Pre and Post-hoc Diagnosis and Interpretation of Malignancy from Breast DCE-MRI

Gabriel Maicas[a,1,*], Andrew P. Bradley[b,1], Jacinto C. Nascimento[c,1], Ian Reid[a,2], Gustavo Carneiro[a,1]

[a] *Australian Institute for Machine Learning, The University of Adelaide, Australia*
[b] *Science and Engineering Faculty, Queensland University of Technology, Australia*
[c] *Institute for Systems and Robotics, Instituto Superior Tecnico, Portugal*

Abstract

We propose a new method for breast cancer screening from DCE-MRI based on a post-hoc approach that is trained using weakly annotated data (i.e., labels are available only at the image level without any lesion delineation). Our proposed post-hoc method automatically diagnosis the whole volume and, for positive cases, it localizes the malignant lesions that led to such diagnosis. Conversely, traditional approaches follow a pre-hoc approach that initially localises suspicious areas that are subsequently classified to establish the breast malignancy – this approach is trained using strongly annotated data (i.e., it needs a delineation and classification of all lesions in an image). We also aim to establish the advantages and disadvantages of both approaches when applied to breast screening from DCE-MRI. Relying on experiments on a breast DCE-MRI dataset that contains scans of 117 patients, our results show that the post-hoc method is more accurate for diagnosing the whole volume per patient, achieving an AUC of 0.91, while the pre-hoc method achieves an AUC of 0.81. However, the performance for localising the malignant lesions remains challenging for the post-hoc method due to the weakly labelled dataset employed during training.

*Corresponding author

Email address: gabriel.maicas@adelaide.edu.au (Gabriel Maicas)

[1]This work was partially supported by the Australian Research Council project (DP180103232). We would like to thank Nvidia for the donation of a TitanXp that supported this work.

[2]IR acknowledges the Australian Research Council: ARC Centre for Robotic Vision (CE140100016) and Laureate Fellowship (FL130100102)

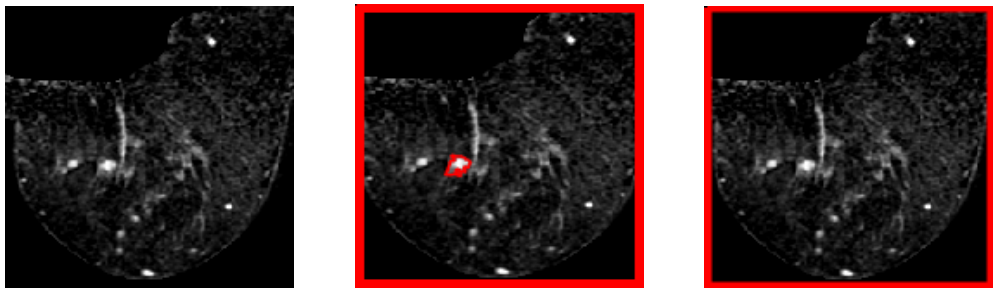
Keywords: magnetic resonance imaging, breast screening, meta-learning, few-shot learning, weakly supervised learning, strongly supervised learning, model interpretation, lesion detection, deep reinforcement learning.

1. Introduction

Breast cancer is amongst the most diagnosed cancers (AIHW, 2007; Siegel et al., 2017) affecting women worldwide (DeSantis et al., 2015; Torre et al., 2015). One of the most effective ways of increasing the survival rate for this disease is based on early detection (Saadatmand et al., 2015; Welch et al., 2016). Screening programs aim to provide such early detection by diagnosing at-risk, asymptomatic patients, allowing for an early intervention and treatment. The most widely employed image modality for population-based breast screening is mammography. High risk patients are also recommended to undergo screening with dynamically contrast enhanced magnetic resonance imaging (DCE-MRI) (Mainiero et al., 2017; Smith et al., 2017). DCE-MRI is known to increase the sensitivity, compared to mammography, especially in young patients that have denser breasts (Kriege et al., 2004).

However, the diagnosis and interpretation of DCE-MRI is a challenging and time consuming task that involves the interpretation of large amounts of data (Behrens et al., 2007) and is prone to high inter-observer variability (Grimm et al., 2015; Lehman et al., 2013). Computer-aided diagnosis (CAD) systems are designed to reduce the analysis time (Gubern-Mérida et al., 2016; Wood, 2005), increase sensitivity (Vreemann et al., 2018) and specificity (Meinel et al., 2007), and serve as a second (automated) reader (Shimauchi et al., 2011). Designing such systems is challenging due to the variability in location, appearance (Levman et al., 2009), size and shape (Song et al., 2016), and the low signal-to-noise ratio (Kousi et al., 2015) of lesions. In general, such CAD systems can be categorised as pre-hoc or post-hoc, depending on how the processing stages are organised, as explained below.

Fully automated pre-hoc CAD methods for breast screening (Amit et al., 2017b; Dalmış et al., 2018; Gubern-Mérida et al., 2015) from DCE-MRI compute the confidence score of malignancy of a breast using the following two-stage sequential approach: 1) detection of suspicious lesions, and 2) classification of the detected lesions. During detection (i.e., the first stage), the algorithm localises benign and malignant lesions, and possibly false positive detections, in the image, which are then classified as malignant or non-malignant in the second stage. Four important challenges arise with this



(a) (b) (c)

Figure 1: Example of a DCE-MRI breast image and annotation types. Image (1a) shows a slice of a breast DCE-MRI volume. Image (1b) shows the same slice with the strong annotations: lesion delineation classification as malignant. Image (1c) shows the weak annotation (i.e., whole image) of the same breast volume as malignant.

pre-hoc approach. Firstly, the modelling of the detector requires strong labels, i.e., precise voxel-wise annotation of lesions (see Fig. 1 for an example of different types of annotations). Strong annotation is expensive because it requires experts to label a relatively large number of training volumes; in addition, given the difficulties involved in such manual labelling process, this annotation may contain noise (this happens partly because experts are generally not trained to provide such precise annotations in regular practice). Secondly, the classifier may be trained using incorrect manually annotated lesion class labels. Such manual annotation is usually produced by biopsy analysis, but if there are benign and malignant lesions jointly present in the same breast, this analysis may not determine the correct association. Thirdly, apart from rare exceptions that need large annotated training sets (Ribli et al., 2018), pre-hoc diagnosis systems are generally trained in a two-stage process (Gubern-Mérida et al., 2015; Mcclymont, 2015). This pipeline is not the optimal way to maximise classification diagnosis performance because the final classification depends on the detection, but the detection optimality does not warrant classification optimality. Finally, the fourth challenge is that the classification accuracy is limited by the detector performance, where it is impossible for the classifier to recover from a missing lesion detection because it can not be classified.

An alternative approach that is starting to gain traction (Esteva et al., 2017; Maicas et al., 2018; Wang et al., 2017a) reverses these stages. The first stage aims to classify the whole breast scan directly, followed by a second

stage that localizes regions in the scan that can explain the classification – for instance, if the first stage outputs a malignant diagnosis, then the second stage aims to find malignant lesions in the scan. We term this a *post-hoc* approach. This approach is of special interest for the problem of breast screening from DCE-MRI because the whole-scan diagnosis can, for example, analyse regions other than lesions that may contain relevant information for the diagnosis (Kostopoulos et al., 2017). The main advantage of these systems compared to pre-hoc systems is the possibility of using scan-level labels (referred to as weak labels in the rest of the paper). Such labels are already present in many Picture Archiving and Communication Systems (PACS) or can be automatically extracted from radiology reports (Wang et al., 2017a), eliminating almost completely the effort needed for the manual annotation described above for the pre-hoc approach. Also, the use of scan-level labels overcomes the limitations in annotations required by pre-hoc approaches. Firstly, there is no need for lesion delineation avoiding such costly process. Secondly, the incorrect labelling of lesions explained above is reduced as the most likely lesion to be malignant is biopsied and therefore the label is more likely to be correct –there is no need to associate labels with lesions. The main challenge of post-hoc systems resides in highlighting the scan regions that can justify a particular classification (e.g., in the case of a malignant classification, it is expected that the regions represent the malignant areas of the scan), given that such manual annotation is not available. This challenge is important for the deployment of post-hoc systems in clinical practice (Caruana et al., 2015).

In this paper, we propose a new post-hoc method and a systematic comparison between pre-hoc and post-hoc approaches for breast screening from DCE-MRI. We aim to answer the following research questions: 1) which approach should be chosen if the goal is to optimally classify a whole scan in terms of malignant or non-malignant findings, and 2) how accurate is the localisation of malignant lesions produced by post-hoc approaches when compared with the localisation of malignant lesions produced by pre-hoc methods. The pre-hoc system considered in this paper is based on our recent detection model (Maicas et al., 2017b) that achieves state-of-the-art (SOTA) lesion localisation, while reducing the inference time needed by traditional exhaustive search methods. For the post-hoc system, we rely on our recently proposed approach based on meta-learning (Maicas et al., 2018) that holds the SOTA performance for the problem of breast screening from DCE-MRI. Decision interpretation is based on our recent 1-class saliency

detector (Maicas et al., 2019), especially designed for the weakly supervised lesion localisation problem after performing volume diagnosis. See Fig. 2 for an overview of the pre-hoc and post-hoc pipelines.

Experiments on a breast DCE-MRI dataset containing 117 patients and 141 lesions show that the post-hoc system achieves better malignancy classification accuracy than the pre-hoc method. In terms of lesion localisation, the post-hoc approach shows less accurate performance compared to the pre-hoc system, which we infer that is mostly due to the weak annotation used in the training phase of the post-hoc method.

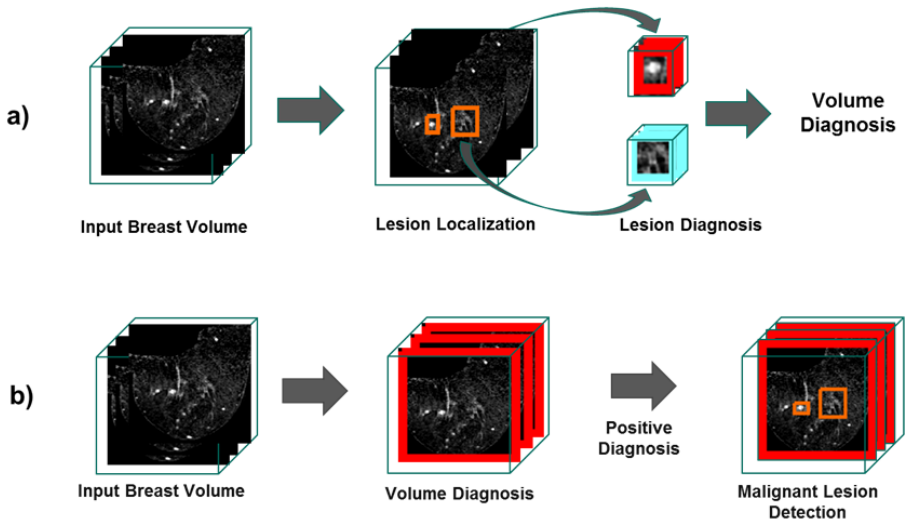


Figure 2: Pre-hoc and post-hoc approaches for breast screening. **a)** The pre-hoc approach first localises lesions in the input breast volume (e.g., detections in orange), and then these lesions are classified to decide about their malignancy (e.g., red indicates positive and blue means negative malignancy classification). Finally, the breast volume is diagnosed according to the classification scores of the lesions. **b)** The post-hoc approach first diagnoses the input breast volume (e.g., red means positive malignancy classification). If the diagnosis is positive, then malignant lesions are localised in the breast (e.g., detections in orange).

2. Literature Review

2.1. Pre-hoc Approaches

Pre-hoc approaches are assumed to contain two sequential stages: 1) detection of regions of interest (ROI) containing suspicious tissue, and 2)

classification of ROIs into malignant or not malignant (benign and/or false positive) tissue.

Traditional pre-hoc approaches for breast screening from breast DCE-MRI were based on manual (Agner et al., 2014; Gallego-Ortiz and Martel, 2015; Mus et al., 2017; Soares et al., 2013) or semi-automated (Chen et al., 2006; Dalmış et al., 2016; Meinel et al., 2007; Milenković et al., 2017; Platel et al., 2014) ROI detection. In addition, the classification in these traditional approaches was based on support vector machine (SVM), random forest, or artificial neural network models, using hand-designed features (e.g., dynamic, morphological, textural or multifractal) (Dalmış et al., 2016; Meinel et al., 2007; Milenković et al., 2017; Platel et al., 2014).

Aiming at reducing user intervention to reduce the number of ROIs (Liu et al., 2017), pre-hoc systems evolved to be fully automated. Such automated pre-hoc approaches generally employed an exhaustive search method or clustering to detect ROIs in the scan using hand-designed features (Gubern-Mérida et al., 2015; Mcclymont, 2015; Renz et al., 2012; Wang et al., 2014). The classification of ROIs into false positive, benign or malignant findings is then performed with a new set of hand-designed features extracted from the ROIs (Gubern-Mérida et al., 2015; Mcclymont, 2015; Renz et al., 2012; Wang et al., 2014). These fully automated methods generally suffer from two issues: 1) the sub-optimality of hand-designed features needed at both ROI localization and ROI classification, and 2) the high computational cost of the exhaustive search to detect ROIs.

Both limitations have been addressed after the introduction of deep learning methodologies (Krizhevsky et al., 2012) in the field of medical image analysis. Initially, feature sub-optimality was addressed either for ROI detection (Maicas et al., 2017a,b) or classification (Amit et al., 2017a,b; Rasti et al., 2017), but it was recently solved for both detection and classification (Dalmış et al., 2018). Dalmış et al. (2018) also reduced the inference time of the exhaustive search by directly computing a segmentation map from the scan using a U-net (Ronneberger et al., 2015).

Although each step of the pipeline has been individually optimized, there is no guarantee that the full pipeline is optimal in terms of classification accuracy. This was addressed with the formation of large datasets that has enabled the use of SOTA one-stage detection and classification computer vision techniques, such as Faster R-CNN (Ren et al., 2015) or Mask RCNN (He et al., 2017). The main advantage of these methods lies in the optimality of the end-to-end training, effectively merging the detection and classification

tasks (Dalmış et al., 2018). For example, Ribli et al. (2018) applied Faster R-CNN to detect tumours from mammograms and they showed that this approach is quite efficient in terms of inference time. However, Faster R-CNN generalises poorly, which means that the training set must contain a large annotated set of ROIs and, at the same time, be rich enough to comprise all possible lesion variations. Besides the need for large datasets, which are difficult to acquire for DCE-MRI breast screening, these systems suffer from the need for strong annotations (i.e., the accurate delineation of the lesions). Li et al. (2018) partially addressed this issue by developing a semi-supervised system, alleviating the need of lesion annotations. However, a large number of annotated images (880) is still required to train the system.

2.2. Post-hoc Approaches

Post-hoc systems aim to overcome the need for strong annotations by training models with only scan-level labels (i.e., weak labels). This is especially useful for the problem of breast screening, where the analysis of adjacent regions to lesions may be important (Kostopoulos et al., 2017). In addition, the classification accuracy of post-hoc systems are not constrained by the lesion detection, which is the case in pre-hoc systems.

Several post-hoc systems have been proposed (Wang et al., 2017a; Zhu et al., 2017). For instance, Wang et al. (2017a) use a deep learning model to produce classification scores from whole scans and Zhu et al. (2017) propose a deep multiple instance learning. However, these approaches still require large datasets to achieve good performance. This issue was addressed by Maicas et al. (2018), who proposed a new meta-learning methodology to learn from a small number of annotated training images. Their work established a new SOTA classification accuracy for breast screening from DCE-MRI.

The main challenge for post-hoc models arises from the fact that they do not use manually annotated ROIs for training, which makes the ROI localisation (and delineation) a hard task. Such ROI localisation is important for explaining the classification made by the CAD system in clinical settings (e.g., for a scan classified as malignant, doctors are likely to know where the lesions are located). Solving this lesion localisation problem is a research problem that is being actively investigated in the field (Dubost et al., 2017; Feng et al., 2017; Maicas et al., 2019; Wang et al., 2017b; Yang et al., 2017). The approach proposed by Maicas et al. (2019) achieves SOTA detection performance by properly defining saliency for the problem of weakly supervised

lesion localisation, which assures that salient regions represent malignant lesions in the image.

However, the literature does not provide any studies comparing pre and post-hoc diagnosis approaches. The main reason for this absence of comparison among the methods described in this literature review is that such analysis is not straightforward due to (Maicas et al., 2017b): 1) the lack of publicly available datasets that can be used to compare new approaches to the current state-of-the-art, 2) the criteria to decide if an ROI is a true positive detection, and 3) the criteria to decide if lesions labelled as the challenging BIRADS=3 should be included into the benign category (Gubern-Mérida et al., 2015). In addition, not all assessments of pre-hoc fully automated methodologies consider false positives in the diagnostic stage as they only differentiate between benign and malignant (Mcclymont, 2015). We propose to compare both types of automated approaches for the problem of breast screening from breast DCE-MRI. With the use of a common dataset and well-defined criteria to satisfy the issues described above, we investigate which approach performs better for breast diagnosis and lesion localisation.

3. Methods

This section provides a formal description of the dataset in Sec. 3.1, the pre-hoc method in Sec. 3.2, and the post-hoc approach in Sec. 3.3.

3.1. Dataset

Let $\mathcal{D} = \left\{ \left(\mathbf{b}_i, \mathbf{x}_i, \mathbf{t}_i, \{\mathbf{s}_i^{(j)}\}_{j=1}^{M}, \{\mathbf{l}_i^{(j)}\}_{j=1}^{M}, \mathbf{y}_i \right)_i \right\}_{i \in \{1,\dots,|\mathcal{D}|\}, \mathbf{b}_i \in \{\text{left,right}\}}$ denote the 3D DCE-MRI dataset, where $\mathbf{b}_i \in \{\text{left}, \text{right}\}$ specifies the left or right breast of the i^{th} patient; $\mathbf{x}_i, \mathbf{t}_i : \Omega \rightarrow \mathbb{R}$ represent the first 3D DCE-MRI subtraction volume and the T1-weighted MRI volume used for preprocessing, respectively, with $\Omega \in \mathbb{R}^3$ representing the volume lattice of size $w \times h \times d$; $\mathbf{s}_i^{(j)} : \Omega \rightarrow \{0, 1\}$ is the voxelwise annotation of the j^{th} lesion present in the breast $\mathbf{b}_i$ ($\mathbf{s}_i^{(j)}(\omega) = 1$ indicates the presence of lesion in voxel $\omega \in \Omega$, and $\mathbf{s}_i^{(j)}(\omega) = 0$ denotes the absence of lesion); $\{\mathbf{l}_i^{(j)}\}_{j=1}^{M} \in \{0, 1\}$ indicates the classification of lesion j as benign or malignant, respectively; and $\mathbf{y}_i$ is a scan-level label with the following values: $\mathbf{y}_i = 0$ if there is no lesion in breast $\mathbf{b}_i$, $\mathbf{y}_i = 1$ if all the lesion(s) in breast $\mathbf{b}_i$ are benign or $\mathbf{y}_i = 2$ if there is at least one malignant lesion. The dataset is patient-wise split into train $\mathcal{T}$, validation $\mathcal{V}$ and test $\mathcal{U}$ sets, such that images of each patient only belong to one of the sets. Note that the

voxelwise lesion annotations $\{\mathbf{s}_i^{(j)}\}_{j=1}^{M}$ and $\{\mathbf{l}_i^{(j)}\}_{j=1}^{M}$ are not employed during the training of the post-hoc system – they are only used to train and test the pre-hoc system and in the quantification of the results for both systems. Finally, the motivation behind the use of the first subtraction image $\mathbf{x}$ lies in the reduction of cost and time for image acquisition and analysis (Gilbert and Selamoglu, 2018; Mango et al., 2015).

3.2. Pre-hoc Method

Our proposed pre-hoc approach is based on the following steps:

1. **Lesion detection** (Sec. 3.2.1): an attention mechanism based on deep reinforcement learning (DRL) (Mnih et al., 2015) searches for lesions using a method that analyses large portions of the breast volume and iteratively focuses the search on the appropriate regions of the input volume.
2. **Lesion diagnosis** (Sec. 3.2.2): a state-of-the-art deep learning classifier (Huang et al., 2017) analyses the lesions detected in the previous step in order to classify them as malignant or non-malignant (note that non-malignant regions are represented by benign lesions or normal tissue, i.e. false positive detections). The confidence score of malignancy for the breast volume is defined as the maximum probability of malignancy among the detected lesions.

3.2.1. Lesion Detection

We propose an attention model that is capable of reducing the inference time of previous methods for lesion detection (Gubern-Mérida et al., 2015; McClymont et al., 2014) in pre-hoc systems. This attention mechanism searches for lesions by progressively transforming relatively large initial bounding volumes (BV) (*i.e.* sub-regions of the *MRI volume*) into smaller regions containing a more focused view of potential lesions (Maicas et al., 2017b). The transformation process is guided by a policy π that indicates how to optimally change the current BV to detect a lesion. The policy is represented by a deep neural network, called deep Q-net (DQN), that receives as input an embedding vector $\mathbf{o} \in \mathbb{R}^O$ of the current BV and outputs a measurement (i.e., the Q-value (Q)), representing the optimality associated with each of the possible transformations to find a lesion. See Figure 3 for a block diagram of this process. The aim of the learning phase is to model such policy, i.e., find the optimal parameters of the DQN. The inference exploits the policy to detect the lesions present in a breast DCE-MRI volume.

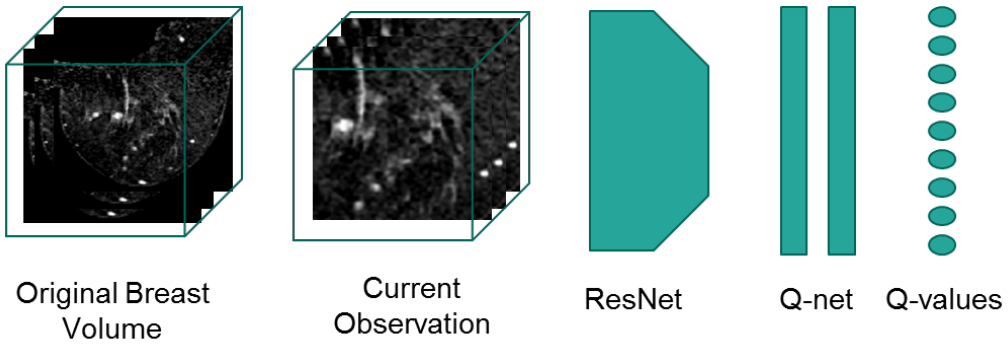


Figure 3: Overview of the proposed lesion detection method. The bounding volume of the current observation is extracted from the input breast volume and fed to the 3D ResNet to obtain the embedding of the observation. The embedding is then forwarded through the Q-net to obtain the Q-values for each of the actions.

The training process of the DQN follows that of a traditional Markov Decision Process (MDP), which models a sequence of decisions to accomplish a goal from an initial state. At every time step, the current BV, represented by the observations $\mathbf{o}$, will be transformed by an action a, yielding a reward r – this reward indicates the effectiveness of the chosen transformation for detecting a lesion. The goal is to learn what actions should be applied to transform the current observation to another one with larger Dice coefficient measured with respect to the target lesion. In an MDP set-up, this translates into choosing the action that maximizes the expected sum of discounted future rewards (Mnih et al., 2015): $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$, where $\gamma \in (0, 1)$ is a discount factor.

Let $Q^\star(\mathbf{o}, a)$ be the optimal Action-Value Function representing the expected sum of discounted future rewards by choosing action a to transform the observation $\mathbf{o}$. The optimal Action-Value function follows the policy π, as in:

$$Q^\star(\mathbf{o}, a) = \max_{\pi} \mathbb{E}[R_t | \mathbf{o}_t = \mathbf{o}, a_t = a, \pi]. \quad (1)$$

Intuitively, $Q^\star(.)$ represents the *quality* of performing the action a given the current observation $\mathbf{o}$ to achieve the final goal. Therefore, the goal of the training process is to learn $Q^\star(.)$, which maximizes the commulative sum of expected discounted rewards.

The optimal $Q^\star(\mathbf{o}, a)$ can be computed iteratively using the *Bellman* equa-

tion and the Q-Learning algorithm (Sutton and Barto, 1998):

$$Q_{i+1}(\mathbf{o}_t, a_t) = \mathbb{E}_{\mathbf{o}_{t+1}}\Big[r_t + \gamma \max_{a_{t+1}} Q(\mathbf{o}_{t+1}, a_{t+1}) | \mathbf{o}_t, a_t\Big]. \tag{2}$$

However, since it is impractical to compute $Q(\mathbf{o}_t, a_t)$ due to the large size of the observation-action space, a DQN function approximator, represented by $Q(\mathbf{o}, a, \boldsymbol{\theta})$, can be used. The weights $\boldsymbol{\theta}$ of the DQN $Q(\mathbf{o}_t, a_t, \boldsymbol{\theta}_t)$ can be learned by minimizing the mean square error of the Bellman equation:

$$L(\boldsymbol{\theta}_t) = \mathbb{E}_{(\mathbf{o}_t, a_t, r_t, \mathbf{o}_{t+1}) \sim U(\mathcal{E})}\Big[\Big(\underbrace{r_t + \gamma \max_{a_{t+1}} Q(\mathbf{o}_{t+1}, a_{t+1}; \boldsymbol{\theta}_t^-)}_{\text{target}} - Q(\mathbf{o}_t, a_t; \boldsymbol{\theta}_t)\Big)^2\Big], \tag{3}$$

where $\boldsymbol{\theta}_t$ are the parameters of the DQN at iteration t, $\boldsymbol{\theta}_t^-$ are the weights of the target network (defined below) used to compute the target value at iteration t, and $U(\mathcal{E})$ is a batch of experiences uniformly sampled from the experience replay memory $\mathcal{E}_t$ (also defined below). The target network is used to compute the target values for each update of the weights of the DQN. The architecture of this target network is the same as that of the DQN and its parameters $\boldsymbol{\theta}_t^-$ contain the weights of the DQN at a previous iteration of the optimization process. The weights $\boldsymbol{\theta}_t^-$ are updated after every iteration through the entire training set from the parameters $\boldsymbol{\theta}_t$ at the iteration $t-1$ and maintained constant between updates: $\boldsymbol{\theta}_t^- = \boldsymbol{\theta}_{t-1}$. The experience-replay memory $\mathcal{E}_t = \{e_1, ..., e_t\}$ stores previous experiences denoted by $e_t = \{\mathbf{o}_t, a_t, r_t, \mathbf{o}_{t+1}\}$, where each e_t is collected at time step t by choosing the action a_t to transform from $\mathbf{o}_t$ into $\mathbf{o}_{t+1}$, yielding the reward r_t. We describe in the next paragraphs how to obtain the observations, to choose the actions and to compute the reward function.

The embedding $\mathbf{o}$ of the current BV is computed as:

$$\mathbf{o} = f_{ResNet}(\mathbf{x}(\mathbf{b}), \theta_{\mathbf{ResNet}}) \tag{4}$$

where $\mathbf{b} = [b_x, b_y, b_z, b_w, b_h, b_d] \in \mathbb{R}^6$ is a bounding volume, with the triplets (b_x, b_y, b_z) and (b_w, b_h, b_d) denoting the top-left-front and the lower-right-back corners of the bounding volume, respectively; the DCE-MRI data is represented by $\mathbf{x}$; and $f_{ResNet}(.)$ represents a 3D Residual Network (ResNet) (He et al., 2016). The training of the 3D ResNet in (4) relies on a binary loss function that differentiates between input bounding volumes with and with-

out lesions. The dataset to train this 3D ResNet is built by sampling random BVs that are labelled as positive if the Dice Coefficient with a ground truth lesion is larger than 0.6, and negative otherwise. We empirically found that such a relatively large threshold of 0.6 helped the model to focus more tightly on the lesions during the training process. Note that the training of the 3D ResNet with a potentially infinite number of BVs from different scales, sizes and locations allows us to obtain a rich collection of BVs without the need for a large training set.

The set $\mathcal{A} = \{l_x^+, l_x^-, l_y^+, l_y^-, l_z^+, l_z^-, s^+, s^-, w\}$ represents the actions to modify the current BV, where $\{l, s, w\}$ represent the translation, scale and trigger (to terminate the search for lesions) actions, respectively; the subscripts $\{x, y, z\}$ denote the horizontal, vertical and depth translation, and the superscripts $\{+, -\}$ represent the positive/negative translation or up/down scaling.

The reward function depends on the improvement in the lesion localisation process after selecting a specific action. For action $a \in \mathcal{A} \setminus \{w\}$, we measure the improvement in terms of the variation of the *Dice coefficient* after applying action a to transform the observation $\mathbf{o}_t$ to $\mathbf{o}_{t+1}$:

$$r(\mathbf{o}_t, a, \mathbf{o}_{t+1}) = sign(d(\mathbf{o}_{t+1}, \mathbf{s}) - d(\mathbf{o}_t, \mathbf{s})), \tag{5}$$

where $d(.)$ is the Dice coefficient between the bounding volume $\mathbf{o}$ and the ground truth $\mathbf{s}$. The intuition behind (5) is that the reward is positive if the Dice coefficient from observation $\mathbf{o}_t$ to observation $\mathbf{o}_{t+1}$ increases, and the reward is negative otherwise. The quantization in (5) avoids a deterioration of the training convergence due to small changes in $d(.)$ (Caicedo and Lazebnik, 2015).

The reward for the trigger action, $a = w$, is defined as:

$$r(\mathbf{o}_t, a, \mathbf{o}_{t+1}) = \begin{cases} +\eta & \text{if } \ d(\mathbf{o}_{t+1}, \mathbf{s}) \geq \tau_w \\ -\eta & \text{otherwise} \end{cases} \tag{6}$$

where $\eta > 1$ encourages the trigger action to finalize the search for lesions if the Dice coefficient with the ground truth $\mathbf{s}$ is larger than a pre-defined threshold τ_w.

Actions during the training process are selected according to a modified ϵ-greedy strategy to balance *exploration* and *exploitation* (Maicas et al., 2017b): with probability ϵ, a random action will be explored, and with probability

$1-\epsilon$, the action will be chosen from the current policy. During *exploration*, with probability κ, a random action is selected, and with probability $1-\kappa$, a random action from the actions that will produce a positive reward is selected. During *exploitation*, the action is selected according to the current policy: $a_t = \arg\max_{a_t} Q(\mathbf{o}_t, a_t; \boldsymbol{\theta}_t)$. The training process starts with $\epsilon = 1$, which decreases linearly, transitioning from pure exploration to mostly exploitation following the current policy as the model learns to detect lesions.

During **inference**, we exploit the learned policy to detect lesions. In practice, we propose several initial bounding volumes covering different relatively large portions of the DCE-MRI volume. Each initialization is processed independently and is iteratively transformed according to the action $a_t^\star$ indicated by the optimal action-value function:

$$a_t^\star = \arg\max_{a_t} Q(\mathbf{o}_t, a_t; \boldsymbol{\theta}^\star). \tag{7}$$

where $\boldsymbol{\theta}^\star$ represents the parameter vector of the trained DQN model learned with (3).

We define the set of detected lesions as $\mathcal{D}^{pre} = \{\mathcal{D}_i^{pre}\}_{i=1}^{|\mathcal{D}^{pre}|}$, where $\mathcal{D}_i^{pre}$ represents the i^{th} bounding volume, when the trigger action is selected to stop the inference process. If the trigger action is not selected after 20 iterations, the search for a lesion is stopped yielding no detection.

3.2.2. Lesion diagnosis

The detected lesions in $\mathcal{D}^{pre}$, formed during the lesion localization stage, are classified in terms of their malignancy. This binary classification is performed with a 3D DenseNet (Huang et al., 2017), trained using the detections from the training set to differentiate normal tissue and benign lesions (i.e., negative diagnoses) from malignant lesions (positive diagnosis). During inference, each detection $\mathcal{D}_i^{pre}$ is fed through the 3D DenseNet to obtain its probability of malignancy. Finally, the confidence score of malignancy of a breast is defined as the maximum of the malignancy probabilities computed from all the detected regions in such breast. The confidence score of malignancy for the breast volume with no detections is set to zero.

3.3. Post-hoc Method

Our proposed post-hoc approach is characterised by the following steps:

1. **Diagnosis** (Sec. 3.3.1): the classifier outputs the probability that a breast DCE-MRI volume contains a malignant lesion. Given the small

training dataset, the model is first meta-trained with a teacher-student curriculum learning strategy to learn to solve several tasks. Then, the classifier is fine-tuned to solve the breast screening diagnosis task.

2. **Lesion Localization** (Sec. 3.3.2): the detector is weakly-trained to localise malignant lesions on breast DCE-MRI volumes that have been positively classified in the diagnosis stage above. This lesion localisation process can be used to interpret the decision from the diagnosis stage.

3.3.1. Breast Volume Diagnosis

Meta-training aims to learn a model that can solve new given tasks (classification problems) as opposed to traditional classifiers that solve a specific classification problem. Traditionally, models for solving new tasks have been achieved by fine-tuning pre-trained models (Tajbakhsh et al., 2016). However, these pre-trained models are rarely available for 3D volumes and large datasets are still required. These limitations can be overcome by including a meta-training phase before training, where the model is presented with several classification tasks that need to be solved, where each task has a small training set. Eventually, the model learns to solve new tasks that contain small training sets.

As noted in our previous work (Maicas et al., 2018), the order in which to present classification tasks during meta-training influences the ability of the model to solve new tasks. Therefore, we propose to use the teacher-student curriculum learning strategy (Matiisen et al., 2017) that has been shown to outperform other strategies (Maicas et al., 2018).

We propose to meta-train the model to solve several related classification tasks, each containing a relatively small number of training images instead of training a classifier to distinguish volumes with any malignant findings from others containing no malignant lesions. Firstly, during the meta-training phase, our model learns to solve different tasks that are formed from our breast DCE-MRI datasets. The tasks to be presented to the model are selected via the teacher-student curriculum learning strategy and contain a small training set. Secondly, the training phase is similar to that of any traditional classifier and solves the breast screening task using the samples available from the training set. The difference in our approach lies in the employment of the meta-trained model as the initialization for the training process. As a result, when the meta-trained model is fine-tuned on the breast screening task with the small training set, it is able to efficiently

and effectively classify previously unseen volumes containing malignant findings (Maicas et al., 2018). Finally, the inference phase (or breast diagnosis) consists of feeding the input volumes to the classifier to estimate the probability that they contain a malignant finding. See Figure 4 for an overview of the volume diagnosis process.

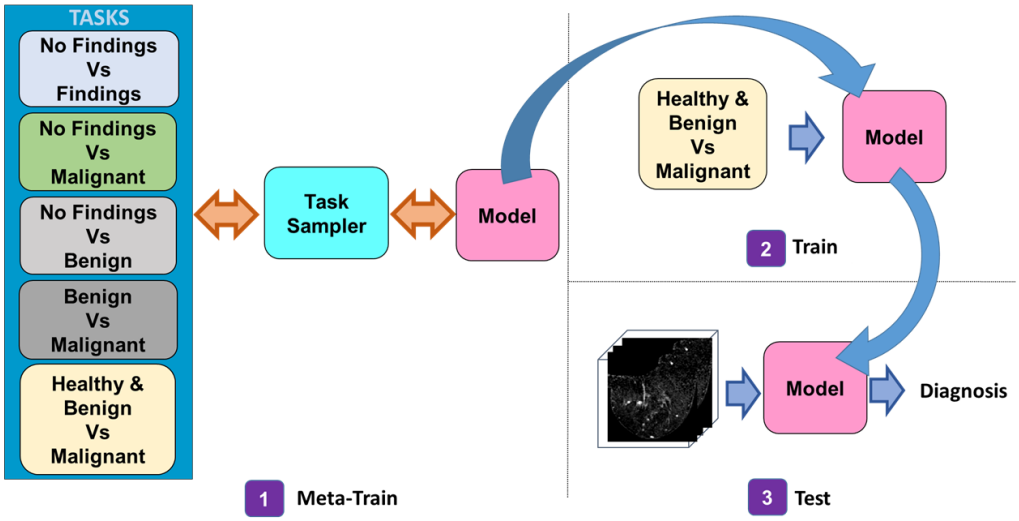


Figure 4: Volume diagnosis process. Firstly, the model is meta-trained on several related classification tasks. Secondly, the model is trained in the breast screening task. Finally, the model is tested on the breast screening task.

During **meta-training**, the model is meta-trained to solve the following five classification tasks:

1. K_1 : findings (lesions) versus no findings,
2. K_2 : malignant findings versus no findings,
3. K_3 : benign findings versus no findings,
4. K_4 : benign findings versus malignant findings,
5. K_5 : malignant findings versus no malignant findings (i.e., breast screening).

Let $K = \cup_{i=1}^{5} K_i$, where each task K_i is associated with a dataset $\mathcal{D}_i$ that contains the volumes from the training set that are relevant for the task K_i. We define the meta-training set $\mathcal{D} = \cup_{i=1}^{5} \mathcal{D}_i$.

Let the model to be meta-trained be defined by g_θ and the meta-update step be indexed by t. For each meta-update, a meta-batch $\mathcal{K}_t$ of tasks is sampled and contains $|\mathcal{K}_t|$ tasks from K (see below for a description of the task sampling method). For each of the tasks $K_j \in \mathcal{K}_t$, $N = N^{tr} + N^{val}$ volume-label samples are sampled from the corresponding meta-training set $\mathcal{D}_j$ to form $\mathcal{D}_j^{tr}$. Let $\mathcal{D}_j^{tr}$ contain N^{tr} samples that will be used as training set and $\mathcal{D}_j^{val}$ contain N^{val} samples that will be used as validation set during the t^{th} meta-update for the j^{th} task.

For every task $K_j \in \mathcal{K}_t$ in the meta-batch, the model is trained with $\mathcal{D}_j^{tr}$ to adapt to the task by performing several gradient descent updates. For simplicity, the adaptation of the model with one gradient descent update is defined by:

$$\theta_j'^{(t)} = \theta^{(t)} - \alpha \frac{\partial \mathcal{L}_{K_j}\left(g_{\theta^{(t)}}\left(\mathcal{D}_j^{tr}\right)\right)}{\partial \theta}, \tag{8}$$

where $\theta^{(t)}$ are the parameters of the model at meta-iteration t, $\mathcal{L}_{K_j}(g_{\theta^{(t)}}\left(\mathcal{D}_j^{tr}\right))$ is the cross-entropy loss computed from $\mathcal{D}_j^{tr}$ for task K_j, α is the learning rate for model adaptation, and $\theta_j'^{(t)}$ are the adapted parameters after performing model adaptation for task K_j.

The adapted models $g_{\theta_j'^{(t)}}$ are subsequently evaluated with the validation pairs $\mathcal{D}_j^{val}$ of the corresponding task. The loss produced by the validation set on each of the tasks is used to compute the meta-gradient associated to each task. Finally, the model parameters θ are updated using the average of the meta-gradients associated to each of the tasks in the meta-batch:

$$\theta^{(t+1)} = \theta^{(t)} - \beta \sum_{K_j \in \mathcal{K}_m} \frac{\partial \mathcal{L}_{K_j}\left(g_{\theta_j'^{(t)}}\left(\mathcal{D}_j^{val}\right)\right)}{\partial \theta}, \tag{9}$$

where β is the meta-learning rate and $\mathcal{L}_{K_j}\left(g_{\theta_j'^{(t)}}\left(\mathcal{D}_j^{val}\right)\right)$ is the cross entropy loss of the validation volumes in $\mathcal{D}_j^{val}$ for task K_j. This procedure is repeated for M meta-iterations, as shown in Alg. 1.

The breast screening **training** process is initialised by the meta-trained model. Using the entire training set $\mathcal{T}$, the model adapts to the breast screening task by performing several gradient descent updates, similarly to the training of a traditional deep learning classifier. We use the validation set $\mathcal{V}$ for model selection. The **inference** of the model is similar to that of

Algorithm 1 Overview of the meta-training procedure presented in (Maicas et al., 2018)

procedure META-TRAIN($\{K_1 \dots K_5\}$, $\{\mathcal{D}_1 \dots \mathcal{D}_5\}$, model g_θ)
  Initialise parameters θ from g_θ
  for $t = 1$ *to* T **do**
    Sample meta-batch $\mathcal{K}_t$ by sampling $|\mathcal{K}_t|$ tasks from $\{K_1 \dots K_5\}$
    for each task $K_j \in$ meta-batch $\mathcal{K}_t$ **do**
      Adapt model using (8) with samples from $\mathcal{D}_j^{tr}$
      Evaluate adapted model using with samples from $\mathcal{D}_j^{val}$
    Meta-update model parameters with (9)

any standard classifier and consists of feeding the testing volume through the network to obtain the probability of malignancy of each of the input volumes. The confidence score of malignancy corresponds to the probability of malignant output by the classifier.

During the **meta-learning process**, the **task sampling** process to form a meta-batch of tasks depends on the past observed performance improvements of the model in each of the tasks. This has been shown to outperform other alternative approaches (Maicas et al., 2018). A partially observable Markov decision process (POMDP) solved using reinforcement learning with Thompson Sampling can model such an approach. A POMDP is characterized by observations, actions, and rewards. In our set-up, we define an observation O_{K_j} as the variation in the area under the receiving operating characteristic curve (AUC) of the adapted model $\theta_j'^{(t)}$ compared to the initial AUC before the model $\theta^{(t)}$ was adapted to the task $K_j \in \mathcal{K}_t$ – in both cases, the AUC is measured using the sampled validation set $\mathcal{D}_j^{val}$. The actions correspond to sampling a particular task. The reward is defined as the difference between the current and previous observations during the last time that the task was sampled. The goal is to decide which action to apply, i.e. which task should be sampled for the next meta-training iteration. We use Thompson sampling to decide the next task to be sampled, which allows us to balance between sampling new tasks, and sampling tasks for which the improvement of performance is currently higher (similar to the exploration-exploitation dilemma in reinforcement learning) (Matiisen et al., 2017).

Let $\mathcal{B}_j$ be a buffer of recent rewards for task K_j – this buffer stores the last B rewards for this task. To perform Thompson sampling, a random recent reward $R_{\mathcal{B}_j} \in \mathcal{B}_j$ is uniformly sampled. The next task K_j to be included in

the meta-batch $\mathcal{K}_t$ of iteration t is selected with $j = \arg\max_i |R_{\mathcal{B}_i}|$. This process is repeated for $|\mathcal{K}_t|$ times to form $\mathcal{K}_t$. The intuition behind this is that for tasks where performance is increasing rapidly (i.e. yielding higher rewards) they will be sampled more frequently until mastered (i.e. the reward will tend to zero as the variation in AUC after adaptation will tend to be smaller in consecutive iterations). Then, a different task will be sampled more frequently. However, if the model reduces the performance in the previously mastered task, it will be sampled again more frequently because the absolute value of the reward will tend to be higher again.

3.3.2. Malignant Region Localization

A breast volume is diagnosed as malignant in the previous step if its confidence score of malignancy is higher than the equal error rate (EER) of the proposed classifier on the validation set. The EER as threshold is chosen to avoid any preference between sensitivity and specificity. For positively classified volumes, we aim to generate a saliency map represented by a binary mask indicating the localization of lesions that can explain the decision made by the classifier; while for negatively classified volumes, no salient region is produced. Therefore, we propose a 1-class saliency detector (Maicas et al., 2019) that has been specifically designed to satisfy these conditions.

Our 1-class saliency detector is modelled with a weakly-supervised training process to detect salient regions in positively classified volumes, where these regions denote malignant lesions. The detector follows an encoder-decoder architecture that generates a mask $\mathbf{m} : \Omega \rightarrow [0, 1]$ of the same size as the input volume, where this mask localizes the most salient regions of the input volume that are involved in the positive classification. The encoder is the classifier from Sec. 3.3.1, which produces the diagnosis. The decoder upsamples the output from the encoder to the original resolution from the lowest resolution feature maps by concatenating four blocks of feature map resize, convolution layer, batch normalization layer and ReLU activation (Zeiler and Fergus, 2014). Skip connections are used to connect corresponding layers of the same resolution in the encoder and decoder. During training, the parameters of the encoder are fixed and the parameters of the decoder are updated using the gradient corresponding to the following loss for each volume $\mathbf{x}_i$:

$$\ell_i(\mathbf{m}) = \lambda_1 \ell_{TV}(\mathbf{m}) + \lambda_2 \ell_A(\mathbf{m}) - y_i \lambda_3 \ell_P(\mathbf{m}, \mathbf{x}_i) + y_i \lambda_4 \ell_D(1 - \mathbf{m}, \mathbf{x}_i), \quad (10)$$

where ℓ_{TV} measures the total variation of the mask forcing the boundary of

salient regions to be relatively smooth, ℓ_A measures the area of the salient regions and aims to reduce the total area of regions, ℓ_P measures the confidence in the classification of the input volume $\mathbf{x}_i$ masked with $\mathbf{m}$, and ℓ_D measures the confidence in the classification of the input volume $\mathbf{x}_i$ masked with the inverse of the generated mask, i.e $(1 - \mathbf{m})$.

By **training** the mask generator model with the loss function (10), there is an explicit relationship between saliency and malignant lesions (Maicas et al., 2019). By setting $y_i = 0$ for negative volumes, they are forced to have no salient regions. For positives volumes, salient regions are forced to have the following characteristics: 1) be small and smooth, 2) when used to mask the input volume, the classification result is positive; and 3) when its inverse is used to mask the input volume, the classification result is negative. During **inference**, volumes diagnosed as positive are fed forward through the decoder to produce a mask, where each voxel has values in $[0, 1]$. This mask is thresholded at ζ to obtain the malignant lesions.

4. Experiments

In this section, we describe the dataset and experimental set-up used to assess the proposed methods for the problems of breast screening and malignant lesion detection.

4.1. Dataset

Our methods are evaluated with a dataset containing MRI scans from 117 patients. The dataset is split in a patient-wise manner into training, validation and test sets using the same split as previous approaches (Maicas et al., 2017a,b, 2018, 2019), so we can directly compare our results with previous works. The training set contains scans from 45 patients, where these scans show 38 malignant lesions and 19 benign lesions – the scans also show that 29 of the patients have at least one malignant lesion while 16 only have benign lesion(s). The validation set has scans from 13 patients, with 11 malignant and 4 benign lesions – these scans show that 9 of the patients have at least one malignant lesion while 4 patients have only benign lesion(s). The test set contains scans from 59 patients, with 46 malignant and 23 benign lesions – the scans show that 37 of the patients have at least one malignant lesion while 22 have only benign lesion(s). A biopsy is performed to characterize the lesion in a breast. If there are multiple lesions in the same breast, the lesion

that seems to have the higher chance of malignancy is biopsied. An experienced breast radiographer annotated the remaining lesions by analyzing the image based on the pathology report. The types of benign lesions included in this work are (Mcclymont, 2015): fibrocystic change (22%), fibroadenoma (18%) and other (60%); and the types of malignant lesions included in this work are (Mcclymont, 2015): ductal carcinoma in situ (31%), invasive ductal carcinoma (33%), invasive lobular carcinoma (11%) and other (25%). Note that BIRADS=3 lesions (fibroadenomas (Lee et al., 2018)) are considered benign in our study. Every patient has at least one lesion, but not every breast contains lesions.

There are 42, 13, and 58 breasts with no lesions in the training, validation and testing sets, respectively. Likewise, 18, 4, and 22 breasts contain only benign lesions (i.e. are considered "benign") and 30, 9, and 38 contain at least one malignant lesion (i.e. are considered "malignant"). For the breast screening problem, "Malignant" breasts are considered positive while "benign" and breasts with no lesions are considered negative.

The MRI dataset (McClymont et al., 2014) contains T1-weighted and two dynamic contrast enhanced (pre-contrast and first post-contrast) volumes for each patient acquired with a 1.5 Tesla GE Signa HDxt scanner. The T1-weighted anatomical volumes were acquired without fat suppression and with an acquisition matrix of 512×512. The DCE-MRI images are based on T1-weighted volumes with fat suppression, with an acquisition matrix of 360×360 and a slice thickness of 1 mm. Firstly, a pre-contrast volume was acquired before a contrast agent was injected. The first post-contrast volume was acquired after a delay of 45 seconds after the acquisition of the pre-contrast. The first subtraction volume is formed by subtracting the pre-contrast volume to the first post-contrast volume. Both T1-weighted and DCE-MRI were acquired axially.

The dataset was preprocessed using the T1-weighted volume to segment the breast region from the chest wall using Hayton's method (Hayton et al., 1997; McClymont et al., 2014). This involves removing the pectoral muscle which may produce false positive detections. In addition, the breast region was divided into left and right breasts by splitting the volume in halves, as the breast region was initially centred. Each breast volume was resized to a size of $100 \times 100 \times 50$ voxels. Note that we operate the proposed methods breast-wise.

4.2. Experimental Set-up

The aim of the experiments is to assess our pre-hoc and post-hoc approaches in terms of their performance for diagnosing malignancy and localising malignant lesions from breast DCE-MRI. Firstly, we individually evaluate the components of our proposed pre-hoc and post-hoc methods. Secondly, we compare the performance of both approaches in terms of diagnosis accuracy and malignant lesion localisation. For the full pipeline comparison, we additionally provide an estimate of the standard errors of our results. We estimate the standard errors as follows: for the AUC of the diagnosis, we utilised an estimate based on the Wilcoxon test (Hanley and McNeil, 1983; Bradley, 1997); and for the diagnosis ROC and malignant lesion localisation FROC curves, we applied a jackknife estimate (Bishop, 2006) on the test set that provides both the average and standard error results. Note that in every localisation evaluation we consider a region to be true positive if the Dice coefficient measured between a candidate region and the ground truth lesion is at least 0.2 (Dhungel et al., 2015; Maicas et al., 2017b, 2019).

4.2.1. Pre-hoc System

The lesion detection step in the pre-hoc approach is evaluated in terms of the free response operating characteristic (FROC) curve measured in a patient-wise manner (as in previous detection works (Gubern-Mérida et al., 2015; Maicas et al., 2017b, 2019)), which compares the true positive rate (TPR) against the number of false positive detections per patient (FPP). We also measure the inference time in a computer with the following configuration: Intel Core i7, 12 GB of RAM and a GPU Nvidia Titan X 12 GB. As in previous diagnosis work (Maicas et al., 2017b), the diagnosis step in the pre-hoc method is evaluated in terms of the area under the receiving operation characteristic curve (AUC), which compares true positive diagnosis rate against false positive diagnosis rate. The AUC is measured in a breast-wise way and considers all the breasts in the testing set when computing the true positive diagnosis rate and false positive diagnosis rate.

The **lesion detection** uses a 3D ResNet trained from scratch with random bounding volumes sampled from the training volumes. More specifically, we sample 8000 positive and 8000 negative patches that are resized to $100 \times 100 \times 50$ (the input size to the 3D ResNet). The choice of the input size of the ResNet is $100 \times 100 \times 50$ so that every lesion is visible – some tiny lesions disappear at finer resolutions. The architecture of the 3D ResNet (He et al., 2016) comprises 5 Residual Blocks (Huang et al., 2016),

each of them preceded by a convolutional layer. After the last residual block, the model contains two additional convolutional layers and a fully connected (FC) layer. The embedding of the observation "$\mathbf{o}$" is the output of the second to last convolutional layer, before the FC layer and it has 2304 dimensions.

The DQN is a 2-layer multi-layer perceptron, with each layer containing 512 nodes. It outputs the Q-value for 9 actions: translation by one third of the size of the observation in the positive or negative direction on each of the dimensions (i.e. 6 actions), scaling by one sixth of the size of the observation and is applied in every dimension (i.e. 2 actions) and the trigger action. The reward value for the trigger action has been empirically defined as $\eta = 10$ if $\tau_w = 0.2$ (i.e., the Dice coefficient is at least 0.2 during the trigger action), and the discount factor is $\gamma = 0.9$. The DQN is trained with batches of 100 experiences from the experience replay memory $\mathcal{E}$, which can store 10000 experiences. We use Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-6}.

During training, the model is initialized with one centred large observation covering 75% of the input breast volume. During inference, the lesion detection algorithm is launched from 13 different initializations in order to increase the chances of finding all possible lesions present in a breast. In addition to the same initialization used during training, eight initializations are placed in each of the eight $50 \times 50 \times 25$ corner volumes, and four $50 \times 50 \times 25$ initializations are placed centred between the previous 8 initializations. The balance between exploration and exploitation during training is given by ϵ, which decreases linearly from $\epsilon = 1$ to $\epsilon = 0.1$ after 300 epochs, and by $\kappa = 0.5$.

Detected regions are resized to $24 \times 24 \times 12$, which is the median value of the size of all detections in the training set. The decision behind this size is based on the following empirical finding: we noticed that larger sizes added too much noise due to the upsampling of tiny lesions and smaller sizes removed important details. The **lesion diagnosis** uses a 3D DenseNet (Huang et al., 2017) composed of three dense blocks of two dense layers each. Each dense layer comprises a batch normalization, ReLu and a convolutional layer. In the particular DenseNet implementation used in this paper, we use a compression of 0.5 and a growth rate of 6. Global average pooling of $6 \times 6 \times 3$ is applied after the last dense block and before the fully connected layer. The DenseNet is optimized with stochastic gradient descent with a learning rate of 0.01. The dataset used to train the 3D DenseNet is composed of all detections obtained from the training set. Model selection is performed us-

ing the detections from the validation set based on the breast-wise AUC for breast screening. Note that detections that correspond to malignant lesions are labelled as positive while detections that correspond to benign lesions or false positives are labelled as negative.

4.2.2. Post-hoc System

The **diagnosis** step in the post-hoc approach is evaluated with the breast-wise AUC. The malignant lesion localization step in the post-hoc approach is evaluated in terms of the patient-wise FROC curve under two different scenarios: 1) all the patients in the test set are considered to compute the FROC (A), and 2) only the patients in the test set that had at least one breast diagnosed as malignant are considered (+) – this scenario allows the computation of the performance of the 1-class saliency detector in an isolated manner. Note that for both scenarios (A) and (+), region detections from non-malignant breasts classified as malignant are considered false positive detections.

The breast volume diagnosis meta-training algorithm uses as the underlying model a 3D DenseNet (Huang et al., 2017). The architecture was decided based on the optimization of a 3D DenseNet (trained with the training set $\mathcal{T}$) to achieve the best results for the breast screening task on the validation set $\mathcal{V}$ and consists of 5 dense blocks with 2 dense layers each. Each dense layer comprises a batch normalization, ReLu and convolutional layer, where compression was 0.5 and growth rate 6. No data augmentation or dropout were used since they did not improve the performance of this 3D model. For meta-training, the learning rate is $\alpha = 0.01$ and the meta-learning rate is $\beta = 0.001$. The number of gradient descent steps during adaptation is 5 and the number of meta-iterations is $M = 3000$. The meta-batch size contained $|\mathcal{K}_t| = 5$ tasks, where each task had $N^{tr} = 4$ samples for training and $N^{val} = 4$ for validation. Each buffer $\mathcal{B}_j$ stored 40 recent rewards.

The localisation of **malignant lesions** in positively classified volumes is achieved by thresholding the generated saliency map at $\zeta = 0.8$ – this threshold was decided based on the detection performance in the validation set. The parameters for training the 1-class saliency detector in (10) are: $\lambda_1 = 0.1$, $\lambda_2 = 3$, $\lambda_3 = 1$, and $\lambda_4 = 2.5$.

4.2.3. Comparison Between Pre- and Post-Hoc

Using the set-up described above for pre-hoc and post-hoc approaches, we compare the performance of both methods. We evaluate diagnosis in

	Inference Time Per Patient
DQN (13 Initializations)	$92 \pm 21s$
MS-SL	$164 \pm 137s$
Cascade	$\mathcal{O}(60)min$

Table 1: Inference time per patient of our proposed pre-hoc detection method (DQN using 13 initializations per breast), the MS-SL (mean-shift structured learning), and the multi-scale cascade baselines.

breast-wise and patient-wise manners in terms of the area under the receiving operation characteristic curve (AUC). The standard error for the AUC is estimated with the Wilcoxon test (Hanley and McNeil, 1983; Bradley, 1997). We also evaluate the performance of malignant lesion localisation for each approach in a patient-wise way using the FROC curve as in previous works (Gubern-Mérida et al., 2015; Maicas et al., 2017b, 2019). We estimate the average TPR and standard error using a jackknife estimate on the test set (Bishop, 2006). Note that the TPR for breast-wise malignant lesion localization is the same as patient-wise, while the breast-wise FPP is the same as the one for patient-wise divided by two. For the post-hoc method, we also plot the two scenarios (A) and (+), explained above.

4.2.4. Experimental Results for the Pre-hoc System

We compare the performance of our **lesion detection** step against an improved version of exhaustive search, namely a multi-scale cascade based on deep learning features (Maicas et al., 2017a), and a mean-shift clustering method followed by structured learning (McClymont et al., 2014) (note that only one operating point is available for this approach), which is evaluated on the same dataset using a different training and testing data split. Figure 5 shows the FROC curve with the detection results and Table 1 contains the inference times per patient needed by each of the methods.

The **diagnosis** of breast volumes, based on the classification of the detected regions, achieves an AUC of 0.85, if all volumes in the test dataset are considered.

4.2.5. Experimental Results for the Post-hoc System

We evaluate the performance of our post-hoc diagnosis against three state-of-the-art classifiers. The first baseline is the 3D DenseNet (Huang et al., 2017) that has been optimized to solve the breast screening problem

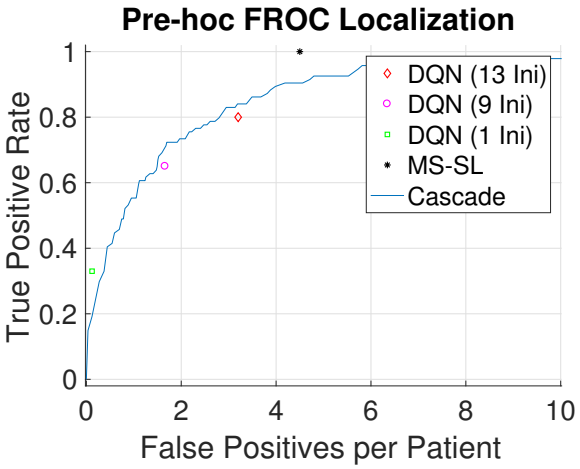


Figure 5: FROC curve per patient for the lesion detection step of our pre-hoc method, labelled as DQN, where the information in brackets refers to the number of initialisations per breast used during the inference process. MS-SL refers to the mean-shift structured learning approach, and Cascade denotes the multi-scale cascade method based on deep learning features.

(as explained in Sec. 4.2.2). The second baseline is the same 3D DenseNet fine-tuned using a multiple instance learning (MIL) set-up (Zhu et al., 2017), which holds the state-of-the-art for the breast screening problem from mammography. Finally, we compare against a 3D DenseNet trained from scratch using multi-task learning (Xue et al., 2018), such that the model is jointly trained to solve all the tasks defined in Sec. 3.3.1. See Table 2 for the AUC diagnosis results.

Our 1-class saliency detector specially designed to **detect malignant**

Baseline	**AUC**
Meta-Training(**Ours**)	**0.90**
Multi-task (Xue et al., 2018)	0.85
MIL (Zhu et al., 2017)	0.85
DenseNet (Huang et al., 2017)	0.83

Table 2: Breast-wise AUC for diagnosis in post-hoc systems. Our proposed post-hoc diagnosis method based on meta-training is labelled as Meta-Training, while the baseline based on multiple instance learning is labelled as MIL and the one based on multi-task learning is denoted as Multi-task.

lesions in positively classified volumes is compared against the following baselines: CAM (Zhou et al., 2016), and Grad-CAM and Guided Grad-CAM (Selvaraju et al., 2017). Figure 6 shows the FROC curves for our proposed methods and baselines in each of the two scenarios (A) and (+).

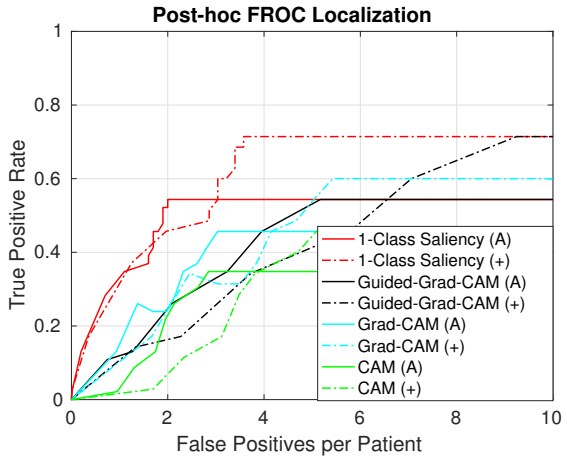


Figure 6: Patient-wise FROC curves for post-hoc malignant lesion detection, where our method is denoted as 1-Class Saliency. Baselines are denoted as CAM (Zhou et al., 2016), and Grad-CAM and Guided Grad-CAM (Selvaraju et al., 2017). For each method, we present two scenarios: (A) all the volumes in the test set are considered to compute the FROC, and (+) only positively classified volumes are considered.

4.2.6. Experimental Results for the Comparison Between Pre- and Post-Hoc

Table 3 contains the AUC for the malignancy diagnosis measured breast-wise and patient-wise for the pre-hoc and post-hoc approaches. Figure 7 shows the ROC curves used in the computation of the AUC in Table 3. Figure 8 shows the FROC curves for malignant lesion detection of pre-hoc and post-hoc ((A) and (+)) methods. Figures 9, 10, and 11 display examples of breast diagnosis and lesion localizations obtained from the proposed pre-hoc and post-hoc methods, where both methods correctly performed diagnosis (Fig. 9), only the pre-hoc method correctly diagnosed the breast (Fig. 10), and only the post-hoc method correctly diagnosed the breast (Fig. 11).

5. Discussion

The localization step in the pre-hoc method achieves similar accuracy to the baseline methods. As shown in Figure 5, the TPR and FPP directly

	Pre-Hoc	Post-Hoc
Breast-wise	0.85 ± 0.04	$\mathbf{0.90 \pm 0.04}$
Patient-wise	0.81 ± 0.05	$\mathbf{0.91 \pm 0.04}$

Table 3: AUC comparing the diagnosis performance between pre-hoc and post-hoc measured breast-wise and patient-wise.

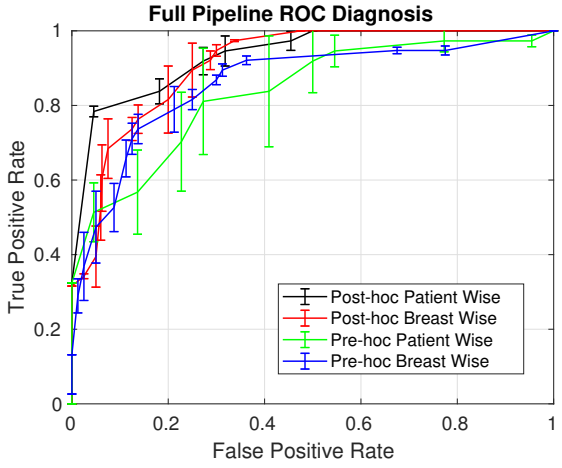


Figure 7: ROC curves for malignancy diagnosis of pre-hoc and post-hoc full pipelines measures breast and patient-wise. See Table 3 for the AUCs and estimated error.

depends on the number of initializations used by the reinforcement learning algorithm. In addition, the performance of our localization step is very similar to the baseline based on a multi-scale cascade using exhaustive search with deep features. However, multi-scale cascade (164s) and clustering+structure learning (several hours) methods require large inference times compared to our attention model (92s) as shown in Table 1.

The post-hoc diagnosis step improves over several baseline methods, as shown in Table 2. These baseline methods are based on a DenseNet (Huang et al., 2017), specifically optimised for the breast screening classification, and on extensions derived from multiple instance learning (Zhu et al., 2017) and multi-task learning (Xue et al., 2018). These results show that meta-training the model to solve tasks with small training sets is an important step to improve the learning of methods when only small datasets are available. Baseline approaches (Huang et al., 2017; Xue et al., 2018) only show a limited improvement over the DenseNet baseline.

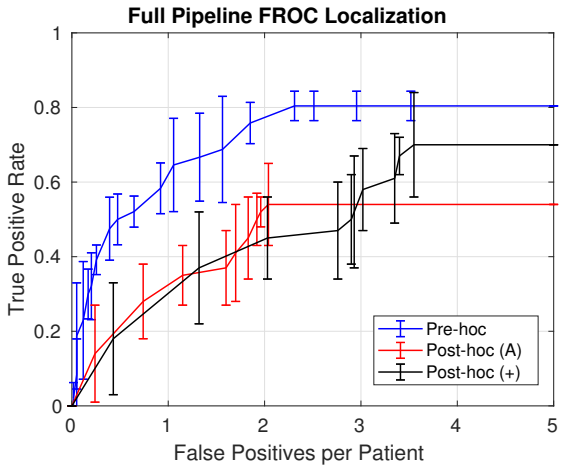


Figure 8: Patient-wise FROC curve for malignant lesion detection of pre-hoc and post-hoc full pipeline methods. For the post-hoc method, we present the two scenarios (A) and (+).

The localization step in our post-hoc method benefits from our definition of saliency, as shown in Figure. 6. In contrast, baseline methods show activations that do not correlate well with the target classification. In addition, baseline methods, such as CAM (Zhou et al., 2016) and Grad-Cam (Selvaraju et al., 2017), suffer from the low resolution of the activation feature maps, despite the improvement achieved by Guided Grad-Cam (Selvaraju et al., 2017). Measuring results only on positively classified volumes ((+) curves in Figure 6) discounts the mistakes made by the diagnosis step and provides an evaluation that isolates the lesion detection ((A) curves in Figure 6). Note that there is no straightforward comparison between the localization steps in post-hoc methodologies (that only detects malignant lesions) and the localization step in pre-hoc methodologies (that detects benign and malignant lesions). Such malignant lesion detection comparison only makes sense in terms of the full pre-hoc and post-hoc pipelines, which is detailed below.

In terms of the full pipeline, we observe in Table 3 and Figure 7 that the post-hoc system has a higher classification AUC than the pre-hoc for breast screening from breast DCE-MRI. The difference between these two methods is higher when measured patient-wise compared to breast-wise. It seems reasonable to think that the reason behind such discrepancy is an effect of the missed detections in pre-hoc. In difficult (small and low contrast lesions) cases with missed detections, the confidence score of malignancy of a breast is considered 0. While this effect is smaller when the AUC is measured

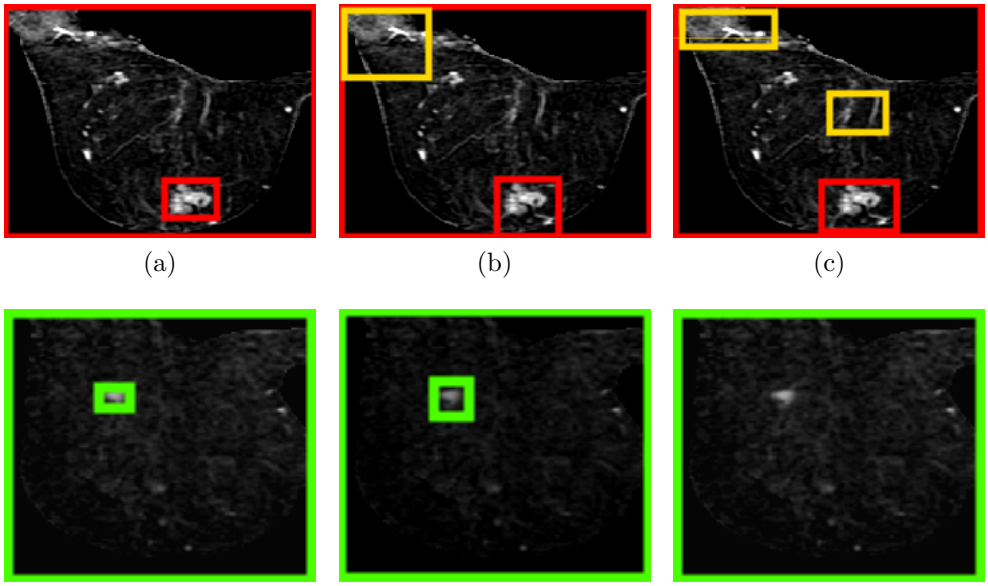


Figure 9: Example of two correct diagnosis by both pre-hoc and post-hoc full pipeline methods. Left column is the ground truth, middle column is the result of the pre-hoc method and right column is the result of the post-hoc method. Red image frames indicate malignant diagnosis, green frames indicate non-malignant diagnosis. Detections in red indicates TP malignant detections, yellow detections indicate FP malignant detections, detections in green indicate benign lesions. **First row:** pre-hoc and post-hoc correct positive diagnosis with the malignant lesion detected. **Second row:** pre-hoc and post-hoc correct negative diagnosis where the pre-hoc method correctly classified as negative a detected benign lesion and the post-hoc method did not localize any malignant lesion.

breast-wise (as there are 118 samples of breasts), it is larger when measured patient-wise (59 samples of patients). Furthermore, the better results of the post-hoc method suggest that the analysis of the whole image allows it to find indications for malignancy that are located in other areas of the image (Kostopoulos et al., 2017).

Regarding the localization of malignant lesions, the pre-hoc system achieves better accuracy, compared with the post-hoc. This suggests that the strong annotations used to train the pre-hoc method gives it an advantage for the localisation of lesions, when compared with the weak annotation used to train the post-hoc approach. This issue is exemplified in Figure 9 (Row 1), where although both approaches present a correct diagnosis, the post-hoc method yields a higher number of false positive malignant lesion detections. A simi-

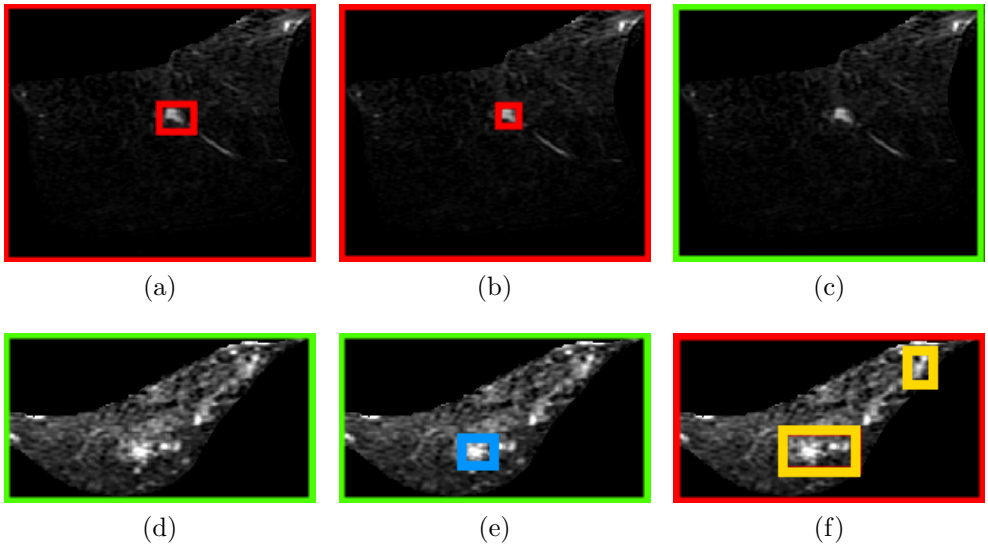


Figure 10: Example of two correct diagnosis by the pre-hoc system, but wrongly diagnosed by the post-hoc method. Left column is the ground truth, middle column is the result of the pre-hoc method and right column is the result of the post-hoc method. Red image frames indicate malignant diagnosis, green frames indicate non-malignant diagnosis. Detections in red indicate TP malignant detections, yellow detections indicate FP malignant detections, detection in blue indicates a ROI detection correctly classified as negative (non-malignant). **First row:** correct positive diagnosis by the pre-hoc method with the malignant lesion correctly detected but incorrect non-malignant diagnosis by the post-hoc method. **Second row:** correct negative diagnosis by the pre-hoc method, but incorrect positive diagnosis by the post-hoc system – yielding the potential malignant regions in the rectangles shown in yellow.

lar behaviour can be seen in Figure 10 (Row 2), where the post-hoc produces an incorrect diagnosis and additionally yields two false positive detections. In addition, the detection step for the pre-hoc system is mainly designed to achieve good performance when only a small training set is available. On the contrary, the malignant lesion localization step in the post-hoc approach is not particularly focused on being able to perform well from a small dataset. This difference in design focus is likely to be influencing the detection results too. Finally, it is worth noting that the FROC for the post-hoc approach (Post-hoc (A) curve in Fig. 8) is affected by the diagnosis process. However, if we remove the effect of the diagnosis step and consider the performance of malignant lesion localization in positively classified volumes, we observe a closer performance compared to the pre-hoc method, even though the post-

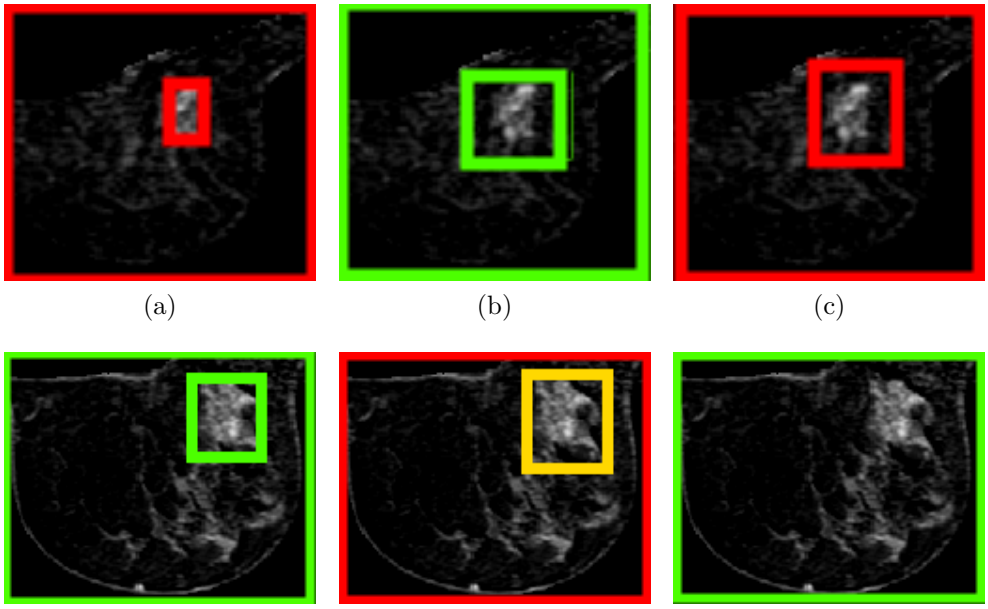


Figure 11: Example of two incorrect diagnosis by the pre-hoc, but correctly diagnosed by the post-hoc method. Left column is the ground truth, middle column is the result of the pre-hoc method and right column is the result of the post-hoc method. Red image frames indicate malignant diagnosis, green frames indicate non-malignant diagnosis. Detections in red indicate TP malignant detections, yellow detections indicate FP malignant detections, detection in green indicates a benign ROI detection. **First Row:** the pre-hoc system incorrectly diagnoses as negative, while post-hoc system correctly diagnoses as positive and yields the malignant lesion. **Second row:** the post-hoc method correctly diagnoses as negative, but pre-hoc incorrectly diagnoses as positive due to the wrong positive classification of a detected lesion.

hoc system is trained with weak annotations.

6. Limitations and Future Work

The main limitation of our work comes from the small dataset available. In addition to a larger test set, we aim to increase our dataset to include patients where no lesions are found in order to better recreate the scenario of a screening population. Ideally, this dataset will contain scanners from different vendors too.

Future work involves the development of a method that can combine both strongly and weakly annotated data to exploit the advantages of each

approach for lesion detection and diagnosis. We will also focus on the improvement of the malignant lesion localization in post-hoc methodologies by designing a new method specifically for the small training set available. We believe that the lesion localization step in pre-hoc approaches could also be improved in terms of inference time by running the different initializations of the detection algorithm in parallel and by optimizing the resizing operation of the current bounding volume (Maicas et al., 2017b). In addition, the use of a U-net (Ronneberger et al., 2015) would allow the implementation of a faster segmentation map maintaining the detection accuracy. We also plan to extend the diagnosis stage of the pre-hoc method by building a classifier that is trained similarly to the proposed for the diagnosis step of the post-hoc approach. Finally, it would be interesting to design a method that could diagnose based on the combined analysis of MRI and mammography.

7. Conclusion

We introduced and compared two different approaches for breast screening from breast DCE-MRI: pre-hoc and post-hoc methods. The pre-hoc method localizes suspicious regions (benign and malignant lesions) using an attention model based on deep reinforcement learning. Detected regions were subsequently classified into malignant or non-malignant lesions using a 3D DenseNet. The post-hoc method diagnoses a DCE-MRI breast volume using a classifier that, before being trained to solve the breast screening task, has been meta-trained to solve several breast-related tasks where only small training sets are available. Malignant regions are then localized with a 1-class saliency detector specifically designed for post-hoc systems that perform diagnosis. Results showed that the post-hoc method can achieve better performance for malignancy diagnosis, whereas the pre-hoc method could more precisely localize malignant lesions. However, this improvement of the pre-hoc detection method relies on the employment of strong annotations during the training process. On the other hand, post-hoc methods only use weak labels during the training phase and outperforms pre-hoc methods in diagnosis, which is the main aim of a breast screening system. In conclusion, we believe that future research should focus on the development and improvement of post-hoc diagnosis methods.

References

Agner, S.C., Rosen, M.A., Englander, S., Tomaszewski, J.E., Feldman, M.D., Zhang, P., Mies, C., Schnall, M.D., Madabhushi, A., 2014. Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced mr images: a feasibility study. Radiology 272, 91–99.

AIHW, 2007. Cancer in Australia 2017. Technical Report. The Australian Institute of Health and Welfare.

Amit, G., Ben-Ari, R., Hadad, O., Monovich, E., Granot, N., Hashoul, S., 2017a. Classification of breast mri lesions using small-size training sets: comparison of deep learning approaches, in: Medical Imaging 2017: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 101341H.

Amit, G., Hadad, O., Alpert, S., Tlusty, T., Gur, Y., Ben-Ari, R., Hashoul, S., 2017b. Hybrid mass detection in breast mri combining unsupervised saliency analysis and deep learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 594–602.

Behrens, S., Laue, H., Althaus, M., Boehler, T., Kuemmerlen, B., Hahn, H.K., Peitgen, H.O., 2007. Computer assistance for mr based diagnosis of breast cancer: present and future challenges. Computerized medical imaging and graphics 31, 236–247.

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Bradley, A.P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition 30, 1145–1159.

Caicedo, J.C., Lazebnik, S., 2015. Active object localization with deep reinforcement learning, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2488–2496.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N., 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 1721–1730.

Chen, W., Giger, M.L., Bick, U., 2006. A fuzzy c-means (fcm)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced mr images. Academic radiology 13, 63–72.

Dalmış, M.U., Gubern-Mérida, A., Vreemann, S., Karssemeijer, N., Mann, R., Platel, B., 2016. A computer-aided diagnosis system for breast dce-mri at high spatiotemporal resolution. Medical physics 43, 84–94.

Dalmış, M.U., Vreemann, S., Kooi, T., Mann, R.M., Karssemeijer, N., Gubern-Mérida, A., 2018. Fully automated detection of breast cancer in screening mri using convolutional neural networks. Journal of Medical Imaging 5, 014502.

DeSantis, C.E., Bray, F., Ferlay, J., Lortet-Tieulent, J., Anderson, B.O., Jemal, A., 2015. International variation in female breast cancer incidence and mortality rates. Cancer Epidemiology and Prevention Biomarkers 24, 1495–1506.

Dhungel, N., Carneiro, G., Bradley, A.P., 2015. Automated mass detection in mammograms using cascaded deep learning and random forests, in: 2015 international conference on digital image computing: techniques and applications (DICTA), IEEE. pp. 1–8.

Dubost, F., Bortsova, G., Adams, H., Ikram, A., Niessen, W.J., Vernooij, M., De Bruijne, M., 2017. Gp-unet: Lesion detection from weak labels with a 3d regression network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 214–221.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115.

Feng, X., Yang, J., Laine, A.F., Angelini, E.D., 2017. Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 568–576.

Gallego-Ortiz, C., Martel, A.L., 2015. Improving the accuracy of computer-aided diagnosis for breast mr imaging by differentiating between mass and nonmass lesions. Radiology 278, 679–688.

Gilbert, F., Selamoglu, A., 2018. Personalised screening: is this the way forward? Clinical radiology 73, 327–333.

Grimm, L.J., Anderson, A.L., Baker, J.A., Johnson, K.S., Walsh, R., Yoon, S.C., Ghate, S.V., 2015. Interobserver variability between breast imagers using the fifth edition of the bi-rads mri lexicon. American Journal of Roentgenology 204, 1120–1124.

Gubern-Mérida, A., Martí, R., Melendez, J., Hauth, J.L., Mann, R.M., Karssemeijer, N., Platel, B., 2015. Automated localization of breast cancer in dce-mri. Medical image analysis 20, 265–274.

Gubern-Mérida, A., Vreemann, S., Martí, R., Melendez, J., Lardenoije, S., Mann, R.M., Karssemeijer, N., Platel, B., 2016. Automated detection of breast cancer in false-negative screening mri studies from women at increased risk. European journal of radiology 85, 472–479.

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839–843.

Hayton, P., Brady, M., Tarassenko, L., Moore, N., 1997. Analysis of dynamic mr breast images using a model of contrast enhancement. Medical image analysis 1, 207–224.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. pp. 2980–2988.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q., 2016. Deep networks with stochastic depth, in: European Conference on Computer Vision, Springer. pp. 646–661.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kostopoulos, S.A., Vassiou, K.G., Lavdas, E.N., Cavouras, D.A., Kalatzis, I.K., Asvestas, P.A., Arvanitis, D.L., Fezoulidis, I.V., Glotsos, D.T., 2017. Computer-based automated estimation of breast vascularity and correlation with breast cancer in dce-mri images. Magnetic resonance imaging 35, 39–45.

Kousi, E., Borri, M., Dean, J., Panek, R., Scurr, E., Leach, M.O., Schmidt, M.A., 2015. Quality assurance in mri breast screening: comparing signal-to-noise ratio in dynamic contrast-enhanced imaging protocols. Physics in Medicine & Biology 61, 37.

Kriege, M., Brekelmans, C.T., Boetes, C., Besnard, P.E., Zonderland, H.M., Obdeijn, I.M., Manoliu, R.A., Kok, T., Peterse, H., Tilanus-Linthorst, M.M., et al., 2004. Efficacy of mri and mammography for breast-cancer screening in women with a familial or genetic predisposition. New England Journal of Medicine 351, 427–437.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Lee, K.A., Talati, N., Oudsema, R., Steinberger, S., Margolies, L.R., 2018. Bi-rads 3: current and future use of probably benign. Current radiology reports 6, 5.

Lehman, C.D., Blume, J.D., DeMartini, W.B., Hylton, N.M., Herman, B., Schnall, M.D., 2013. Accuracy and interpretation time of computer-aided detection among novice and experienced breast mri readers. American Journal of Roentgenology 200, W683–W689.

Levman, J.E., Causer, P., Warner, E., Martel, A.L., 2009. Effect of the enhancement threshold on the computer-aided detection of breast cancer using mri. Academic radiology 16, 1064–1069.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Fei-Fei, L., 2018. Thoracic disease identification and localization with limited supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8290–8299.

Liu, H., Zheng, Y., Liang, D., Tang, P., Ren, F., Zhang, L., Zhao, Z., 2017. Total variation based dce-mri decomposition by separating lesion from background for time-intensity curve estimation. Medical physics 44, 2321–2331.

Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G., 2018. Training medical image analysis systems like radiologists, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham. pp. 546–554.

Maicas, G., Carneiro, G., Bradley, A.P., 2017a. Globally optimal breast mass segmentation from dce-mri using deep semantic segmentation as shape prior, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 305–309.

Maicas, G., Carneiro, G., Bradley, A.P., Nascimento, J.C., Reid, I., 2017b. Deep reinforcement learning for active breast lesion detection from dce-mri, in: Medical Image Computing and Computer-Assisted Intervention MICCAI 2017, Springer International Publishing, Cham. pp. 665–673.

Maicas, G., Snaauw, G., Bradley, A.P., Reid, I., Carneiro, G., 2019. Model agnostic saliency for weakly supervised lesion detection from breast dce-mri, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1057–1060.

Mainiero, M.B., Moy, L., Baron, P., Didwania, A.D., Green, E.D., Heller, S.L., Holbrook, A.I., Lee, S.J., Lewin, A.A., Lourenco, A.P., et al., 2017. Acr appropriateness criteria® breast cancer screening. Journal of the American College of Radiology 14, S383–S390.

Mango, V.L., Morris, E.A., Dershaw, D.D., Abramson, A., Fry, C., Moskowitz, C.S., Hughes, M., Kaplan, J., Jochelson, M.S., 2015. Abbreviated protocol for breast mri: are multiple sequences needed for cancer detection? European journal of radiology 84, 65–70.

Matiisen, T., Oliver, A., Cohen, T., Schulman, J., 2017. Teacher-student curriculum learning. arXiv preprint arXiv:1707.00183 .

Mcclymont, D., 2015. Computer assisted detection and characterisation of breast cancer in MRI. Ph.D. thesis.

McClymont, D., Mehnert, A., Trakic, A., et al., 2014. Fully automatic lesion segmentation in breast MRI using mean-shift and graph-cuts on a region adjacency graph. JMRI 39, 795–804.

Meinel, L.A., Stolpen, A.H., Berbaum, K.S., Fajardo, L.L., Reinhardt, J.M., 2007. Breast mri lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (cad) system. Journal of magnetic resonance imaging 25, 89–95.

Milenković, J., Dalmış, M.U., Žgajnar, J., Platel, B., 2017. Textural analysis of early-phase spatiotemporal changes in contrast enhancement of breast lesions imaged with an ultrafast dce-mri protocol. Medical physics 44, 4652–4664.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al., 2015. Human-level control through deep reinforcement learning. Nature 518, 529.

Mus, R.D., Borelli, C., Bult, P., Weiland, E., Karssemeijer, N., Barentsz, J.O., Gubern-Mérida, A., Platel, B., Mann, R.M., 2017. Time to enhancement derived from ultrafast breast mri as a novel parameter to discriminate benign from malignant breast lesions. European journal of radiology 89, 90–96.

Platel, B., Mus, R., Welte, T., Karssemeijer, N., Mann, R., 2014. Automated characterization of breast lesions imaged with an ultrafast dce-mr protocol. IEEE transactions on medical imaging 33, 225–232.

Rasti, R., Teshnehlab, M., Phung, S.L., 2017. Breast cancer diagnosis in dce-mri using mixture ensemble of convolutional neural networks. Pattern Recognition 72, 381–390.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

Renz, D.M., Böttcher, J., Diekmann, F., Poellinger, A., Maurer, M.H., Pfeil, A., Streitparth, F., Collettini, F., Bick, U., Hamm, B., et al., 2012. Detection and classification of contrast-enhancing masses by a fully automatic

computer-assisted diagnosis system for breast mri. Journal of Magnetic Resonance Imaging 35, 1077–1088.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with deep learning. Scientific reports 8, 4165.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Saadatmand, S., Bretveld, R., Siesling, S., Tilanus-Linthorst, M.M., 2015. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. Bmj 351, h4901.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. pp. 618–626.

Shimauchi, A., Giger, M.L., Bhooshan, N., Lan, L., Pesce, L.L., Lee, J.K., Abe, H., Newstead, G.M., 2011. Evaluation of clinical breast mr imaging performed with prototype computer-aided diagnosis breast mr imaging workstation: reader study. Radiology 258, 696–704.

Siegel, R.L., Miller, K.D., Jemal, A., 2017. Cancer statistics, 2017. CA: A Cancer Journal for Clinicians .

Smith, R.A., Andrews, K.S., Brooks, D., Fedewa, S.A., Manassaram-Baptiste, D., Saslow, D., Brawley, O.W., Wender, R.C., 2017. Cancer screening in the united states, 2017: a review of current american cancer society guidelines and current issues in cancer screening. CA: a cancer journal for clinicians 67, 100–121.

Soares, F., Janela, F., Pereira, M., Seabra, J., Freire, M.M., 2013. 3d lacunarity in multifractal analysis of breast tumor lesions in dynamic contrast-enhanced magnetic resonance imaging. IEEE Transactions on Image Processing 22, 4422–4435.

Song, J.L., Chen, C., Yuan, J.P., Sun, S.R., 2016. Progress in the clinical detection of heterogeneity in breast cancer. Cancer medicine 5, 3475–3488.

Sutton, R., Barto, A.G., 1998. Reinforcement learning: An introduction. volume 2. MIT press.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A., 2015. Global cancer statistics, 2012. CA: a cancer journal for clinicians 65, 87–108.

Vreemann, S., Gubern-Merida, A., Lardenoije, S., Bult, P., Karssemeijer, N., Pinker, K., Mann, R.M., 2018. The frequency of missed breast cancers in women participating in a high-risk mri screening program. Breast Cancer Research and Treatment .

Wang, L., Harz, M., Boehler, T., Platel, B., Homeyer, A., Hahn, H.K., 2014. A robust and extendable framework towards fully automated diagnosis of nonmass lesions in breast dce-mri, in: International Symposium on Biomedical Imaging, IEEE. pp. 129–132.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017a. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097–2106.

Wang, Z., Yin, Y., Shi, J., Fang, W., Li, H., Wang, X., 2017b. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 267–275.

Welch, H.G., Prorok, P.C., OMalley, A.J., Kramer, B.S., 2016. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. New England Journal of Medicine 375, 1438–1447.

Wood, C., 2005. Computer aided detection (cad) for breast mri. Technology in cancer research & treatment 4, 49–53.

Xue, W., Brahm, G., Pandey, S., Leung, S., Li, S., 2018. Full left ventricle quantification via deep multitask relationships learning. Medical image analysis 43, 54–65.

Yang, X., Wang, Z., Liu, C., Le, H.M., Chen, J., Cheng, K.T.T., Wang, L., 2017. Joint detection and diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 426–434.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer. pp. 818–833.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.

Zhu, W., Lou, Q., Vang, Y.S., Xie, X., 2017. Deep multi-instance networks with sparse label assignment for whole mammogram classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 603–611.