimodal data. For instance, Vogl et al. [34] included DWI and PET data, and there are also few multiparametric approaches incorporating T2-weighted images and diffusion image information, such as the work of Chen et al. [51].

Another difference lies in the choice between 2D or 3D approaches, especially concerning the latter CNN possibilities. Douglas et al. [46] compared 2D and 3D U-Net models, showing higher performances for the 2D version; however, the results come mainly from comparisons with FCM-produced ground truth rather than 3D ground truth manually segmented by radiologists. Other studies, such as Rahimpour et al. [39], suggest upgrading the models to handle 3D data, since tumour volume is an important predictor for patient survival, and interslice information could also capture more detail in the case of small lesions. 3D approaches are often computationally expensive and require larger datasets; however, patch-based approaches could represent a possible solution to increase dataset size while maintaining a lower computational burden, as experienced by Vidal et al. [38]. Thus, the choice of 2D models often reflects the lack of large datasets or computational resources, preferring to use single 2D slices during the training phase and reconstruct the whole segmentation by fusion only at the end of the process. Further investigation should be conducted to understand whether the potential advantages of 3D models are actually realised with the current dataset sizes.

Another crucial point is the use of only lesion-positive slices or 3D patches in training and evaluation in many of the reviewed works. This choice is prevalent in studies that use segmentations as a baseline for radiomic analysis, where the focus is exclusively on lesion slices. However, it represents a clear limitation in performance evaluation for more general segmentation tasks. As remarked by Spuhler et al. [36], models that have never been exposed to lesion-free slices might not be immediately generalisable and may require

further tuning with normal tissue before being applied to different tasks and datasets. Galli et al. [37] further highlighted this issue, showing that scores obtained solely on lesion slices tend to overestimate the model's performance on broader segmentation tasks. To ensure comparability, results should be reported consistently, either using the convention of lesion-only slices/patches or by including normal tissue portions in the evaluation. However, not all studies explicitly state their approach, making direct interpretation of the results challenging. The inclusion of normal tissue patches and slices should be accounted for in model customisation by selecting appropriate loss functions that handle lesion-free samples effectively and by employing balancing techniques to manage class imbalance, as demonstrated in the training scheme proposed by Galli et al. [37].

Despite the remarked differences, general agreement in terms of preprocessing steps was found in most of the studies. Instead of manually selecting a bounding box around the lesion, to achieve a fully automatic procedure, it is useful to first perform breast segmentation and then perform lesion segmentation only within the breast ROI. This step is essential to prevent false positives due to enhancement in the heart region and other possible confounding organs with similar enhancement. Moreover, this step also serves to reduce the computational burden of considering the whole image. For this purpose, some studies also employed division into left and right breasts to double the training sample size in a natural manner or the use of even smaller boxes. However, this approach, while benefiting from one side, removes the possibility to use bilateral breast symmetry to evaluate possible false positives due to intrinsic parenchymal enhancement, as mentioned by Min [35] and Zhang [47]. A good compromise might be represented by more complex models involving multiple streams for lesion detection, boundary characterisation, and segmentation, as for instance the one proposed by Sun et al. [49]. Moreover, when using multiple images, registration, normalisation, and resampling are often necessary. For deep learning approaches, data augmentation also represents a key step to train the model with its full potential and to improve its generalisation capabilities. Data augmentation usually involves small random rotations, scaling, shifting, mirroring, elastic deformations, and variations in grey-level distribution. In terms of postprocessing, in some cases, small components were ignored by posing a threshold on lesion size or keeping only the larger fragments. However, traditional ML approaches to classify the proposed lesions as normal or suspicious could be used, as for instance made by Min et al. [35]. For DL approaches, the tuning of the binarisation threshold is also essential to correctly balance false positives and undetected lesions.

Considering the whole corpus of studies, large attention was placed on CNN based on U-Nets, which, through continuous improvements, are achieving higher and higher performances. To improve segmentation in such unbalanced images, different loss functions were introduced apart from the classical binary cross-entropy, such as dice loss, focal loss, Tversky loss, and combinations of multiple loss functions (mainly binary cross-entropy and dice), together with the introduction of L2 regularisation terms. To further improve performance, attention modules were also employed, including channel attention, spatial attention, edge feature extraction modules, and even modules to weight information from different modalities, as proposed by Chen et al. [51], such as Modality-Specific Attention and Modality Trusty Gating modules. However, a comprehensive quantitative comparison of all these alternatives is still missing, requiring further assessment to determine the most beneficial choices.

Overall, in view of all the variability, a direct quantitative comparison of model performances is difficult. Most studies evaluate segmentation performance using the Dice score as a metric, obtaining values between 0.7 and 0.8 in most cases, lying within a range that is highly comparable with commonly encountered inter-operator variabilities. Some of the

more recent models, trained with larger datasets, have also achieved impressive performances, with Dice scores exceeding 0.8. Despite the various proposed architectures and the high scores, some issues remain, such as missing very small lesions or false positive detections due to blood vessels, parenchymal enhancement, and axillary lymphadenopathies. These underline the potential for further improving the models, both by increasing the training sample sizes and variability, and by refining finer details of the architecture, as well as the preprocessing and postprocessing stages. Moreover, the use of the Dice score as the sole metric might be limiting. Rahimpour[39] suggests including other metrics, such as HD95, to provide more comprehensive evaluations that are not solely correlated with segmentation loss.

Considering instead the application of segmentation, while some studies focused exclusively on segmentation, there was also an important component of the examined studies that used the generated segmentation mainly for radiomic purposes. In particular, automatic segmentation was a useful baseline to extract radiomic features aiming at discrimination of malignant and benign lesions, characterization of molecular subtype, characterization of lymph node status, and prediction of pCR in view of neoadjuvant therapies. The use of automatic segmentation achieved comparable performances also in the radiomic analysis, suggesting it as a tool to obtain larger samples for advanced analysis. The increasing interest in radiomic studies prompts in parallel an increasing need for automatic unbiased tools to produce reliable and reproducible segmentation, showing overall the importance of this field.

In summary, to provide a general course of action, the most common steps to achieve a segmentation approach are outlined in Figure 3. First, it is essential to define the input information that enables the segmentation, selecting the modality and the spatial and temporal dimensions of the data. If the dataset size is large enough, multimodal imaging and information from multiple time points can enhance model capabilities, especially when dealing with lesions of different types, benefiting from enhancement kinetics and complementary information from various modalities. Ideally, 3D images better encode small lesions between different slices; however, this comes with the disadvantage of a larger computational burden, making the choice dependent on the available hardware. Once the input is clearly defined, it is of utmost importance to determine whether the model will use only lesion information or the full information of both lesion and normal areas. The suggestion is to use full information in both training and testing to ensure a consistent estimation of performance, enabling the model to be used both on pathological and normal cases without issues. To enhance information, as the lesion is usually smaller than the whole image, different preprocessing steps can be applied, such as extracting bounding boxes around the lesion, pre-segmenting the breast ROI, dividing the left and right breasts, and extracting even smaller patches of tissue. This could help any model in identifying local information for a problem where the target is significantly smaller than the whole image. When using multimodal images and different time points, it is also recommended to check and perform image registration. Further basic operations involve normalization and resampling to a standard spacing. Additional strategies that can improve model performance rely on data augmentation, involving not only geometric transformations (rotations, scaling, shifting, mirroring, elastic deformations) but also intensity-based transformations, such as histogram matching and other possible variations in the distribution of gray levels. These techniques can help improve the generalizability of the model, allowing it to handle multicentric data more effectively. Currently, the main segmentation approaches rely on supervised DL, however, unsupervised, semi-supervised, and combined techniques should be further investigated. Most cases rely on a single DNN; however, the possible use of multiple streams for detection, segmentation, and refinement

can be a potential alternative. Considering the training strategy, it is essential to carefully balance healthy and lesion inputs. In the case of supervised learning, data imbalance can be addressed by incorporating specific loss functions, such as Dice loss, focal loss, Tversky loss, or even combinations of these losses with basic ones, and using attention mechanisms to maintain focus on the lesion. Thereafter, as misleading tissues such as lymph nodes, microvessels, or small enhanced regions can sometimes lead to false positives and produce multiple fragments, thresholding based on size and analysis of connected components can be used as preprocessing. Moreover, morphological operations can further refine the final output. Ultimately, after validating the segmentation model, clinical applications can also be considered, using the produced segmentation as an input mask for radiomic models, lesion classification tasks, prediction of disease progression, and response to treatment, thereby contributing to improved precision and personalised medicine.

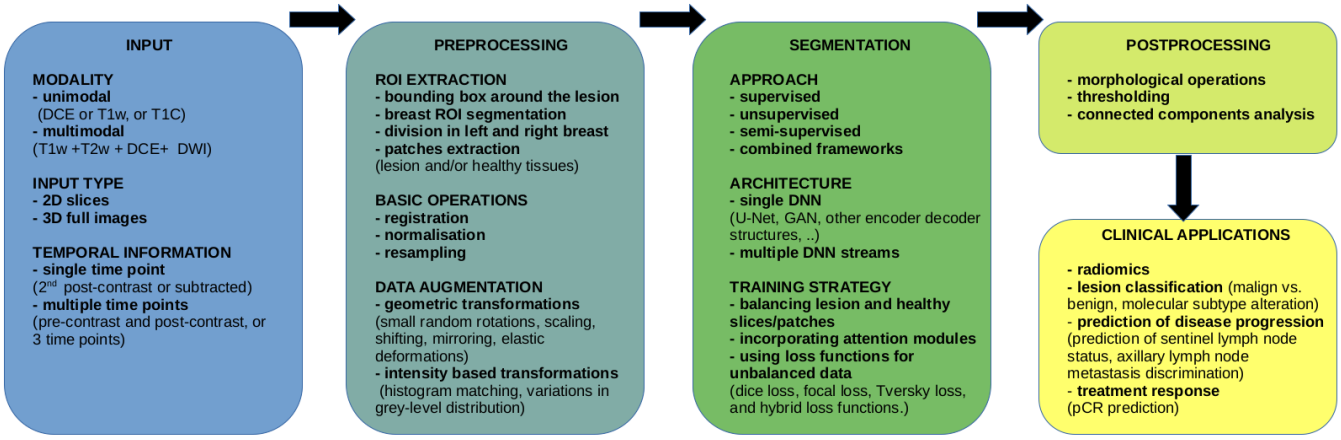


Figure 3. Summary of the main stages and options in the breast lesion segmentation process.

5. Conclusions

In conclusion, from the analysis of the existing corpus of studies, several fundamental gaps have emerged. In particular, the use of DL techniques demands increasingly large amounts of data for both training and testing. However, there is a significant lack of comprehensive, large public datasets that are essential for preventing and reducing bias, as well as improving model generalisation. Collaboration between different centres and institutions is necessary to share data and successfully validate models, paving the way for practical clinical applications. One of the main challenges in building large datasets is the time-consuming nature of manual segmentation required to provide accurate labels. A potential solution is to further investigate weakly supervised models that require minimal effort to produce a significantly smaller amount of labelled slices, along with the implementation of fully unsupervised deep learning approaches. Operator variability remains a source of bias in model training, further reinforcing the need to reduce human involvement by adopting unsupervised models capable of learning without strong human bias. Public datasets should include multicentric data from different scanners, offering variability in techniques, multiparametric data, and full images containing both normal tissue and lesion slices. Such datasets should also feature a variety of mass and non-mass lesions, variations in parenchymal composition, diverse ethnicities, cancers of small size, and patients with implants or different breast morphologies. To better capture inter-slice features and obtain comprehensive information, the use of 3D models with multiparametric data is recommended. However, this approach further underscores the need for larger and more diverse datasets. Achieving generalisability will require coordinated efforts to build shared datasets, along with the production of open-source codes and the availabil-

ity of pre-trained models. These resources could facilitate transfer learning and domain adaptation, ultimately advancing the field toward more robust and generalisable solutions. The landscape of ML applications for segmentation is vast. However, the possibility of any model being clinically applied must go through several essential steps. Before integration into clinical practice, further testing of the model's capabilities on external and public datasets is necessary to ensure bias reduction and to identify potential limitations in generalisation. Once models are properly validated and achieve CE/FDA marking, indicating their conformity, they should be implemented as custom modules in pre-existing local software. This ensures they are easy to use and effectively assist clinicians not only with diagnosis and prognosis but also by saving time.

This review reported the state of the art for breast lesion segmentation on MR images, exploring the studies published in the last decade, encompassing early unsupervised approaches, traditional classifiers, and diving into deep learning approaches. An increasing interest has grown towards deep learning approaches, especially CNN based on U-Net; however, comparison between studies is still difficult due to differences in images, protocols, and training procedures. However, this review has also some limitations, as it offers only a partial point of view, including results only from the PubMed database; a more complete overview on the topic should also include works that were not indexed on the PubMed database or account for a wider initial keyword selection.

Funding: This research was funded by Project—PNC-E3-2022-23683266—Advanced Diagnostic (HLS-DA), titled INNOVA

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADC	Apparent Diffusion Coefficient
AI	Artificial Intelligence
AUC	Area Under the Curve
BCE	Binary Cross-Entropy
CNN	Convolutional Neural Networks
CV	Cross-Validation
DCE	Dynamic Contrast Enhanced
DL	Deep Learning
DNN	Deep Neural Networks
DSC	Dice-Sørensen Coefficient
DWI	Diffusion-Weighted Imaging
FCM	Fuzzy C-Means
LOOCV	Leave-One-Out Cross-Validation
ML	Machine Learning
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
NAC	Neoadjuvant Chemotherapy
ODE	Ordinary Differential Equations
pCR	Pathological Complete Response
RF	Random Forest

ROI	Region of Interest
RUS	Random Under-Sampling
T1C	T1-Weighted with Contrast
T1w	T1-Weighted
T2w	T2-Weighted
TIC	Time-Intensity Curve

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Reis-Filho, J.S.; Pusztai, L. Gene expression profiling in breast cancer: Classification, prognostication, and prediction. *Lancet* **2011**, *378*, 1812–1823. [CrossRef] [PubMed]
3. Pace, L.E.; Keating, N.L. A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA* **2014**, *311*, 1327–1335. [CrossRef]
4. Morrow, M.; Waters, J.; Morris, E. MRI for breast cancer screening, diagnosis, and treatment. *Lancet* **2011**, *378*, 1804–1811. [CrossRef] [PubMed]
5. Radhakrishna, S.; Agarwal, S.; Parikh, P.M.; Kaur, K.; Panwar, S.; Sharma, S.; Dey, A.; Saxena, K.; Chandra, M.; Sud, S. Role of magnetic resonance imaging in breast cancer management. *South Asian J. Cancer* **2018**, *7*, 069–071. [CrossRef] [PubMed]
6. Martincich, L.; Montemurro, F.; De Rosa, G.; Marra, V.; Ponzone, R.; Cirillo, S.; Gatti, M.; Biglia, N.; Sarotto, I.; Sismondi, P.; et al. Monitoring response to primary chemotherapy in breast cancer using dynamic contrast-enhanced magnetic resonance imaging. *Breast Cancer Res. Treat.* **2004**, *83*, 67–76. [CrossRef] [PubMed]
7. Fangberget, A.; Nilsen, L.; Hole, K.H.; Holmen, M.; Engebraaten, O.; Naume, B.; Smith, H.J.; Olsen, D.R.; Seierstad, T. Neoadjuvant chemotherapy in breast cancer-response evaluation and prediction of response to treatment using dynamic contrast-enhanced and diffusion-weighted MR imaging. *Eur. Radiol.* **2011**, *21*, 1188–1199. [CrossRef]
8. Mann, R.M.; Cho, N.; Moy, L. Breast MRI: State of the art. *Radiology* **2019**, *292*, 520–536. [CrossRef] [PubMed]
9. Agrawal, G.; Su, M.Y.; Nalcioglu, O.; Feig, S.A.; Chen, J.H. Significance of breast lesion descriptors in the ACR BI-RADS MRI lexicon. *Cancer Interdiscip. Int. J. Am. Cancer Soc.* **2009**, *115*, 1363–1380. [CrossRef] [PubMed]
10. Mori, N.; Sheth, D.; Abe, H. Nonmass enhancement breast lesions: Diagnostic performance of kinetic assessment on ultrafast and standard dynamic contrast-enhanced MRI in comparison with morphologic evaluation. *Am. J. Roentgenol.* **2020**, *215*, 511–518. [CrossRef] [PubMed]
11. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **2016**, *278*, 563–577. [CrossRef] [PubMed]
12. Altabella, L.; Benetti, G.; Camera, L.; Cardano, G.; Montemezzi, S.; Cavedon, C. Machine learning for multi-parametric breast MRI: Radiomics-based approaches for lesion classification. *Phys. Med. Biol.* **2022**, *67*, 15TR01. [CrossRef] [PubMed]
13. Sutton, E.J.; Onishi, N.; Fehr, D.A.; Dashevsky, B.Z.; Sadinski, M.; Pinker, K.; Martinez, D.F.; Brogi, E.; Braunstein, L.; Razavi, P.; et al. A machine learning model that classifies breast cancer pathologic complete response on MRI post-neoadjuvant chemotherapy. *Breast Cancer Res.* **2020**, *22*, 1–11. [CrossRef]
14. Montemezzi, S.; Benetti, G.; Bisighin, M.V.; Camera, L.; Zerbato, C.; Caumo, F.; Fiorio, E.; Zanelli, S.; Zuffante, M.; Cavedon, C. 3T DCE-MRI radiomics improves predictive models of complete response to neoadjuvant chemotherapy in breast cancer. *Front. Oncol.* **2021**, *11*, 630780. [CrossRef] [PubMed]
15. Li, Y.; Fan, Y.; Xu, D.; Li, Y.; Zhong, Z.; Pan, H.; Huang, B.; Xie, X.; Yang, Y.; Liu, B. Deep learning radiomic analysis of DCE-MRI combined with clinical characteristics predicts pathological complete response to neoadjuvant chemotherapy in breast cancer. *Front. Oncol.* **2023**, *12*, 1041142. [CrossRef] [PubMed]
16. Granzier, R.W.; Verbakel, N.M.; Ibrahim, A.; Van Timmeren, J.; Van Nijnatten, T.; Leijenaar, R.; Lobbes, M.; Smidt, M.; Woodruff, H. MRI-based radiomics in breast cancer: Feature robustness with respect to inter-observer segmentation variability. *Sci. Rep.* **2020**, *10*, 14163. [CrossRef] [PubMed]
17. Pistel, M.; Brock, L.; Laun, F.B.; Erber, R.; Weiland, E.; Uder, M.; Wenkel, E.; Ohlmeyer, S.; Bickelhaupt, S. Stability of Radiomic Features against Variations in Lesion Segmentations Computed on Apparent Diffusion Coefficient Maps of Breast Lesions. *Diagnostics* **2024**, *14*, 1427. [CrossRef]
18. Zhao, X.; Bai, J.W.; Guo, Q.; Ren, K.; Zhang, G.J. Clinical applications of deep learning in breast MRI. *Biochim. Biophys. Acta (BBA)-Rev. Cancer* **2023**, *1878*, 188864. [CrossRef]
19. Balkenende, L.; Teuwen, J.; Mann, R.M. Application of deep learning in breast cancer imaging. In *Proceedings of the Seminars in Nuclear Medicine*; Elsevier: Amsterdam, The Netherlands, 2022; Volume 52, pp. 584–596.

20. Lei, Y.M.; Yin, M.; Yu, M.H.; Yu, J.; Zeng, S.E.; Lv, W.Z.; Li, J.; Ye, H.R.; Cui, X.W.; Dietrich, C.F. Artificial intelligence in medical imaging of the breast. *Front. Oncol.* **2021**, *11*, 600557. [CrossRef] [PubMed]
21. Reig, B.; Heacock, L.; Geras, K.J.; Moy, L. Machine learning in breast MRI. *J. Magn. Reson. Imaging* **2020**, *52*, 998–1018. [CrossRef] [PubMed]
22. Mehta, P.; Bukov, M.; Wang, C.H.; Day, A.G.; Richardson, C.; Fisher, C.K.; Schwab, D.J. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **2019**, *810*, 1–124. [CrossRef]
23. Dhanachandra, N.; Manglem, K.; Chanu, Y.J. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Comput. Sci.* **2015**, *54*, 764–771. [CrossRef]
24. Christ, M.J.; Parvathi, R. Fuzzy c-means algorithm for medical image segmentation. In Proceedings of the 2011 3rd International Conference on Electronics Computer Technology, Kanyakumari, India, 8–10 April 2011; Volume 4, pp. 33–36.
25. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [CrossRef] [PubMed]
26. Rayed, M.E.; Islam, S.S.; Niha, S.I.; Jim, J.R.; Kabir, M.M.; Mridha, M. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Inform. Med. Unlocked* **2024**, *47*, 101504. [CrossRef]
27. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 1–28. [CrossRef] [PubMed]
28. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **2020**, *295*, 328–338. [CrossRef] [PubMed]
29. Pasini, G.; Russo, G.; Mantarro, C.; Bini, F.; Richiusa, S.; Morgante, L.; Comelli, A.; Russo, G.I.; Sabini, M.G.; Cosentino, S.; et al. A Critical Analysis of the Robustness of Radiomics to Variations in Segmentation Methods in 18F-PSMA-1007 PET Images of Patients Affected by Prostate Cancer. *Diagnostics* **2023**, *13*, 3640. [CrossRef] [PubMed]
30. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
31. Jayender, J.; Chikarmane, S.; Jolesz, F.A.; Gombos, E. Automatic segmentation of invasive breast carcinomas from dynamic contrast-enhanced MRI using time series analysis. *J. Magn. Reson. Imaging* **2014**, *40*, 467–475. [CrossRef]
32. McClymont, D.; Mehnert, A.; Trakic, A.; Kennedy, D.; Crozier, S. Fully automatic lesion segmentation in breast MRI using mean-shift and graph-cuts on a region adjacency graph. *J. Magn. Reson. Imaging* **2014**, *39*, 795–804. [CrossRef]
33. Hu, L.; Cheng, Z.; Wang, M.; Song, Z. Image manifold revealing for breast lesion segmentation in DCE-MRI. *Bio-Med. Mater. Eng.* **2015**, *26*, S1353–S1360. [CrossRef] [PubMed]
34. Vogl, W.D.; Pinker, K.; Helbich, T.H.; Bickel, H.; Grabner, G.; Bogner, W.; Gruber, S.; Bago-Horvath, Z.; Dubsky, P.; Langs, G. Automatic segmentation and classification of breast lesions through identification of informative multiparametric PET/MRI features. *Eur. Radiol. Exp.* **2019**, *3*, 1–13. [CrossRef] [PubMed]
35. Min, H.; McClymont, D.; Chandra, S.S.; Crozier, S.; Bradley, A.P. Automatic lesion detection, segmentation and characterization via 3D multiscale morphological sifting in breast MRI. *Biomed. Phys. Eng. Express* **2020**, *6*, 065027. [CrossRef] [PubMed]
36. Spuhler, K.D.; Ding, J.; Liu, C.; Sun, J.; Serrano-Sosa, M.; Moriarty, M.; Huang, C. Task-based assessment of a convolutional neural network for segmenting breast lesions for radiomic analysis. *Magn. Reson. Med.* **2019**, *82*, 786–795. [CrossRef]
37. Galli, A.; Marrone, S.; Piantadosi, G.; Sansone, M.; Sansone, C. A pipelined tracer-aware approach for lesion segmentation in breast DCE-MRI. *J. Imaging* **2021**, *7*, 276. [CrossRef]
38. Vidal, J.; Vilanova, J.C.; Martí, R. A U-Net Ensemble for breast lesion segmentation in DCE MRI. *Comput. Biol. Med.* **2022**, *140*, 105093.
39. Rahimpour, M.; Saint Martin, M.J.; Frouin, F.; Akl, P.; Orlhac, F.; Koole, M.; Malhaire, C. Visual ensemble selection of deep convolutional neural networks for 3D segmentation of breast tumors on dynamic contrast enhanced MRI. *Eur. Radiol.* **2023**, *33*, 959–969. [CrossRef] [PubMed]
40. Qin, C.; Lin, J.; Zeng, J.; Zhai, Y.; Tian, L.; Peng, S.; Li, F. Joint Dense Residual and Recurrent Attention Network for DCE-MRI Breast Tumor Segmentation. *Comput. Intell. Neurosci.* **2022**, *2022*, 3470764. [CrossRef]
41. Zhou, H.; Hua, Z.; Gao, J.; Lin, F.; Chen, Y.; Zhang, S.; Zheng, T.; Wang, Z.; Shao, H.; Li, W.; et al. Multitask Deep Learning-Based Whole-Process System for Automatic Diagnosis of Breast Lesions and Axillary Lymph Node Metastasis Discrimination from Dynamic Contrast-Enhanced-MRI: A Multicenter Study. *J. Magn. Reson. Imaging* **2023**, *59*, 1710–1722. [CrossRef]
42. Zhu, J.; Geng, J.; Shan, W.; Zhang, B.; Shen, H.; Dong, X.; Liu, M.; Li, X.; Cheng, L. Development and validation of a deep learning model for breast lesion segmentation and characterization in multiparametric MRI. *Front. Oncol.* **2022**, *12*, 946580. [CrossRef]
43. Liu, H.Q.; Lin, S.Y.; Song, Y.D.; Mai, S.Y.; Yang, Y.D.; Chen, K.; Wu, Z.; Zhao, H.Y. Machine learning on MRI radiomic features: Identification of molecular subtype alteration in breast cancer after neoadjuvant therapy. *Eur. Radiol.* **2023**, *33*, 2965–2974. [CrossRef] [PubMed]

44. Zhang, J.; Zhan, C.; Zhang, C.; Song, Y.; Yan, X.; Guo, Y.; Ai, T.; Yang, G. Fully automatic classification of breast lesions on multi-parameter MRI using a radiomics model with minimal number of stable, interpretable features. *La Radiol. Med.* **2023**, *128*, 160–170. [CrossRef] [PubMed]
45. Liu, Y.; Chen, Z.; Chen, J.; Shi, Z.; Fang, G. Pathologic complete response prediction in breast cancer lesion segmentation and neoadjuvant therapy. *Front. Med.* **2023**, *10*, 1188207. [CrossRef] [PubMed]
46. Douglas, L.; Bhattacharjee, R.; Fuhrman, J.; Drukker, K.; Hu, Q.; Edwards, A.; Sheth, D.; Giger, M. U-Net breast lesion segmentations for breast dynamic contrast-enhanced magnetic resonance imaging. *J. Med. Imaging* **2023**, *10*, 064502. [CrossRef]
47. Zhang, Y.; Chan, S.; Park, V.Y.; Chang, K.T.; Mehta, S.; Kim, M.J.; Combs, F.J.; Chang, P.; Chow, D.; Parajuli, R.; et al. Automatic detection and segmentation of breast cancer on MRI using mask R-CNN trained on non–fat-sat images and tested on fat-sat images. *Acad. Radiol.* **2022**, *29*, S135–S144. [CrossRef]
48. Peng, C.; Zhang, Y.; Meng, Y.; Yang, Y.; Qiu, B.; Cao, Y.; Zheng, J. LMA-Net: A lesion morphology aware network for medical image segmentation towards breast tumors. *Comput. Biol. Med.* **2022**, *147*, 105685. [CrossRef]
49. Sun, L.; Zhang, Y.; Liu, T.; Ge, H.; Tian, J.; Qi, X.; Sun, J.; Zhao, Y. A collaborative multi-task learning method for BI-RADS category 4 breast lesion segmentation and classification of MRI images. *Comput. Methods Programs Biomed.* **2023**, *240*, 107705. [CrossRef] [PubMed]
50. Ru, J.; Lu, B.; Chen, B.; Shi, J.; Chen, G.; Wang, M.; Pan, Z.; Lin, Y.; Gao, Z.; Zhou, J.; et al. Attention guided neural ODE network for breast tumor segmentation in medical images. *Comput. Biol. Med.* **2023**, *159*, 106884. [CrossRef] [PubMed]
51. Chen, Q.; Zhang, J.; Meng, R.; Zhou, L.; Li, Z.; Feng, Q.; Shen, D. Modality-Specific Information Disentanglement From Multi-Parametric MRI for Breast Tumor Segmentation and Computer-Aided Diagnosis. *IEEE Trans. Med. Imaging* **2024**, *43*, 1958–1971. [CrossRef] [PubMed]
52. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
53. Vrudhula, A.; Kwan, A.C.; Ouyang, D.; Cheng, S. Machine Learning and Bias in Medical Imaging: Opportunities and Challenges. *Circ. Cardiovasc. Imaging* **2024**, *17*, e015495. [CrossRef] [PubMed]
54. Zhao, X.; Liao, Y.; Xie, J.; He, X.; Zhang, S.; Wang, G.; Fang, J.; Lu, H.; Yu, J. BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification. *Comput. Biol. Med.* **2023**, *164*, 107255. [CrossRef]
55. Li, X.; Abramson, R.G.; Arlinghaus, L.R.; Chakravarthy, A.B.; Abramson, V.G.; Sanders, M.; Yankeelov, T.E. Data from QIN-breast. *Cancer Imaging Arch.* **2016**, *5*. [CrossRef]
56. Huang, W.; Li, X.; Chen, Y.; Li, X.; Chang, M.C.; Oborski, M.J.; Malyarenko, D.I.; Muzi, M.; Jajamovich, G.H.; Fedorov, A.; et al. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: A multicenter data analysis challenge. *Transl. Oncol.* **2014**, *7*, 153–166. [CrossRef] [PubMed]
57. Diogo, P.; Morais, M.; Calisto, F.M.; Santiago, C.; Aleluia, C.; Nascimento, J.C. Weakly-supervised diagnosis and detection of breast cancer using deep multiple instance learning. In Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 18–21 April 2023; pp. 1–4.
58. Xun, S.; Li, D.; Zhu, H.; Chen, M.; Wang, J.; Li, J.; Chen, M.; Wu, B.; Zhang, H.; Chai, X.; et al. Generative adversarial networks in medical image segmentation: A review. *Comput. Biol. Med.* **2022**, *140*, 105063. [CrossRef] [PubMed]
59. Xia, X.; Kulis, B. W-net: A deep model for fully unsupervised image segmentation. *arXiv* **2017**, arXiv:1711.08506.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.