Computers in Biology and Medicine 140 (2022) 105093

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# A U-Net Ensemble for breast lesion segmentation in DCE MRI

Roa'a Khaled [a], Joel Vidal [a], Joan C Vilanova [b,c,d], Robert Martí [a,*]

[a] Computer Vision and Robotics Institute, University of Girona, Campus Montilivi, Girona, 17003, Spain
[b] Department of Radiology, Clinica Girona, Girona, 17002, Spain
[c] Institute for Diagnostic Imaging (IDI), Girona, 17007, Spain
[d] Faculty of Medicine, University of Girona, Girona, 17003, Spain

ARTICLE INFO

*Keywords:*
Breast lesions segmentation
DCE-MRI
Deep learning
3D U-Net
Ensemble methods
Breast cancer

ABSTRACT

Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) has been recognized as an effective tool for Breast Cancer (BC) diagnosis. Automatic BC analysis from DCE-MRI depends on features extracted particularly from lesions, hence, lesions need to be accurately segmented as a prior step. Due to the time and experience required to manually segment lesions in 4D DCE-MRI, automating this task is expected to reduce the workload, reduce observer variability and improve diagnostic accuracy.

In this paper we propose an automated method for breast lesion segmentation from DCE-MRI based on a U-Net framework. The contributions of this work are the proposal of a modified U-Net architecture and the analysis of the input DCE information. In that sense, we propose the use of an ensemble method combining three U-Net models, each using a different input combination, outperforming all individual methods and other existing approaches.

For evaluation, we use a subset of 46 cases from the TCGA-BRCA dataset, a challenging and publicly available dataset not reported to date for this task. Due to the incomplete annotations provided, we complement them with the help of a radiologist in order to include secondary lesions that were not originally segmented. The proposed ensemble method obtains a mean Dice Similarity Coefficient (DSC) of 0.680 (0.802 for main lesions) which outperforms state-of-the art methods using the same dataset, demonstrating the effectiveness of our method considering the complexity of the dataset.

## 1. Introduction

Breast cancer (BC) begins with an uncontrolled change and division of cells in the breast, forming a mass (lesion) that can either grow and spread to other parts of the body (as the case of a malignant lesion), or just grow without spreading (as the case of a benign lesion). It is most easy to treat BC when the lesion is small, however, no symptoms normally appear at that stage. Hence, screening is very crucial for the early detection [28].

According to the World Health Organization, BC is impacting 2.1 million women each year and causing the greatest number of cancer deaths among women. In the US, according to the estimates of the American Cancer Society (ACS) for the year 2020, BC cases were expected to form 30% of all diagnosed cancer cases and 15% of all cancer deaths were expected to be caused by BC [27]. These statistics indicate that the spread of BC is indeed one of the main health challenges in the world. Despite that, statistics in the US have shown an increase in the five-year survival rate from 75% to 91% between 1975 and 2015, as well as a continuous decrease in death rate. This is mostly due to the early detection of BC and the expanding access to high-quality prevention and treatment services [27].

Imaging modalities have been playing a vital role in all phases of BC management, starting from screening and early detection to diagnosis and treatment follow-up. In addition, due to the fact that cancer is a complex disease with varied pathology, improvements of existing techniques and new imaging modalities have been continually introduced in order to improve the detection efficiency and hence BC outcomes and survival [28]. However, each of these modalities has different clinical advantages and disadvantages. The choice of the modalities and techniques is also affected by the patient's stage, age and the density of the breast tissue. Currently, the main clinical breast imaging modalities used for BC detection and diagnosis are: Mammography, Ultrasound, and Magnetic Resonance Imaging (MRI). Although other breast diagnostic methods exist, such as: tomosynthesis, elastography,

* Corresponding author.
*E-mail addresses:* rua.khaled@yahoo.com (R. Khaled), joel.vidal@udg.edu (J. Vidal), kvilanova@comg.cat (J.C. Vilanova), robert.marti@udg.edu (R. Martí).

https://doi.org/10.1016/j.compbiomed.2021.105093
Received 3 July 2021; Received in revised form 26 November 2021; Accepted 26 November 2021
Available online 30 November 2021

photoacoustics, and optical imaging, they have different degrees of adoption due to technological and clinical limitations, hence they are not as widely used as the main ones mentioned earlier.

Mammography is currently the gold standard method for BC screening and early detection. However, its sensitivity (true positive rate) declines from 75% to 50% in middle aged patients who have increased breast density. On the other hand, ultrasound imaging is used as an adjunct tool to mammography to detect the location and nature of the suspicious lesion, which improves the diagnostic yield for women with dense breasts and those at higher risk of BC, but at the expense of being operator dependent and having an increased false positive rate [28]. MRI is widely used for both the early detection and diagnosis of BC. It has higher sensitivity than mammography and ultrasound, specially for particular patient groups, such as higher risk women and women with dense breasts. Moreover, it allows the simultaneous evaluation of both breasts and has no side effects as there is no radiation involved [28].

Dynamic contrast enhanced MRI (DCE-MRI) has been adopted due to its effectiveness for the diagnosis of BC as it visualizes both physiological tissue characteristics and anatomical structures. However, it is less specific (has more false positives) compared to mammography [33]. In DCE-MRI the changes of T1 in tissues are measured in order to observe and quantify the contrast enhancement over time after the administration of a contrast agent (Gadolinium). The change of contrast enhancement depends on several factors, such as: regional blood flow, size and number of blood vessels, and their permeability, which are strongly related to cancer tissues.

Nevertheless, DCE-MRI analysis is time consuming and requires experienced radiologists to evaluate and interpret the large amount of 4D information for each patient. Therefore, many methods have been developed to automatically extract features and interpret those images. Proposed features include lesion morphology, texture, and enhancement kinetics and have been proven by recent studies to be useful for the identification of genomic composition of BC lesions and for patient outcomes prediction [1,20,33]. However, the extraction of these features requires the lesions to be accurately segmented first. Therefore, the accurate segmentation of breast lesions in DCE-MRI is a critically significant task for automated BC analysis, diagnosis and treatment follow-up [33].

The most straightforward way to achieve this task is to manually annotate lesion regions by radiologists, but this is also time-consuming and error-prone. Therefore, automating this challenging task will help radiologists to reduce the high manual workload and obtain more accurate lesion segmentation.

In this work we propose an automated segmentation method for breast lesions in DCE-MRI. This method is based on our previous work in which a ROI guided, 3D patch based U-Net framework was proposed [19]. In this paper we improve the method by using a modified U-Net architecture that incorporates residual basic blocks. Additionally, we investigate the use of different inputs and propose an ensemble method combining three different models each of them utilizes a different combination of inputs. Furthermore, we complement (and make publicly available) the provided annotations with the help of an experienced radiologist in order to address the problem of having incomplete annotations for cases with multiple lesions. Finally, a comparison with a state-of-the-art approach on the same task has been made and results show that a better performance is achieved using the proposed method.

The remainder of this paper is structured as follows: in section 2 related works from the state-of-the-art are outlined. In section 3 we describe the dataset used and the modifications made in order to complement the annotations. In section 4 our proposed method is presented. In section 5 the obtained results are analysed and discussed. Finally, in section 6 we present our conclusions and future work.

## 2. State of the art

Automating the task of breast lesion segmentation in DCE-MRI is a challenging problem and an active area of research. Existing methods for lesion segmentation in general fall into two categories (or combinations of them): 1) Semi-automatic methods, and 2) Learning-based methods.

One of the challenges to automate the task of breast lesion segmentation is the difficulty of identifying them from confounding organs or vessels. In this sense, semi-automatic methods seem to avoid these problems at the expense of the need of user intervention as radiologists need to define lesion regions first (bounding boxes) to make the automatic segmentation task easier [2,29,35]. In fact, there are only a few studies focusing on breast lesion segmentation in DCE-MRI, and in most of them semi-automatic methods were used.

On the other hand, learning-based methods perform automatic lesion segmentation using supervised learning algorithms and they have achieved remarkable performance in many medical applications. Those methods can be further classified into two types: 1) Traditional Machine Learning (ML) methods, and 2) Deep Learning (DL) methods.

### 2.1. Traditional Machine Learning (ML) methods

In traditional ML methods, feature extraction and model training are considered as two separate tasks. This implies that a human intervention is needed in order to choose and hand-engineer the features that are required to train the model. In addition, there is a need to manually choose the model (classifier) to be trained. The algorithm (features and classifier) can be adjusted multiple times until the required results are obtained, at the expense of being less robust and generalisable to unseen data.

Several traditional ML methods have been proposed for breast lesion segmentation but only few of them are focusing on DCE-MRI. Many studies have been conducted in order to identify the definitive set of features and segmentation model for DCE-MRI and in most of these studies similar approaches have been utilized. For instance, kinetic features have been proven in many studies as the most effective way to perform lesion segmentation by means of ML. Kinetic features are defined as characteristics modelling the shape of the Time Intensity Curve (TIC) and apart from segmentation, further analysis of TIC is commonly used to provide several parameters useful for lesion diagnosis. TIC is obtained either for each voxel or for regions of interest, and it shows the absorption and the release of the contrast agent over time according to the vascularisation characteristics of the tissue [21]. Washout and plateau patterns (along with rapid up-slope in the early phase) in TIC are more likely to be associated with malignancy, whereas a persistent pattern is usually linked with benign lesions [8]. Fig. 1 illustrates the difference between TIC of normal tissues and lesions.

Examples of works utilizing kinetic features include the work proposed in Ref. [16], in which authors present an automated localization method for breast lesions in DCE-MRI, by first extracting blob and relative enhancement voxel features for locating initial lesion candidates and then computing a malignancy score for each lesion candidate using region based morphological and kinetic features in order to reduce false positives. In this study, authors investigated the use of different classifiers in the second stage, where the random forests classifier outperformed LDA, kNN, gentleboost and SVM.

Another automatic approach for DCE-MRI breast lesion segmentation was proposed in Ref. [18]. In this approach a high dimensional dataset was first built by collecting the TIC of every pixel in the ROI. Then a nonlinear dimensionality reduction technique (Laplacian Eigenmaps) was employed to map the TIC data from the higher dimensional space to 3 -dimensional space. In other words, the DCE series of one image slice was represented by a 3D image which was then represented as an RGB feature image. Finally, k-means clustering technique was performed for lesion segmentation.

In [7] a fuzzy c-means (FCM) clustering-based method was proposed

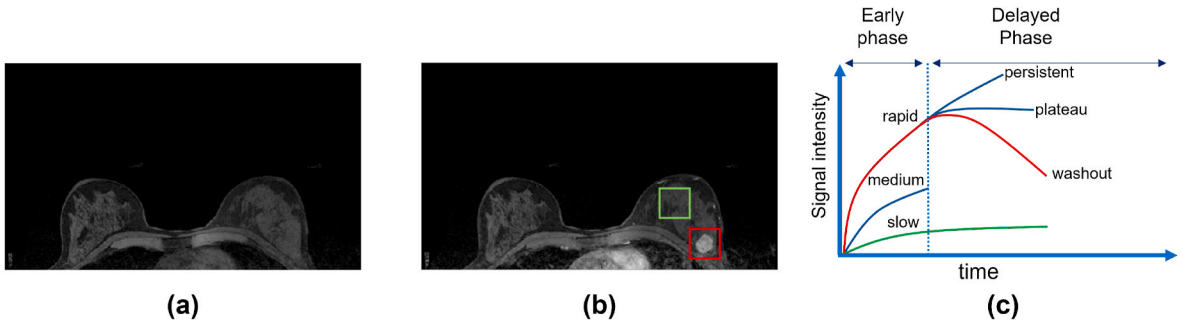


**Fig. 1.** Illustration of the difference between Time Intensity Curves (TIC) of normal tissue (in green) and a malignant lesion (in red). (a) Pre-contrast volume. (b) Second post-contrast volume. (c) Different types of TIC from DCE-MRI including the ones for normal tissue (in green) and a malignant lesion (in red).

for the 3D segmentation of breast lesions in DCE-MRI after performing lesion enhancement within the selected ROI. Lesion enhancement was performed by dividing the intensity value at each voxel in the post-contrast series by the intensity value at the corresponding pre-contrast voxel.

In [30] authors proposed an automated method for both lesion segmentation and classification. The proposed method incorporated a random forest (RF) classifier combined with mpPET/DCE-MRI intensity-based features for lesion segmentation, whereas shape, kinetic and spatio-temporal texture features were utilized for lesion classification.

Despite the good performance reported by traditional ML methods, there are still many limitations that need to be tackled. For instance, most of these works perform 2D segmentation and fail to obtain a good performance for 3D segmentation. Furthermore, almost all of these works segment a breast lesion from the minimum bounding box of the lesion as regions of interest (ROIs), which impose the need for experienced human intervention and, hence, there is still a need for a fully automatic method. On the other hand, if larger ROI is used with such methods, the accuracy will be significantly reduced since the large ROI will include many other surrounding enhanced tissues and the proposed traditional ML algorithms work well only when the ROI is small enough. Also, if larger ROI is used there is an increase in computational demands. Therefore, DL methods are expected to address such limitations.

### 2.2. Deep Learning (DL) based methods

Recently, there are more studies focusing on DL methods for breast lesion segmentation and diagnosis, and such methods have surpassed most traditional ML methods. In contrast to traditional ML methods, DL methods combine both feature extraction and model training into a single cohesive learning framework so that the segmentation task is performed in an end-to-end manner. Emerging Deep CNN (DCNN) models are capable of extracting salient features directly from the data, which means that manual feature design and its associated challenges are now obviated. In addition, with incorporating a combination of imaging and non-imaging digital data (such as patient-level information, and tumor-level information), DCCN models are now able to identify not only known correlations but also novel imaging biomarkers that have huge potential to enhance clinical performance. All that has become possible because of the breakthroughs in computer processing, data storage, and algorithm design in recent years [23,33].

When utilizing DL methods for segmentation tasks, voxel-wise (or pixel-wise) classification models are trained, in which cubic patches centered at a particular voxel can be first extracted, then a patch-wise binary classifier learns to classify voxels into either lesion or non lesion voxels [33]. In general, since the task of lesion segmentation can be considered as a semantic segmentation in which the input image has to be divided into Regions of Interest (ROIs), each referring to a lesion, any CNN segmentation model could be potentially used [21].

One of the first works to address semantic segmentation with CNN was SegNet [3]. SegNet is a deep convolutional encoder-decoder architecture, followed by a pixel-wise classifier. The role of the encoder network is to learn a compact representation of the input data, while the role of the decoder network is to map the encoded features to a segmentation mask. Similarly, the U-Net developed for biomedical image segmentation exploits an encoder-decoder architecture, enhanced by the presence of skipping connections between the two sides with the aim of exploiting encoding information to improve the decoding stage and to reduce the gradient vanishing problem [24]. This model has been widely used in the literature and several modifications have been proposed. Both SegNet and U-Net architectures provide the potential to produce accurate models even with relatively small datasets.

There are several DL based methods that have been recently proposed for breast image analysis, based on mammography, MRI, ultrasound, and whole-slide histology images. However, there are not so many works focusing on DCE-MRI. Moreover, the problems of both class imbalance and confounding regions (which are very common in breast lesion segmentation from DCE-MRI) are rarely taken into account [33].

Existing literature include the work in Ref. [6] where temporal and 3D features were extracted using three stacked parallel ConvLSTM networks over a 4-layer U-Net [26]. In Ref. [34], authors proposed a U-Net based method to segment breast lesions in DCE-MRI scans in both 2D and 3D settings, using a binary cross-entropy loss function. Obtained dice coefficients and false positives indicated a slightly better performance of the 3D network over the 2D one. Limitations of this study include using only second post-contrast scans and using lesion bounding boxes instead of the full-sized scans.

In [15] both SegNet and U-Net were used with a binary cross entropy loss function. Results showed better performance of U-Net over SegNet, which can be explained by the fact that SegNet is more adapted to multi classification tasks as in autonomous cars applications. Due to the lack of data, 2D slices were used instead of 3D volumes, which could be regarded as a limitation. Moreover, the ground truth labels were provided by only one radiologist without accounting for inter-reader variability.

A U-Net based architecture was also used in Ref. [21], which incorporated the well-known Three Time Points approach (3 TP) for the inputs. In the 3 TP approach (proposed by Ref. [13]), three well defined temporal acquisitions were proved to improve breast MRI lesion analysis (t0 = pre-contrast, t1 = 2 min after contrast agent injection, t2 = 6 min after contrast agent injection). In Ref. [21], the network was fed with images obtained at the three specific time points to take into account the fundamental characteristic of DCE-MRI. Segmentation was performed on slices, where the three temporal acquisitions of the same slice were used as channels within the image. Slices were extracted along the projection with the higher resolution (the coronal projection) and to obtain a reliable and fair evaluation, the slices from the same subject were always separated across the cross-validation folds. Additionally, the dice loss function was used.

Finally, authors in Ref. [33] proposed a mask-guided hierarchical

learning (MHL) U-Net framework in which the two issues of confounding organs and class imbalance were addressed. In order to eliminate confounding organs, 3D breast masks were first generated using a U-Net model. Then a coarse-to-fine segmentation was performed using a two-stage U-Net model. The first stage aims to generate over-segmented lesion-like regions using the post-contrast volumes and the difference volumes (between post and pre-contrast), and breast masks as inputs to the first U-Net. The second stage aims to refine the segmentations generated by the first stage using a second U-Net. Moreover, a Dice-Sensitivity-like loss function was proposed and used in the first stage U-Net in order to handle the class-imbalance problem, and in the second stage a Dice-like loss function and a reinforcement sampling strategy were used.

## 3. Materials

### 3.1. Data

For this study a subset of 46 cases from the TCGA-BRCA collection has been used, this subset has the Tissue Source Site code: BH. The TCGA-BRCA collection was collected by the TCGA Breast Phenotype Research Group and made available in The Cancer Imaging Archive (TCIA) [5,11]. All cases are diagnosed with BC by performing image-guided core needle biopsy, in other words, all cases have at least one lesion.

Volumes were acquired at the University of Pittsburgh Medical Center (1999–2004) prior to any treatment using a standard double breast coil on a 1.5T GE whole body MRI system (GE Medical Systems, Milwaukee, Wisconsin, USA). The imaging protocols included one pre and four to six post-contrast volumes obtained using a T1-weighted 3D spoiled gradient echo sequence with a gadolinium-based contrast agent (Omniscan; Nycomed-Amersham, Princeton, NJ). Typical in-plane resolution was 0.53–0.86 mm, and typical spacing between slices was 2–3 mm. The subset data used in this study has a coronal-sagittal size of 512x512 and the number of axial slices ranged between 85 and 112.

Each breast MRI examination was independently reviewed by three expert board-certified breast radiologists blinded to outcome data, each primary breast lesion was then segmented in 3D and subsequently validated. It is important to mention that most of the cases had multiple lesions according to the reviewer radiologists, however the Ground Truth (GT) segments only the primary lesion since the purpose of the TCGA/TCIA study was to map the radiomics (phenotypes) of the primary lesion to the corresponding clinical, histopathology, and genomic data.

Overall, this dataset is complex and challenging. Fig. 2 shows example cases of the subset data used in this study with lesions of various sizes, shapes, locations and intensity enhancements, in addition to multiple lesions (as in Fig. 2-(f)).

### 3.2. Data preparation

Each patient DCE-MRI series was provided as a DICOM file that combines all pre and post-contrast volumes as different channels. For easier use, pre and post-contrast volumes within the series were separated and saved as Nifti files using SimpleITK library. Later, the difference images between pre and post-contrast volumes were obtained and qualitatively analysed, and no significant differences (mis-alignment) due to patient movement could be found. Therefore, no registration was performed on the pre and post-contrast volumes.

As mentioned earlier, only the primary lesion in each case is segmented in the Ground Truth (GT). GT annotations were provided as lesion bounding boxes in the form of binary files, where each file contains the coordinates of the lesion bounding box with respect to the full volume in addition to the binary values (0 or 1) of the bounding box voxels. We use these binary files to generate GT annotations of full volume sizes.

#### 3.2.1. Complementing the annotations

According to the radiologists' reports provided with the dataset, many cases presented multiple lesions. However, only one lesion per case was segmented (primary lesion) since the data was not particularly made for the lesion segmentation task. Having incomplete GT annotations indeed affects the results and causes lower dice values as explained in our previous work [19]. Hence, we complement the annotations of the cases labelled by at least one radiologist (out of three) as multifocal or multicentric lesions with the help of an experienced radiologist in order to include all the other lesions that were not initially segmented (secondary lesions).

In total, the GT of 11 cases (out of 30 multifocal/multicentric cases) were complemented. GT complementing was done without any

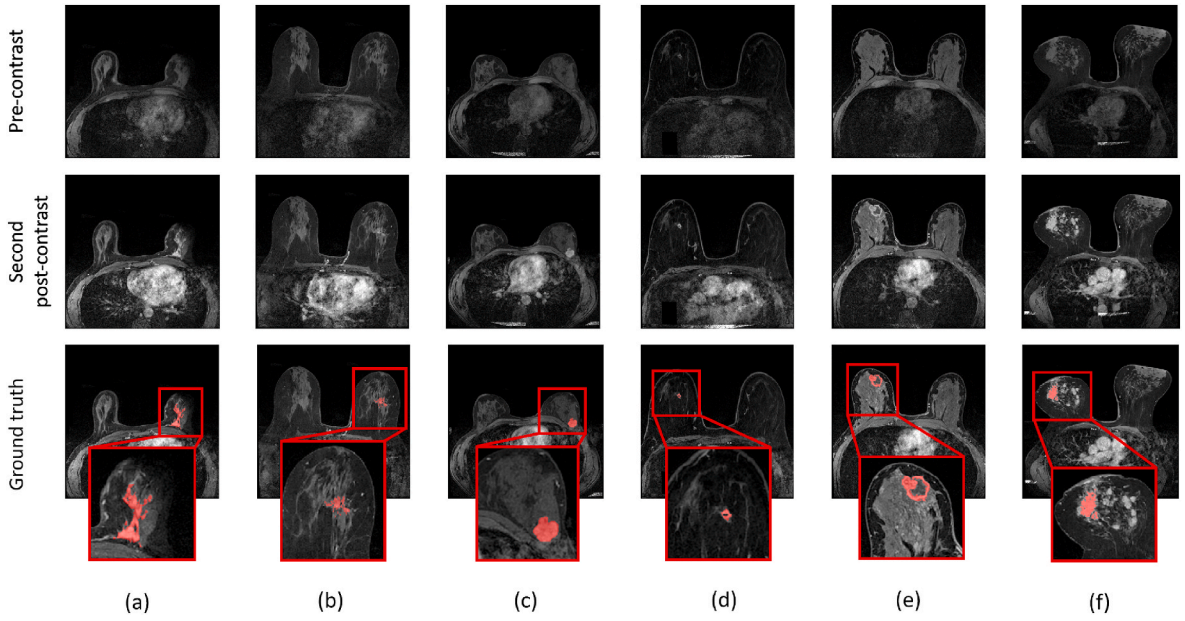


**Fig. 2.** Example cases of the used subset data with lesions of different sizes, shapes, locations, intensity enhancements and multiple lesions. Cases in (a) to (f) are: A0B5, A0DE, A0B6, A0H6, A0DZ and A0B1, respectively.

additional information from pathological and radiological reports, and only using MRI-DCE volumes and knowing that they were cases of multiple lesions. This helped to avoid bias of the annotations towards information not contained in the images. The complemented annotations were made available on: (https://github.com/ICEBERG-VICOROB/Breast-DCE-MRI-Completed-Annotations-).

Fig. 3 shows example cases with multiple lesions for which the GT has been complemented.

## 4. Methods

In this paper we propose an automated method for segmenting breast lesions in DCE-MRI using DL, that is an extension to a preliminary version of this work [19]. Our method is an ensemble of three models, each of them is based on the same 3D patch based modified U-Net, following the workflow shown in Fig. 4 but using a different combination of input volumes. The proposed method takes into account the problems of confounding organs and class imbalance by performing a balanced patch sampling technique restricted by an automatically generated breast ROI to ensure that the two classes are equally distributed in the training set and to avoid having patches from confounding organs. Fig. 4 shows an overview of the proposed method.

Moreover, we compare the proposed method with an existing approach proposed in a recent study where two hierarchical segmentation models are used. The obtained results are presented and discussed in Section 5.

### 4.1. Pre-processing and patch sampling

Several pre-processing steps are performed on input volumes prior to feeding them to the network, as proposed in our earlier work [19]. These steps are outlined in the following:

a) Breast ROI masks generation.
b) Zero padding with padding width equal to half of the patch size.
c) Zero-mean unit-variance intensity normalization.
d) Balanced patch extraction.

In the following subsections we discuss in detail some of these steps.

#### 4.1.1. Breast ROI masks generation

DCE-MRI includes confounding backgrounds such as: vessel structures and internal organs. Such regions mimic the lesions in terms of contrast agent permeability (i.e. high changes in intensity across the time), which makes the task of breast lesion segmentation more challenging and might increase FPs detection. Therefore, utilizing a region of interest (ROI) which includes the breast area only is an important step that has been exploited in many works in the literature. The most straightforward way is to use a breast mask as a ROI (as used in Ref. [33]) which can remove most of the confounding organs.

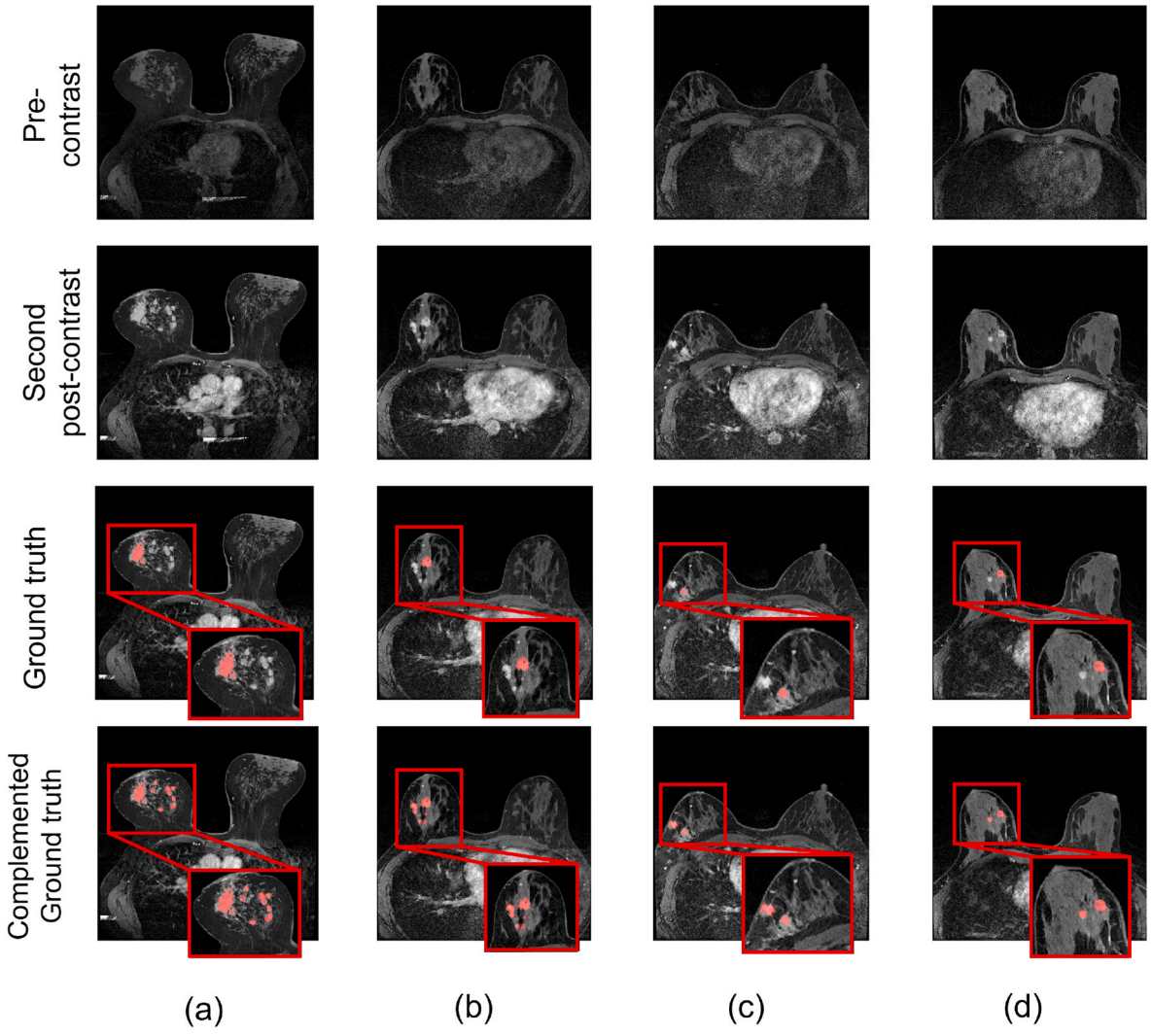


**Fig. 3.** GT before and after complementing for example cases with multiple lesions. Cases in (a) to (d) are: A0B1, A0AZ, A0DG and A0HA, respectively.

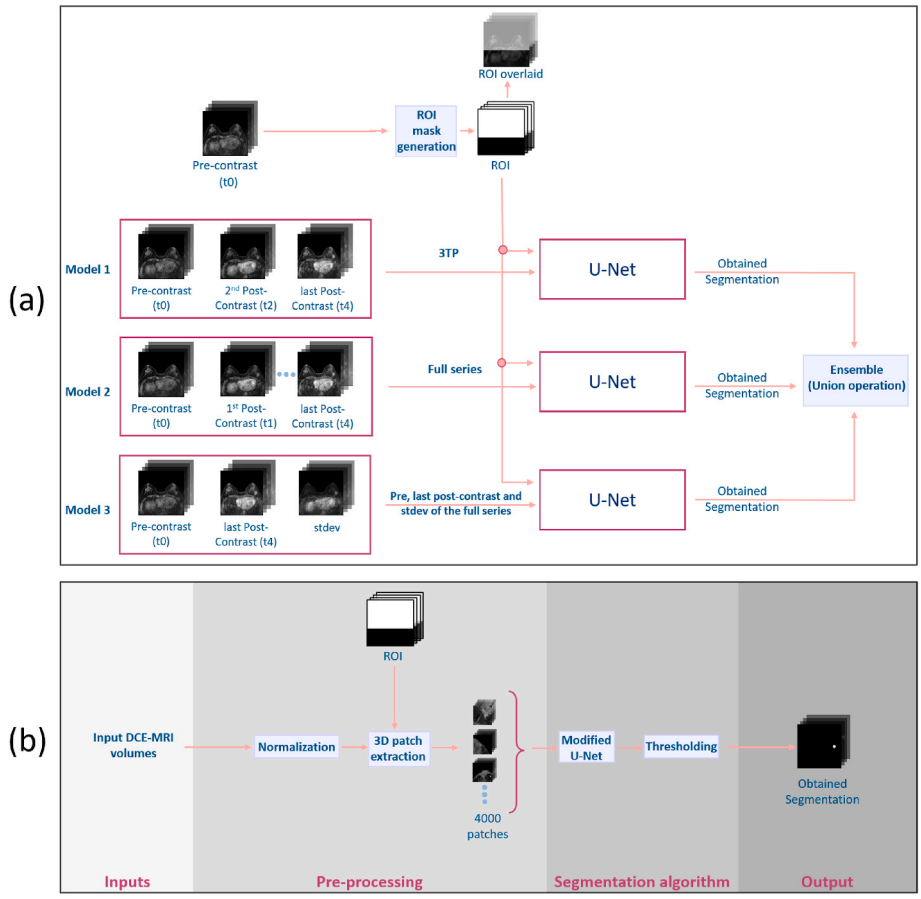


**Fig. 4.** Illustration of the proposed framework. (a) Ensemble of three models. (b) Proposed 3D patch based modified U-Net method used for each of the three models in (a).

Several methods have been proposed to generate breast masks such as thresholding, morphology, region growing, active contours, wavelet filtering and Atlas, or a combination of these methods [22,25,31]. However, breast segmentation is out of the scope of this work. Thus we implemented a simple method to generate ROI masks instead of breast masks, discarding as much of the confounding internal region as possible (such as heart and lung areas). The proposed method relies on a simple landmark detection method, in which we detect the skin-air boundary between the two breasts and then exclude the non-breast part of the volume that lies beyond the detected landmark. This method has been further explained in detail in our previous work [19].

#### 4.1.2. *Balanced patch extraction*

In a uniform patch sampling strategy, patches are extracted from all parts of an image uniformly. However, in the context of lesions segmentation, uniform patch sampling results in a very common issue of class imbalance. This is due to lesions being in general very small compared to the rest of the volume size, hence the number of voxels in the lesion region (positive class) is much smaller than that in the background (negative class) and the number of patches extracted from the lesion class will be very small, which eventually leads to a poor performance of the network and misclassification of lesion voxels. Several studies on lesion segmentation of different organs have addressed this issue [4,9,32].

In [17] authors proposed a method to address the class imbalance issue in which the extracted patches always contain lesion voxels and were randomly shifted so that the center of the patch does not necessarily be a lesion voxel. Another method proposed in Ref. [12], which utilized a balanced sampling strategy such that for each image there are an equal number of patches representing both classes. Additionally, a ROI restricted technique was proposed in which negative patch extraction was restricted to be from a ROI and not background regions.

In our work we utilize a method similar to what is proposed in Ref. [12], such that for each image there are an equal number of patches representing the lesion (positive) minority class, and the background (negative) class. First, the number of positive patches was set to be 2000 per case. Then, the GT was used to obtain coordinates of only positive voxels, which were set as initial patch centers. To ensure having 2000 patches, those obtained coordinates were either replicated or truncated. Later, patch centers were shifted randomly, with shifts set to be less than half of the patch size. Finally, positive patches were extracted using the obtained shifted coordinates of patch centers. Second, the same number (2000) of negative patches were similarly extracted. Furthermore, the patch extraction was restricted to be within the previously generated breast ROI, to avoid the region of confounding organs.

### 4.2. *Segmentation algorithm*

The segmentation algorithm is based on an ensemble of three U-Net models, with a similar architecture but different input information from the DCE-MRI patches. Individual results of each model are then combined into a single segmentation by a simple union operation of the segmentations, as illustrated in Fig. 4-(a). The first model's input is based on the Three Time Point acquisitions (3 TP) proposed in Ref. [21] which, as mentioned earlier, includes the pre-contrast, second post-contrast (2 min) and last post-contrast (6 min). The second model's input consists of the full series provided (i.e. one pre-contrast and four post-contrast volumes). Finally, for the third model's input we propose to include the pre-contrast and last post-contrast volumes, with an additional volume computed from the standard deviation (stdev) of the

intensity signals of the whole DCE-MRI volume series, with the aim to better represent the time-intensity variation. In other words, stdev was obtained voxel wise across the time dimension.

Regarding each individual U-Net model, we propose to use a modified 3D U-Net architecture. U-Net is an encoder-decoder architecture originally designed for biomedical electron microscopy (EM) images multi-class pixel-wise semantic segmentation [10,24]. In the proposed U-Net architecture we make some alterations to the basic U-Net architecture; we use four levels (blocks) in each of the encoding and decoding paths and replace the U-Net convolutional blocks with residual basic blocks. This architecture is illustrated in Fig. 5.

First, data were taken from different temporal volumes and used as different channels of the input, which was then fed into the network. As illustrated in Fig. 5, every step in the contracting path consists of a residual basic block followed by a $(2 \times 2 \times 2)$ max-pooling with stride $= 2$ for down-sampling. Then the contracting path is followed by a latent space consisting of a residual basic block with Rectified Linear Unit (ReLU) activation and an instance normalization. Similarly, every step in the expanding path consists of a $(2 \times 2 \times 2)$ up-convolution with stride $= 2$ followed by concatenation with the feature map from the corresponding level of the contracting path and then followed by a residual basic block. Finally, there is a $(1 \times 1 \times 1)$ output convolution layer with two output channels followed by a softmax layer which returns probabilities for each class.

In our proposed algorithm we use the binary cross-entropy loss function, AdaDelta optimizer, and a threshold of 0.5 for generating the output segmentation. We extract 4000 balanced patches per case with a size of (32,32,32) and a sampling step of (16,16,16). This configuration is based on results from our previous work in which we investigate different parameters [19]. We use this configuration for all experiments explained in section 5 except for 1 and 3 where we use a single model (with pre-contrast, last post-contrast and stdev as inputs) for a simpler method comparison.

## 5. Results and discussion

In this section we report the different experiments performed and the obtained results along with the subsequent discussion. We first investigate the impact of volume resampling from original to isotropic voxel sizes. Then we investigate the impact of different combinations of input volumes as well as the proposed ensemble of three models, each utilizing a different combination of inputs. Additionally, we compare the performance of our proposed architecture with another two different U-Net based architectures: (1) The basic U-Net used in our earlier work [19], and (2) The two hierarchical basic U-Nets proposed in Ref. [33].

Finally, as the annotations were complemented (as explained in section 3.2.1) we compare the performance when using both the original (incomplete) annotations and the complemented ones.

All experiments are performed using 5 fold cross-validation across the provided 46 cases, with 20 epochs per fold. In each fold, 9 cases are used for testing (10 cases in the last fold) and the remaining cases are shuffled and divided into 80% for training and 20% for evaluation. By doing so, we obtain lesion segmentation results for each of the 46 cases.

As an evaluation criteria, we use the Dice Similarity Coefficient (DSC) since it is one of the most used metrics in the state of the art of medical image segmentation based on neural networks. Additionally, we use Hausdorff Distance (HD) and the False Positive Rate (FPR) to provide a comparison of how well the contours were segmented and the false positives detected among the different methods investigated.

We compute the global DSC measure (simply referred to as *DSC*) for all segmented regions compared to the lesion ground truths, which indicates the overall segmentation accuracy. However, as MRI volumes may include secondary lesions (as in the original dataset) and other enhanced regions not regarded as a malignant lesion (vessel enhancements and other non-malignant findings such as axillary lymphadenopathy), the algorithm may segment those additional regions as lesions, which results in increasing the number of false positives detections, hence obtaining a lower global DSC. While this global DSC value is an important metric, we also propose to compute the DSC for the main lesion of the MRI, as an indication of the quality of the segmentation given an individual lesion. We will refer to this as the lesion DSC or $DSC_L$. Similarly, a lesion HD is also provided for the main lesion, referred to as $HD_L$ in addition to the global HD (referred to as *HD*).

### 5.1. Implementation details

The proposed architecture and all other experimented architectures were implemented in Python 3.7.4 using Pytorch 1.4.0 machine learning framework.

All python scripts were executed on Ubuntu on a physical server hosted in our university equipped with 256 GB RAM, and Nvidia GeForce RTX 2080 GPU with 11 GB of memory.

### 5.2. Experiment 1: Volume resampling

Since the original volumes have larger axial voxel size than the coronal-sagittal one, we investigate whether using isotropic volumes could improve the performance. In order to generate isotropic volumes, the original volumes were resampled having the voxel size across the three dimensions set to a specific target voxel size. Volumes with two

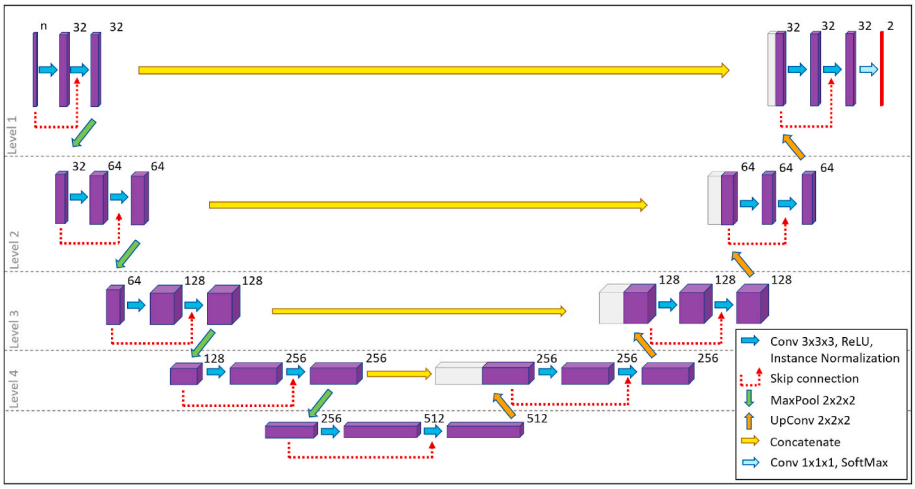


**Fig. 5.** The proposed U-Net with residual basic blocks architecture.

different target isotropic sizes were generated, in the first one (denoted by Isotropic 1) the target voxel size was set to be equal to the original coronal-sagittal voxel size; in other words the axial voxel size was decreased to match the original coronal-sagittal voxel size. In the second one (denoted by Isotropic 2) the target voxel size was set to be in between the original axial and the original coronal-sagittal voxel sizes; in other words the coronal-sagittal voxel size was increased and the axial voxel size was decreased.

Table 1 reports the obtained results, which indicate that using isotropic volumes did not improve the performance.

### 5.3. Experiment 2: input Volumes

As mentioned earlier in Section 2, one of the limitations in most of the existing works is that only one temporal acquisition among the time series is being used. Therefore, it is interesting to use several time acquisitions since the 3D + time data of DCE-MRI involves important information about the Time Intensity Curve (TIC) for each voxel, which if utilized, might improve the performance of lesion segmentation task using DCE-MRI.

In this subsection we evaluate the performance utilizing different input combinations comparing individual performance of these methods (input combinations) with the proposed ensemble method. Two ensemble models are obtained, in the first we perform a majority voting of the three models and in the second we perform the union of the three output segmentations provided by the three methods. Table 2 reports the obtained results.

As Table 2 indicates, the proposed method of using an ensemble of three models (each model utilizing a different input combination) outperforms the individual models. Although similar results of the global DSC were obtained for some individual models (3,4 and 5) and ensemble methods (Majority voting), the ensemble model based on the union of the three methods provides the best segmentation of main lesions indicated by the better values of $DSC_L$ and $HD_L$, with $DSC_L$ being differences significative. This can be qualitatively observed in Fig. 6.

The three methods used to generate an ensemble are the ones that yield the best performance among the five individual experimented methods (i.e. methods (3), (4) and (5)). The better performance of these three methods compared to methods (1) and (2) shows that incorporating information from more temporal volumes has the potential to improve the segmentation task from DCE-MRI.

Results show a better performance of the proposed method (ensemble by taking the union) in detecting very small lesions that are very difficult to segment with most other individual methods, an example case is shown in Fig. 6-(b). The better performance of the ensemble method is explained by the fact that it is using redundant information from multiple segmentations, hence minimising missing lesion voxels (FN) and increasing segmentation quality (higher DSC especially for main lesions), however at the expense of slightly higher FPR.

### 5.4. Experiment 3: U-Net architecture

In this subsection we compare results obtained using the U-Net architecture with residual blocks (described in Section 4.2) with another two U-Net based architectures. The first is a basic U-Net (used in our previous work [19]) with four levels (in order to have a fair comparison with our proposed architecture) along with a cross-entropy loss function. The second one is the two hierarchical basic U-Nets approach proposed in Ref. [33], in which we use dice-sensitivity-like loss in the first stage and dice-like loss in the second stage, as proposed by authors in Ref. [33]. In both stages the basic U-Net with four levels is used. Moreover, in the three experimented architectures we use the pre-contrast, last post-contrast and stdev volumes as inputs. Finally, the obtained results are reported in Table 3, where it should be noted that for a fair comparison of model's architectures, a single model U-Net with residual blocks is reported (not the ensemble results).

**Table 1**
DSC values (mean ± stdev) and p values for original (non-isotropic) volumes VS. isotropic volumes. *DSC* is the global dice and $DSC_L$ is the main lesion dice.

| Voxel Size | *DSC* | $DSC_L$ |
|---|---|---|
| Original (non-isotropic) | **0.649± 0.258** | **0.719 ± 0.250** |
| Isotropic 1 | 0.580 ± 0.274 (p = .008) | 0.693 ± 0.267 (p = .200) |
| Isotropic 2 | 0.579 ± 0.260 (p = .003) | 0.633 ± 0.277 (p<.001) |

**Table 2**
DSC (mean ± stdev), p, HD and FPR for different input combinations. *DSC* and *HD* are the global dice and Hausdorff distance whereas $DSC_L$ and $HD_L$ are the dice and Hausdorff distance of the main lesions.

| Inputs | *DSC* | $DSC_L$ | *HD* | $HD_L$ | FPR |
|---|---|---|---|---|---|
| (1) Pre, last post | 0.591 ± 0.279 (p = .001) | 0.652 ± 0.283 (p<.001) | 253.061 (p = .081) | 23.481 (p = .141) | 1.25E-04 (p = .001) |
| (2) Pre, last post, subtraction (post-pre) | 0.587 ± 0.260 (p<.001) | 0.649 ± 0.248 (p<.001) | 250.651 (p = .036) | 16.291 (p = .099) | 1.42E-04 (p = .010) |
| (3) 3 TP (pre, 2nd post, last post) | 0.654 ± 0.262 (p = .403) | 0.716 ± 0.241 (p<.001) | 219.712 (p<.001) | 11.996 (p = .453) | 1.07E-04 (p<.001) |
| (4) Full series (pre, all posts) | 0.664 ± 0.234 (p = .572) | 0.730 ± 0.222 (p = .005) | 227.654 (p<.001) | 11.879 (p = .135) | 1.14E-04 (p<.001) |
| (5) Pre, last post, stdev of the full series | 0.649 ± 0.258 (p = .220) | 0.719 ± 0.250 (p = .004) | 226.873 (p<.001) | 13.098 (p = .315) | 1.24E-04 (p<.001) |
| Ensemble (Majority voting) of (3), (4) and (5) | 0.679 ± 0.258 (p = .835) | 0.728 ± 0.243 (p = .002) | 135.415 (p<.001) | 11.561 (p = .593) | 1.00E-04 (p<.001) |
| Ensemble (Union) of (3), (4) and (5) | 0.674 ± 0.224 | **0.790 ± 0.172** | 278.548 | **11.256** | 1.73E-04 |

As observed from Table 3, the proposed U-Net architecture using a single model outperforms the other architectures achieving a mean dice of 0.649 (and 0.719 if to consider only main lesions).

The better performance of the proposed network is attributed to the presence of residual blocks. Residual blocks make it possible to train deeper networks and avoid overfitting due to the skip connections that allow the output of some earlier layers to be fed directly to deeper layers.

Fig. 7 shows some qualitative improvements in segmentation results using the proposed network compared to the other two.

### 5.5. Experiment 4: Detection evaluation and complemented GT

One of the main concerns in clinical settings regarding breast cancer is the detection performance. Since the proposed method performs both segmentation and detection in 3D, we investigate in this subsection how well it performs in detecting breast lesions.

Moreover, we investigate both segmentation and detection performance after complementing the GT. As mentioned earlier in section 3.2.1, obtained dices for cases with multiple lesions were affected by the incomplete GT issue since our network does not only segment the primary lesions. It is believed that a better and more fair evaluation (both for segmentation and detection) can be made with complete GT annotations. Therefore, the annotations of 11 cases with multiple lesions were complemented by an experienced radiologist. The proposed framework was then trained again using the same 46 cases but with only 11 annotations being replaced with the complemented ones. We would like to emphasize that complementing the GT does not mean just

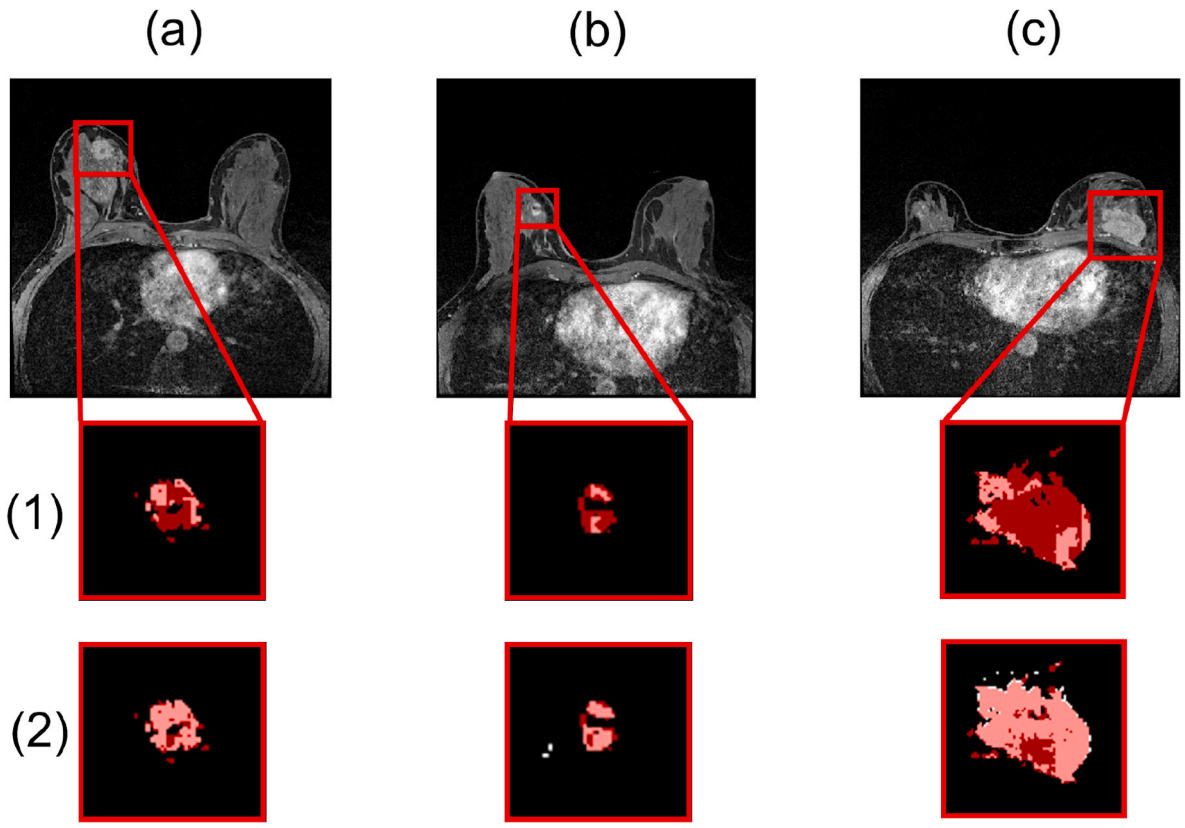


**Fig. 6.** Comparison of obtained segmentation using different methods (input combinations), where GT is represented in red and output segmentation in white. Methods are: (1) Full series (pre, all posts). (2) An ensemble of three methods (3 TP, full series, stdev along with pre and last post). Cases from (a) to (c) are A0DZ, A0HA, and A0E0, respectively. 2D slices shown on the top row are taken from the second post-contrast volumes of each case.

**Table 3**
DSC (mean ± stdev), p values, HD and FPR for different U-Net architectures. *DSC* and *HD* are the global dice and Hausdorff distance whereas $DSC_L$ and $HD_L$ are the dice and Hausdorff distance of the main lesions.

| Network | *DSC* | $DSC_L$ | *HD* | $HD_L$ | FPR |
|---|---|---|---|---|---|
| Single Model Basic U-Net | 0.551 ± 0.286 (p = .001) | 0.590 ± 0.302 (p<.001) | 123.939 (p<.001) | 16.369 (p = .001) | 1.21E-04 (p = .910) |
| Single Model Two hierarchical U-Nets | 0.615 ± 0.266 (p = .277) | 0.669 ± 0.275 (p = .137) | 212.123 (pp = .320) | 15.885 (p = .034) | 1.29E-04 (p = .729) |
| Single Model U-Net with residual blocks | **0.649 ± 0.258** | **0.719 ± 0.250** | 226.873 | **13.097** | 1.24E-04 |

refining the boundaries of the original GT lesions, but also segmenting new lesions that were not originally segmented in the original GT.

Segmentation and detection results before and after GT complementation are reported in Tables 4 and 5 respectively. In Table 4 we compare segmentation performance before and after GT complementation in terms of DSC, HD and FPR.

As expected, results improve, although only marginally, compared to the original GT both in terms of DSC and HD. Having only slight improvement after this step could be explained by the fact that only a small number of cases were complemented (11 cases), and the small size of most of the additional lesions (surrounding the main lesion) that were added after the GT complementation process.

In Table 5 we report the detection performance in terms of the number of correctly detected lesions and FP lesions based on Intersection (I), where two thresholds of I were used (0.20 and 0.50). In addition, the detected FP connected components were filtered to exclude those with size of 10 voxels or smaller. This is due to the majority of FP connected components consisting of only one voxel (or very few voxels). This size (10 voxels) was chosen to be smaller than the smallest lesion in the GT annotations. Finally we compare these detection related metrics before and after GT complementation.

Based on results from Table 5, the proposed method had a high detection rate for main lesions (i.e. before GT complementing). On the other hand, detection rate was decreased after GT complementation which is expected because of the complexity of the dataset and the lesions that were added. In addition, the detection rate after GT complementation was mainly affected by two cases (A0B1 and A0B5), which had complicated multiple lesions with many small lesions. Fig. 8 shows the original and complemented GT for case A0B1 in 3D, and the predicted segmentation with the network trained on both original and complemented GT.

This case has 46 lesions in the complemented GT, with 29 of them being detected by our method. Despite the missed lesions, dice value was largely improved, which is explained by the fact that the model segments the major lesions and misses those which were very small. For completeness, we investigate detection performance after excluding this case (results are reported in Table 5).

Although the number of FP is relatively high for a lesion detection algorithm in clinical practice (but less than 10 FP/volume) it is important to emphasize that the aim of the algorithm is to provide a fully automatic lesion segmentation framework to further characterise lesions and, as shown by the results, the overall segmentation even in complex cases such as in Fig. 8 can be considered as accurate.

Finally, even though this GT complementation process mildly impacted the segmentation results, we believe it has helped to better evaluate our method by adding lesions detected by our model but were

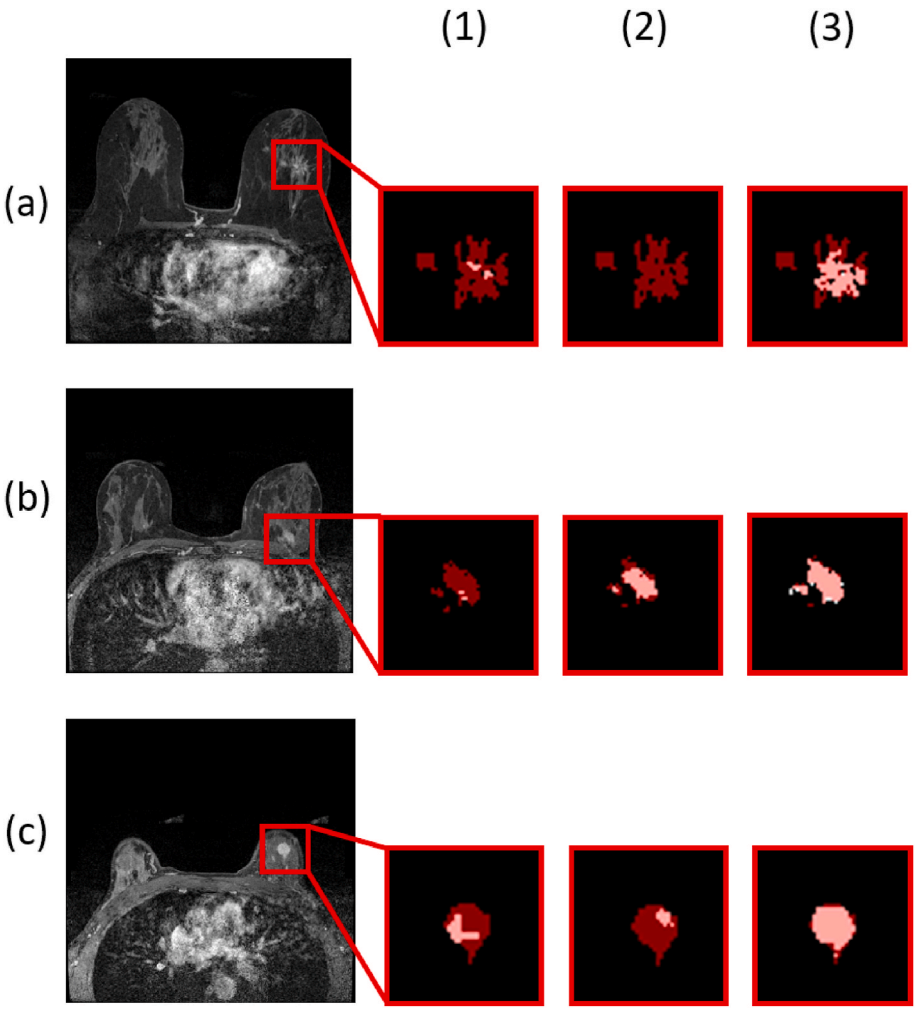


**Fig. 7.** Example cases of improved segmentation achieved using our proposed architecture compared to another two architectures, where GT is represented in red and output segmentation in white. Architectures are: (1) Basic U-Net. (2) Two hierarchical U-Nets. (3) Proposed architecture (single U-Net with residual blocks). Cases from (a) to (c) are: A0DE, A0C0, A0BQ, respectively. 2D slices shown on the leftmost are taken from the second post-contrast volumes of each case.

**Table 4**
DSC (mean ± stdev), p values, HD and FPR for the proposed ensemble method using data before and after complementing. *DSC* and *HD* are the global dice and Hausdorff distance whereas $DSC_L$ and $HD_L$ are the dice and Hausdorff distance of main lesions.

| Data | *DSC* | $DSC_L$ | *HD* | $HD_L$ | FPR |
|---|---|---|---|---|---|
| Before complementing | 0.674 ± 0.224 | 0.790 ± 0.172 | 278.548 | 11.256 | 1.73E-04 |
| After complementing | **0.680 ± 0.221** | **0.802 ± 0.156** | **275.013** | **11.042** | **1.71E-04** |

considered as FP due to their absence in the original GT. Moreover, the lesions that were added after GT complementing include very complicated types of lesions, which provides a more reliable dataset for the DL model to learn from and enhances its ability to learn from realistic cases. This highlights the importance of the quality of the annotations for such supervised learning methods. Therefore, we believe that this step is important for the completeness of the dataset for future use and reproducibility of the results.

### 5.6. Experiment 5: Comparison with non-learning methods

In this subsection we compare the performance of our proposed DL method with a non-learning method, that is Fuzzy C-Means [7]. We implement a standard FCM method with 5 clusters, fuzziness degree = 2, maximum number of iterations = 500 and convergence threshold of 1E-3. Tables 6 and 7 reports the obtained results (both segmentation and detection) with the FCM method.

**Table 5**
Detection performance of the proposed ensemble method using data before and after complementing, in terms of number of correctly detected lesions, number of FP lesions and mean Intersection (I). Filtering FP lesions means excluding small connected components of size 10 voxels or smaller.

| Data | Detected Lesions (I >= 0.2) | Detected Lesions (I >= 0.5) | FP lesions (before filtering) | FP lesions (after filtering) | Mean I |
|---|---|---|---|---|---|
| Before complementing | 97.8% (45/46 lesions) | 89.1% (41/46 lesions) | 1736 | 382 | 0.811 |
| After complementing (all 46 cases) | 78.8% (89/113 lesions) | 69% (78/113 lesions) | 1834 | 426 | 0.761 |
| After complementing (excluding case A0B1) | 89.6% (60/67 lesions) | 82.1% (55/67 lesions) | 1809 | 424 | 0.785 |

Results show that using Fuzzy C-Means directly on the whole DCE MRI volume, in the same conditions as in the proposed approach, obtains very unreliable segmentation results with DSC values being around

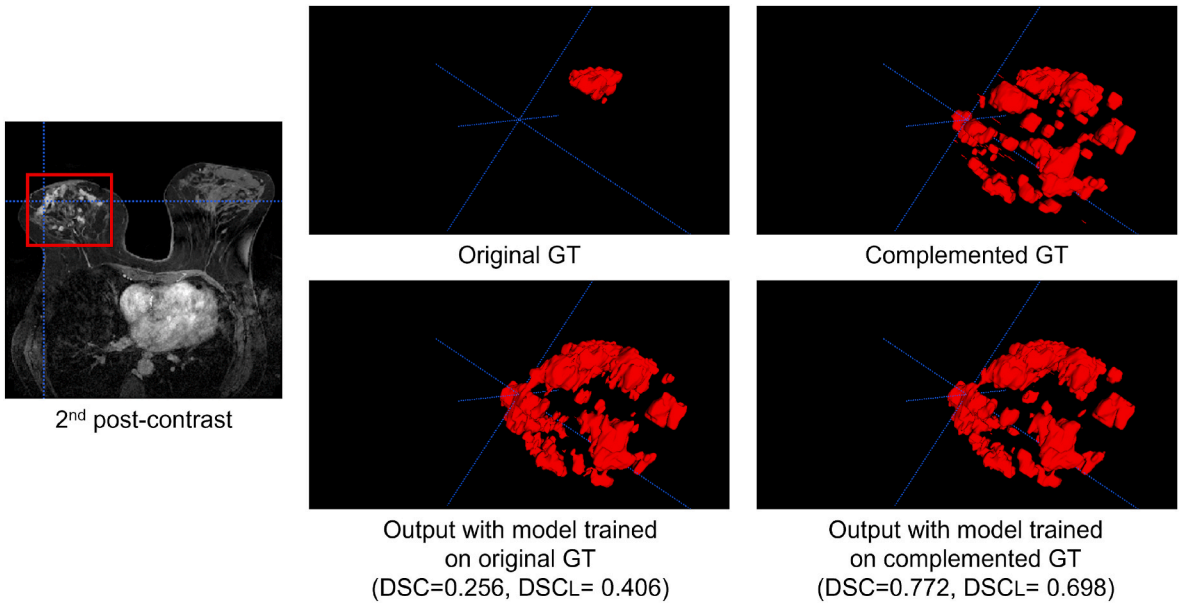


**Fig. 8.** Case (A0B1) with complicated multiple lesions. Shown 3D segmentations are: original and complemented GT, and the predicted segmentation with the network trained on both original and complemented GT.

**Table 6**
DSC (mean ± stdev), p values, HD and FPR for the FCM method using data before and after complementing. *DSC* and *HD* are the global dice and Hausdorff distance whereas $DSC_L$ and $HD_L$ are the dice and Hausdorff distance of main lesions.

| Data | *DSC* | $DSC_L$ | *HD* | $HD_L$ | FPR |
|---|---|---|---|---|---|
| Before complementing | 0.102 ± 0.092 | **0.560 ± 0.340** | 365.340 | 21.439 | 3.57E-03 |
| After complementing | **0.112 ± 0.105** | 0.549 1 ± 0.33 | **360.923** | **21.164** | **3.55E-03** |

**Table 7**
Detection performance of the FCM method using data before and after complementing, in terms of number of correctly detected lesions, number of FP lesions and mean Intersection (I). Filtering FP lesions means excluding small connected components of size 10 voxels or smaller.

| Data | Detected Lesions (I >= 0.2) | Detected Lesions (I >= 0.5) | FP lesions (before filtering) | FP lesions (after filtering) | Mean I |
|---|---|---|---|---|---|
| Before complementing | 97.8% (45/46 lesions) | 82.6% (38/46 lesions) | 113 234 | 12 520 | 0.730 |
| After complementing (all 46 cases) | 96.5% (109/113 lesions) | 78.8% (89/113 lesions) | 113 425 | 12 711 | 0.687 |
| After complementing (excluding case A0B1) | 94% (63/67 lesions) | 70.1% (47/67 lesions) | 112 195 | 12 547 | 0.675 |

0.1. This is explained by the fact that although it may correctly detect the main lesion, it also segments other mass-like regions which are not related to the lesion (i.e. breast boundaries, vessels, non-lesion structures). This is further corroborated by the $DSC_L$ measure (DSC of the main lesion), which improves significantly (0.549) compared to the global *DSC*, but is still lower compared to our proposal (0.802). Similarly, for the other measures (*HD*, $HD_L$ and FPR) the same trend is observed. Those results indicate that Fuzzy C-Means could not be applied on the whole volume and may require the manual selection of the region of interest containing the lesion, which is a clear limitation of the method. Even under these circumstances (i.e. for the $DSC_L$ metric) the proposed method still outperforms Fuzzy C-Means.

Regarding lesion detectability, values of correctly detected lesions are the same with both methods (97.8%) before GT complementing using a threshold of 0.2, with a better detection performance of the proposed algorithm (89.1% vs. 82.6%) when a higher threshold of Intersection (0.50) is used, which indicates better segmentation of the detected lesions with the proposed method. More importantly, the number of FP lesion detections increases significantly per volume (more than 25-fold) with the FCM method, which might explain why FCM detects more lesions after GT complementation, but with the proposed algorithm still achieving a better detection when using a higher threshold.

Finally, apart from the provided qualitative comparison, another aspect to be compared is that FCM is not a fully automatic method as the case of the proposed method. Besides the need for manual selection of lesion ROI for the FCM method to obtain reliable results, there is also a need for parameter tuning and post processing.

### 5.7. Discussion and limitations

The proposed method outperforms current state of the art in terms of DSC and HD measures on a particularly challenging but clinically realistic public dataset. In addition to the proposed algorithm, the analysis, metrics and description of the results and dataset serve as an interesting baseline framework for algorithm evaluation in the task of breast MRI lesion segmentation.

Besides segmentation, and since the proposed method actually performs both detection and segmentation at once, results show that in most cases lesions were detected successfully and their segmentation usually corresponds to the GT lesion area, hence, missing lesions was not a common issue. Furthermore, detected FPs were mostly corresponding to vascular structures or lymph nodes, which are both areas enhanced due to the contrast agent.

Comparing our results to other existing approaches (mentioned in Section 2.2), we show accurate segmentation results using a 3D segmentation approach with full-sized 4D data compared to most of the existing works in which either 2D slice based segmentation is performed, only a single temporal acquisition is used, or have the need to manually provide a region of interest within the lesion. Table 8 shows a

**Table 8**
Comparison of our proposed method and other existing methods.

| | Architecture | 2D/3D | Number of cases (public/private) | Inputs | Loss function | Evaluation criteria and score | Scanner |
|---|---|---|---|---|---|---|---|
| Zhang et al., 2019b [34] | U-Net | 2D | 1246 slices (private) | 2nd post-contrast (lesion bounding boxes) | Cross-entropy | DSC = 0.91 | – |
| | U-Net | 3D | 158 cases (private) | 2nd post-contrast (lesion bounding boxes) | Cross-entropy | DSC = 0.92 | |
| El Adoui et al., 2019 [15] | U-Net | 2D | 5452 slices (private) | Post-contrast | Cross-entropy | IoU = 0.761 4 | 1.5T Siemens |
| | SegNet | 2D | 5452 slices (private) | Post-contrast | Cross-entropy | IoU = 0.688 8 | |
| Piantadosi et al., 2019 [21] | U-Net | 2D | 35 case (256x128x80) (private) | Pre-contrast, 2 min post-contrast, and 6 min post-contrast | Dice | DSC = 0.612 4 | 1.5T Siemens |
| Zhang et al., 2019a [33] | Two hierarchical U-Nets | 3D | 272 cases (private) | Pre-contrast, post-contrast, and subtraction (breast mask guided) | First stage: Dice-sensitivity-like Second stage: Dice-like | DSC = 0.72 | 1.5T GE and 3.0T Siemens |
| Our proposed work | Ensemble of 3 U-Nets with ResNet basic blocks | 3D | 46 cases (public) | Ensemble (Union) of 3 models, each with different inputs (ROI mask guided) | Cross- entropy | DSC = 0.680 (0.802 for primary lesions only) | 1.5T GE |

comparison between the results obtained by our method and results obtained in other works.

As shown in Table 8, dice values reported in the existing literature were not particularly high (ranging between 0.60 and 0.80), reflecting the complexity of this task. One exemption could be the work of Zhang et al., 2019b [34] where high dice values were obtained, which are likely to be explained by the fact that lesion bounding boxes are used as input. Our results of 0.68 were very close to the results obtained in other works, despite the complexity of the dataset with multiple lesions and the fact that ours is a fully automatic proposal using 3D data. It is important to mention that a direct comparison of our method with the existing approaches can not be established as all methods have been using different datasets and input information (3D/2D, whole images or ROIs). In that sense, as we are using a publicly available dataset, the results presented in this work allows reproducibility and comparability of future developed methods in order to mitigate the variability of results in the current state of the art.

As for the limitations of our study, the dataset we used had a small number of cases. Having a larger dataset is believed to improve the performance. Additionally, all cases in our dataset contained at least one lesion which might cause issues when dealing with normal MRIs (i.e. healthy cases). Although MRI studies are usually acquired for high risk women or in cases of suspicious findings (with a higher incidence than screening population), our work assumes that at least a lesion is present in the scan, which may not always be the case. Another limitation related to our dataset is that we did not test with scans acquired from different scanners. All scans used in this study were acquired with the same scanner. Future work will investigate the use of different scanners and how this affects the segmentation accuracy.

In addition, even though the breast ROI obtained and used in our work could eliminate most of the confounding organs region in most cases, there were few cases where the ROI could also include surrounding confounding organs. This is due to the fact that some lesions are located on the body-breast boundary and the ROI is defined in a conservative way to avoid missing any breast lesions, at the expense of including some small surrounding areas such as heart and lungs. This could be considered as a limitation to this work, future work will focus on incorporating breast masks instead of breast ROIs.

Another observation is that our network also segments axillary lymphadenopathy (which appears in several cases among our dataset), even though it is not strictly a malignant lesion, it is often a sign associated with breast cancer. Axillary lymphadenopathy is defined as changes in the size and consistency of lymph nodes in the armpit (axilla) and it is a symptom associated with a range of diseases and conditions from mild infections to breast cancer, including also the COVID-19 vaccination as shown in recent studies [14]. Fig. 9 shows an example case diagnosed with uni-centric lesion and an axillary lymphadenopathy. As shown in Fig. 9, the lesion is segmented well by our algorithm, however the dice is affected due to segmenting the axillary lymphadenopathy which has a bigger size than the lesion.

## 6. Conclusions and future work

In this study an automated breast lesion segmentation method has been proposed for DCE-MRI. Our proposed method is an ensemble of models based on a 3D patch based modified U-Net framework. In this modified U-Net we have introduced residual basic blocks instead of basic U-Net blocks. Additionally, we have utilized a ROI restricted balanced patch extraction in order to address both the class imbalance and confounding organs problems. Differently from most existing works on this topic, full automatic 3D segmentation is performed instead of 2D. Therefore, our method performs both segmentation and detection at the same time.

Additionally we have utilized not only one temporal acquisition (as

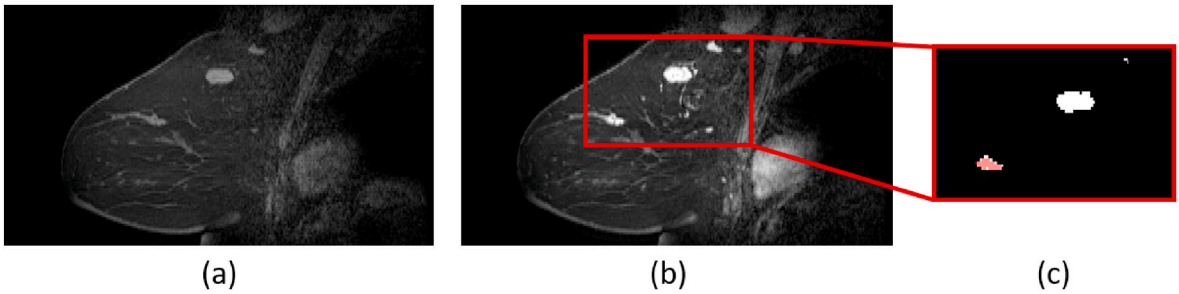


**Fig. 9.** Example case (A18H) with axillary lymphadenopathy. (a) pre-contrast. (b) second post-contrast. (c) Obtained segmentation (in white) and GT (in red).

in most existing works) but different temporal scans instead. Different combinations of inputs have been investigated and a combined model of the best three combinations have been proposed.

Experiments have been performed on 46 cases and different metrics have been used to evaluate the obtained segmentation. We have obtained a mean dice of 0.680 (0.802 for main lesions only) which is promising considering the various issues encountered with the incomplete GT and the complex dataset that included very small, irregular, low enhanced lesions as well as lesions located on the body-breast boundary and confounding background.

Further improvements could be achieved by incorporating a larger dataset with a complete annotation for those cases with multiple lesions. Moreover, using a breast mask instead of a simple ROI could also potentially alleviate the issue of confounding regions as it will help exclude confounding regions of the internal organs without excluding lesions located on the body-breast boundary.

Finally, the deployment of deeper architectures and the deployment of an alternative way to represent the 4D data in one volume as a reduced representation that better captures the TIC of each voxel could also improve the results by reducing computational demands.

## Credit authorship contribution statement

**Roa'a Khaled:** Data processing, data analysis and interpretation, manuscript drafting and final approval. **Joel Vidal:** Data processing and analysis, manuscript revising and final approval. **Joan C Vilanova:** Data annotation and analysis, manuscript revising and final approval. **Robert Martí:** Conception and study design, data analysis and interpretation, manuscript revising and final approval.

## Declaration of competing interest

I would like to submit the original research manuscript entitled "*A U-Net Ensemble for Breast Lesion Segmentation in DCE MRI*" to be considered for publication in the Journal of Computers in Biology and Medicine.

Throughout this study, we propose an automated method for breast lesion segmentation from DCE-MRI based on Deep Learning (DL). The proposed method extends an earlier version of this work based on a U-Net framework. The contributions of this work are the proposal of a modified U-Net architecture that incorporates residual basic blocks and the analysis of the input DCE information. In that sense, we propose the use of an ensemble method that combines three U-Net models, each using a different input combination, outperforming all individual methods and other existing approaches. For evaluation, we use 46 cases from a subset of the TCGA-BRCA dataset, which is a challenging and publicly available dataset not reported to date for MRI lesion segmentation. Due to the incomplete annotations (ground truth) provided, we complement them with the help of a radiologist in order to include secondary lesions that were not originally segmented. The proposed ensemble method outperforms state-of-the art methods using the same dataset, demonstrating the effectiveness of our method considering the complexity of the dataset. In addition to a novel segmentation method, this work provides a framework for method comparison in order to become a publicly available benchmark for this task.

We declare that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere. We also declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors have approved the manuscript and agree with its submission to the Journal of Computers in Biology and Medicine. Thank you very much for your kind consideration.

## Acknowledgments

This work was partially supported by the project ICEBERG: Image Computing for Enhancing Breast Cancer Radiomics (RTI2018-096 333-B-I00, Spanish Ministry). We would like to thank the TCGA Breast Phenotype Research Group for providing the computer-extracted lesion segmentation data used in this study, which comes from the University of Chicago lab of Maryellen Giger.

## References

[1] A.A. Alzaghal, P.J. DiPiro, Applications of advanced breast imaging modalities, Curr. Oncol. Rep. 20 (2018) 57, https://doi.org/10.1007/s11912-018-0700-3.

[2] A.B. Ashraf, S.C. Gavenonis, D. Daye, C. Mies, M.A. Rosen, D. Kontos, A multichannel markov random field framework for tumor segmentation with an application to classification of gene expression-based breast cancer recurrence risk, IEEE Trans. Med. Imag. 32 (2013) 637–648, https://doi.org/10.1109/TMI.2012.2219589.

[3] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495, https://doi.org/10.1109/TPAMI.2016.2644615.

[4] A. Bria, N. Karssemeijer, F. Tortorella, Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications, Med. Image Anal. 18 (2013) 241–252, https://doi.org/10.1016/j.media.2013.10.014.

[5] E. Burnside, K. Drukker, H. Li, E. Bonaccio, M. Zuley, M. Ganott, J. Net, E. Sutton, K. Brandt, G. Whitman, S. Conzen, L. Lan, Y. Ji, Y. Zhu, C. Jaffe, E. Huang, J. Freymann, J. Kirby, E. Morris, M. Giger, Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage, Cancer 122 (2015) 748–757, https://doi.org/10.1002/cncr.29791.

[6] M. Chen, H. Zheng, C. Lu, E. Tu, J. Yang, N. Kasabov, A spatio-temporal fully convolutional network for breast lesion segmentation in DCE-MRI, in: L. Cheng, A. C.S. Leung, S. Ozawa (Eds.), Neural Information Processing, Springer International Publishing, Cham, 2018, pp. 358–368, https://doi.org/10.1007/978-3-030-04239-4_32.

[7] W. Chen, M. Giger, U. Bick, A fuzzy C-means (FCM)-Based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images, Acad. Radiol. 13 (2006) 63–72, https://doi.org/10.1016/j.acra.2005.08.035.

[8] L. Cheng, X. Li, Breast magnetic resonance imaging: kinetic Curve assessment, Gland Surg. 2 (2013) 50–53, https://doi.org/10.3978/j.issn.2227-684X.2013.02.04.

[9] P. Christ, F. Ettlinger, F. Grün, M. Elshaera, J. Lipkova, S. Schlecht, F. Ahmaddy, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, F. Hofmann, M. D'Anastasi, S. A. Ahmadi, G. Kaissis, J. Holch, W. Sommer, R. Braren, V. Heinemann, B. Menze, Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks, CoRR abs/1702.05970. URL: http://arxiv.org/abs/1702.05970, 2017. arXiv:1702.05970.

[10] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-net: learning dense volumetric segmentation from sparse annotation, in: S. Ourselin, L. Joskowicz, M.R. Sabuncu, G. Unal, W. Wells (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, Cham, 2016, pp. 424–432, https://doi.org/10.1007/978-3-319-46723-8_49.

[11] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The cancer imaging archive (TCIA): maintaining and operating a public information repository, J. Digit. Imag. 26 (2013) 1045–1057, https://doi.org/10.1007/s10278-013-9622-7.

[12] A. Clèrigues, S. Valverde, J. Bernal, J. Freixenet, A. Oliver, X. Lladó, Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks, Comput. Biol. Med. 115 (2019) 103487, https://doi.org/10.1016/j.compbiomed.2019.103487.

[13] H. Degani, V. Gusis, D. Weinstein, S. Fields, S. Strano, Mapping pathophysiological features of breast tumors by MRI at high spatial resolution, Nat. Med. 3 (1997) 780–782, https://doi.org/10.1038/NM0797-780.

[14] C.E. Edmonds, S.P. Zuckerman, E.F. Conant, Management of unilateral axillary lymphadenopathy detected on breast MRI in the era of coronavirus disease (COVID-19) vaccination, Am. J. Roentgenol. (2021), https://doi.org/10.2214/AJR.21.25604.

[15] M. El Adoui, S. Mahmoudi, A. Larhmam, M. Benjelloun, MRI breast tumor segmentation using different encoder and decoder CNN architectures, J. Comput. 8 (2019) 52, https://doi.org/10.3390/computers8030052.

[16] A. Gubern-Mérida, R. Martí, J. Melendez, J. Hauth, R. Mann, N. Karssemeijer, B. Platel, Automated localization of breast cancer in DCE-MRI, Med. Image Anal. 20 (2014) 265–274, https://doi.org/10.1016/j.media.2014.12.001.

[17] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M. Valdes-Hernandez, D. Dickie, J. Wardlaw, D. Rueckert, White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks, Neuroimage: Clinical 17 (2017), https://doi.org/10.1016/j.nicl.2017.12.022.

[18] L. Hu, Z. Cheng, M. Wang, Z. Song, Image manifold revealing for breast lesion segmentation in DCE-MRI, Bio Med. Mater. Eng. 26 (2015) S1353–S1360, https://doi.org/10.3233/BME-151433.

[19] R. Khaled, J. Vidal, R. Martí, Deep learning based segmentation of breast lesions in DCE-MRI, in: A. Del Bimbo, R. Cucchiara, S. Sclaroff, G.M. Farinella, T. Mei, M. Bertini, H.J. Escalante, R. Vezzani (Eds.), Pattern Recognition. ICPR International Workshops and Challenges, Springer International Publishing, Cham, 2021, pp. 417–430, https://doi.org/10.1007/978-3-030-68763-2_32.

[20] L. Losurdo, T. Basile, A. Fanizzi, R. Bellotti, U. Bottigli, R. Carbonara, R. Dentamaro, D. Diacono, V. Didonna, A. Lombardi, F. Giotta, C. Guaragnella, A. Mangia, R. Massafra, P. Tamborra, S. Tangaro, D. La Forgia, A gradient-based approach for breast DCE-MRI analysis, BioMed Res. Int. 2018 (2018) 9032408, https://doi.org/10.1155/2018/9032408.

[21] G. Piantadosi, S. Marrone, A. Galli, M. Sansone, C. Sansone, DCE-MRI breast lesions segmentation with a 3TP U-net deep convolutional neural network, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), 2019, pp. 628–633, https://doi.org/10.1109/CBMS.2019.00130.

[22] A. Rampun, B.W. Scotney, P.J. Morrow, H. Wang, J. Winder, Segmentation of breast MR images using a generalised 2D mathematical model with inflation and deflation forces of active contours, in: Artificial Intelligence in Medicine, vol. 97, 2019, pp. 44–60, https://doi.org/10.1016/j.artmed.2018.10.007. URL: https://www.sciencedirect.com/science/article/pii/S0933365718302549.

[23] B. Reig, L. Heacock, K.J. Geras, L. Moy, Machine learning in breast MRI, J. Magn. Reson. Imag. 52 (2020) 998–1018, https://doi.org/10.1002/jmri.26852. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.26852. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.26852.

[24] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[25] M.K. Sharma, M. Jas, V. Karale, A. Sadhu, S. Mukhopadhyay, Mammogram segmentation using multi-atlas deformable registration, Comput. Biol. Med. 110 (2019) 244–253. URL: https://www.sciencedirect.com/science/article/pii/S0010482519301945.

[26] X. Shi, Z. Chen, H. Wang, D.Y. Yeung, W.k. Wong, W.c. Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, in: Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 1, MIT Press, Cambridge, MA, USA, 2015, pp. 802–810, https://doi.org/10.5555/2969239.2969329.

[27] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2020, CA A Cancer J. Clin. 70 (2020) 7–30, https://doi.org/10.3322/caac.21590.

[28] V.S. Subbhuraam, E. Ng, U.R. Acharya, O. Faust, Breast imaging: a survey, World J. Clin. Oncol. 2 (2011) 171–178, https://doi.org/10.5306/wjco.v2.i4.171.

[29] A. Vignati, V. Giannini, M. Luca, L. Morra, D. Persano, L. Carbonaro, I. Bertotto, L. Martincich, D. Regge, A. Bert, F. Sardanelli, Performance of a fully automatic lesion detection system for breast DCE-MRI, J. Magn. Reson. Imag. : JMRI 34 (2011) 1341–1351, https://doi.org/10.1002/jmri.22680.

[30] W.D. Vogl, K. Pinker, T. Helbich, H. Bickel, G. Grabner, W. Bogner, S. Gruber, Z. Bago-Horvath, P. Dubsky, G. Langs, Automatic segmentation and classification of breast lesions through identification of informative multiparametric PET/MRI features, European Radiology Experimental 3 (2019), https://doi.org/10.1186/s41747-019-0096-3.

[31] D. Wei, S. Weinstein, M.K. Hsieh, L. Pantalone, D. Kontos, Three-dimensional whole breast segmentation in sagittal and axial breast MRI with dense depth field modeling and localized self-adaptation for chest-wall line detection, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 66 (2019) 1567–1579, https://doi.org/10.1109/TBME.2018.2875955.

[32] J. Zhang, Y. Gao, S.H. Park, X. Zong, W. Lin, D. Shen, Structured learning for 3-D perivascular space segmentation using vascular features, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 64 (2017) 2803–2812.

[33] J. Zhang, A. Saha, Z. Zhu, M.A. Mazurowski, Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics, IEEE Trans. Med. Imag. 38 (2019) 435–447, https://doi.org/10.1109/TMI.2018.2865671.

[34] L. Zhang, Z. Luo, R. Chai, D. Arefan, J. Sumkin, S. Wu, Deep-learning method for tumor segmentation in breast DCE-MRI, in: P.H. Chen, P.R. Bak (Eds.), Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications, International Society for Optics and Photonics, SPIE, 2019, pp. 97–102, https://doi.org/10.1117/12.2513090.

[35] Y. Zheng, S. Baloch, S. Englander, M.D. Schnall, D. Shen, Segmentation and classification of breast tumor using dynamic contrast-enhanced MR images, in: N. Ayache, S. Ourselin, A. Maeder (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 393–401, https://doi.org/10.1007/978-3-540-75759-7_48.