WEAKLY-SUPERVISED DIAGNOSIS AND DETECTION OF BREAST CANCER USING DEEP MULTIPLE INSTANCE LEARNING

Pedro Diogo[†] *Margarida Morais*[†] *Francisco Maria Calisto*[†] *Carlos Santiago*[†]
Clara Aleluia[⋆] *Jacinto C. Nascimento*[†]

[†]Institute for Systems and Robotics, Instituto Superior Técnico, Portugal
[⋆]Hospital Prof. Doutor Fernando Fonseca, Imaging Services, Portugal

ABSTRACT

The detection and classification of breast cancer lesions with computer-aided diagnosis systems has seen a huge boost in recent years due to deep learning. However, most works focus on 2D image modalities like mammography and ultrasound. Dealing with 3D magnetic resonance imaging (MRI) data adds news challenges, such as data insufficiency and lack of local annotations provided by experts. To handle these issues, this work proposes an new two-stage framework based on deep multiple instance learning, which requires only global label (weak supervision) and provides: 1) classification predictions for the whole volume and for each slice; and 2) 3D localization of lesions, through the selection of consecutive slices and patches that most likely contain the lesion (heatmaps). Results show that the proposed approach achieves classification performances that are competitive with the state of the art, and a qualitative assessment of the heatmaps illustrates the effectiveness of this approach in finding the malignant lesion in the images.

***Index Terms*—** Breast Cancer,Magnetic Resonance Imaging, Multiple Instance Learning

1. INTRODUCTION

Breast Cancer (BC) is the most prevalent form of cancer in women and accounts for a significant number of deaths worldwide [1]. Although many countries have been able to reduce mortality rates through screening programs, radiologists are currently overwhelmed and their increasing workload is a major cause of burnout [1]. This fostered research on computer-aided detection and diagnosis systems (CADs), and several systems have already been developed specifically for breast cancer (BC) [2]. These systems are becoming increasingly popular as second readers, to help radiologists identify all lesions while simultaneously reducing their workload [3].

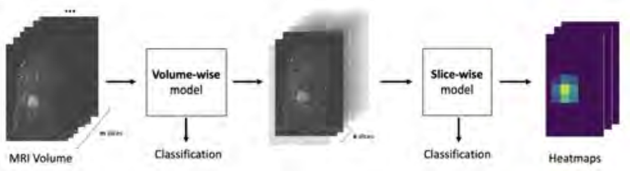

Fig. 1. Overview of the proposed Deep MIL framework.

However, most approaches are designed for mammography [4] and ultrasound [5]. By contrast, the body of work dedicated to the automatic analysis of magnetic resonance imaging (MRI) for BC is considerably smaller [6], despite being the recommended imaging modality in women with higher risk of developing breast cancer [7]. Two reasons that could explain this are: 1) the need for large datasets to properly train deep learning-based systems for 3D data; and 2) the lack of large annotated datasets, especially in regards to the location of lesions that justify the diagnosis, since manually annotating 3D data is substantially more challenging and demanding.

This work overcomes the above limitations by addressing the detection and diagnosis of breast cancer in a weakly-supervised scenario. The proposed system processes the whole MRI volume using a two-stage deep multiple instance learning (MIL) framework, depicted in Fig. 1, where the first stage considers the MRI volume as a collection of 2D images, and the second stage analyzes each image as a collection of patches. This framework is capable of identifying the 3D location of the lesions, by selecting the most relevant slices in a volume and then identifying the patches within each slice that justify the diagnosis. Furthermore, the proposed system does not process 3D data directly, thus avoiding the computational and data overhead.

2. BACKGROUND

One of the main challenges to the development of CAD systems for BC in MRI is the lack of data with annotations about the location of lesions. Previous works have either relied on an annotated dataset to train a region of interest detection module [8], or they require an experienced radiologist to manually provide potential malignant regions prior to clas-

[1] World Health Organization, “Breast cancer facts,” https://gco.iarc.fr/ today/data/factsheets/cancers/

sification [9, 10, 11]. But relying on this type of annotations prevents scaling the training process to larger datasets, due to the tremendous annotation effort they involve.

On the other hand, weakly supervised approaches have the benefit of simultaneously classifying the entire volume and identifying relevant regions that justify the decision, using only global labels which are comparably inexpensive. In particular, MIL is a weakly supervised learning strategy commonly used in binary problems that treats samples as a collection of instances, called bags, where the only label available is assigned to the entire bag, not to individuals instances [12]. It assumes that a bag is positive if at least one instance in that bag is positive, and negative otherwise. Previous works have used MIL approaches in BC classification problems, namely in mammography images [13, 14] and in ultrasound [15]. For instance, W. Zhu et al. [13] used a pooling function that involved ranking instances with the goal of performing end-to-end lesion classification for the whole mammogram. Since each image patch is given a malignancy score, they can detect lesions as a side product of their approach. Conversely, Sarath et al. [14] proposed a two-stage MIL framework where first a localization convolutional neural network (CNN) is trained to extract local candidate patches, and then a MIL strategy is employed to obtain a global image-level feature representation that is classified as benign or malignant. Nonetheless, these approaches rely on a fixed amount of patches (instances) to classify the image.

This work will aim to counter the above shortcoming by adaptively determining the number of instances needed to classify the whole MRI and by performing classification at two levels: volume-level and slice-level.

3. TWO-STAGE DEEP MIL FRAMEWORK

This work proposes a two-stage deep MIL framework to classify 3D MRI data. The framework, depicted in Fig. 1, comprises: 1) a volume-wise model that analysis the whole MRI volume and classifies it using only a subset of the slices; and 2) a slice-wise model that processes the slices selected by the previous model and classifies each of them using patches. The following sections describe each model in detail.

3.1. Volume-wise MIL Model

The volume-wise model is responsible for classifying the whole MRI volume (as malignant or not) and selecting the slices that contributed the most for that decision. To achieve this, a MIL setting is defined, where the model considers the MRI volume (bag) as a collection of several slices (instances).

Fig. 2 illustrates a scheme of this model, where the first step is to extract the most relevant features from each of the S slices (2D images) in the volume using a CNN. The feature vector, f_s, obtained for the s-th image then passes through a logistic regression that assigns it a malignancy probability given by

$$r_s = \sigma(w^\top f_s + b), \tag{1}$$

where w and b are learnable weights and biases, and $\sigma(\cdot)$ is the sigmoid function. Then, an adaptive top-k pooling module is applied, to select a subset of slices (shown in red in Fig. 2) that will contribute to the final volume-wise decision.

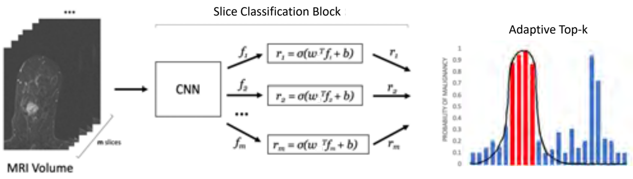
...
MRI Volume
m slices
Slice Classification Block
CNN
f_1
f_2
...
f_m
$r_1 = \sigma(w^T f_1 + b)$
$r_2 = \sigma(w^T f_2 + b)$
$r_m = \sigma(w^T f_m + b)$
r_1
r_2
r_m
Adaptive Top-k
PROBABILITY OF MALIGNANCY
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Fig. 2. Volume-wise MIL model.

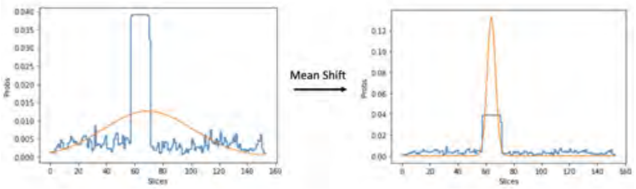

Fig. 3. Gaussian distribution estimate (orange) with and without Mean Shift for a specific set $\{r_1, \ldots, r_S\}$ (blue).

The adaptive top-k pooling module assumes that only a subset of consecutive slices may contain a malignant lesion. To identify this subset of slices, we apply Mean Shift [16] to find a robust Gaussian distribution fit to the probabilities $\{r_1, \ldots, r_S\}$. Specifically, we iteratively update the mean, μ, and standard deviation, σ, of the Gaussian distribution according to

$$\mu^{t+1} = \sum_{s=1}^{S} \omega_s^t \, \tilde{r}_s \, s \quad , \quad \sigma^{t+1} = \sqrt{\sum_{s=1}^{S} \omega_s^t \, \tilde{r}_s \, (s - \mu)^2}, \tag{2}$$

where $\tilde{r}_s = r_s / \sum_{j=1}^{S} r_j$ is the normalized malignancy probability, and ω_s^t is the Mean Shift weight given by

$$\omega_s^t = \frac{1}{K\sigma^t} \exp\left(-\frac{1}{2}\left(\frac{s - \mu^t}{\sigma^t}\right)^2\right), \tag{3}$$

where K is a normalization constant that ensures $\sum_{s=1}^{S} \omega_s^t = 1$. Assigning these weights to each slice is important to ensure that the estimation of the distribution is robust to spurious values, as illustrated in Fig. 3. After T iterations, the slices within $\left[\mu^T - \sigma^T, \mu^T + \sigma^T\right]$ are selected and sent to the slice-wise model, described in the following section. The final volume-wise classification is obtained as an average of the malignancy probabilities of the selected slices.

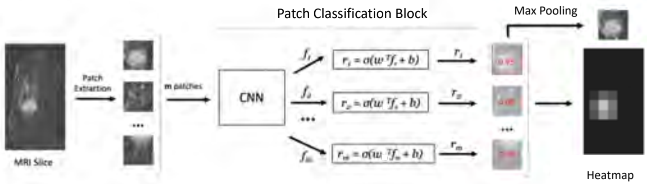
Patch Classification Block
Max Pooling
MRI Slice
Patch Extraction
m patches
CNN
f_1
f_2
f_m
$r_1 = \sigma(w^T f_1 + b)$
$r_2 = \sigma(w^T f_2 + b)$
$r_m = \sigma(w^T f_m + b)$
r_1
r_2
r_m
Heatmap

Fig. 4. Overview of the slice-wise MIL model.

3.2. Slice-wise MIL Model

The slice-wise model aims to classify and detect lesions within the slices chosen by the volume-wise model. Figure 4 shows an overview of the model. The first step is to divide an image into patches, which are then processed by a feature extraction CNN. Similarly to the volume-wise model, the feature vectors obtained for each patch go through a logistic regression, which converts the vectors into a malignancy probability, using specific weights, w, and bias, b, in (1). Finally, the classification of each slice is obtained using max pooling, following the MIL assumption. This means that if there is least one malignant patch, then the entire slice is considered malignant. Additionally, since the patch classification block is assigning a malignancy probability to each patch, a heat map can be computed based on those probabilities, as illustrated in Fig. 4 (right).

4. EXPERIMENTAL SETUP

The proposed framework was evaluated on a private dataset containing 164 MRI scans and corresponding diagnosis as malignant or not, provided by a senior radiologist. The MRI scans correspond to dynamic contrast enhancement data, which has previously been shown to be useful for BC diagnosis [17], since it removes high-intensity signal from background fat and improves lesion conspicuity [18].

Training and validation were performed using a subset of 134 MRI scans (71 malignant and 63 normal) randomly split in a 80%-20% ratio. Additionally, the slice-wise model was trained and evaluated using the prediction given by the slice classification block from the volume-wise model as GT. This resulted in 221 positive slices and 232 negative slices selected by the adaptive top-k pooling approach.

The remaining 30 MRI scans (18 malignant and 12 normal) were used as the test set to evaluate the performance of the volume-wise model. From these 30 volumes, the proposed framework selected 166 positive slices and 254 negative slices for the second stage.

The CNNs used in the proposed framework were MobileNetV2 [19] in both stages. The networks were trained for 50 epochs with binary cross-entropy, using Adam [20] with a learning rate of 1e-3. Due to the large size of the MRI data, the batch size was 4 for the first stage and 8 for the second.

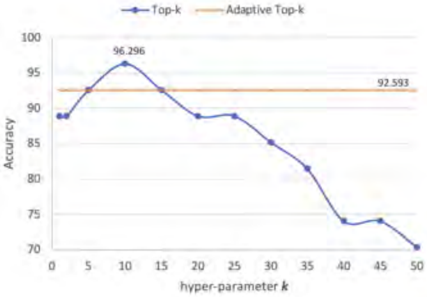

Fig. 5. Accuracy comparison between the proposed adaptive top-k pooling approach and the standard top-k.

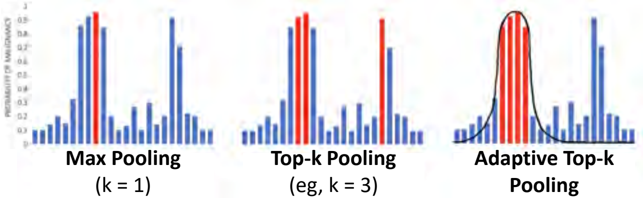

Fig. 6. Comparison between three pooling-based approaches.

5. RESULTS

5.1. Comparing Pooling Approaches

To validate the advantage of the proposed adaptive top-k pooling module, we compared it to the standard top-k pooling for different values of k, including $k = 1$ (max pooling). The results obtained for the validation set are shown in Fig. 5. The proposed adaptive top-k pooling achieves one of the best accuracies, only being outperformed by the top-10 approach. It is also clear that max pooling has much lower performance, and larger values of k also lead to a significant decrease in accuracy. This behavior was expected since, the malignant lesion only covers a small number of slices in the MRI volume, which means that additional slices would be misguiding the final prediction.

Furthermore, it is critical to note that, during training, the slices selected by the volume-wise model are then sent to the slice-wise model with the GT label given by the predicted volume-wise class. As such, sending a fixed number of slices (such as $k = 10$) may generate training label noise for the subsequent model. The proposed adaptive top-k mitigates this drawback by selecting a specific number of slices per volume.

Additionally, our approach ensures that the selected slices are contiguous, which is in accordance to the medical knowledge that the lesion is not spread along arbitrary slices in the MRI volume, but rather in a specific region comprised by consecutive slices. This cannot be guaranteed by the standard top-k, as shown in Fig. 6.

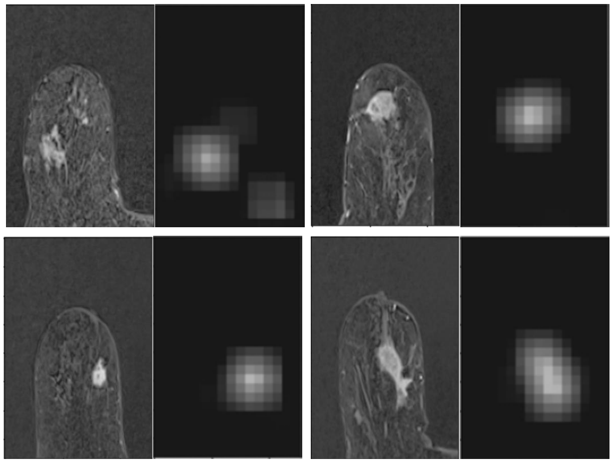

Fig. 7. Examples of images and corresponding heatmaps of detected lesions, according to the probability of malignancy of each patch in the image.

Model	Strategy	Acc	AUC	Sen	Spe	Prec
Volume-wise	Adaptive Top-k	96.67%	0.96	0.94	1.00	1.00
	Top-10	86.66%	0.91	0.78	1.00	1.00
Slice-wise	Using probs.	91.43%	0.98	0.82	0.98	0.96

Table 1. Classification results for the final versions of the models

	Supervision	AUC	Acc	Sen	Spe
[21]	Strong	0.91	-		
[22]	Strong	-	0.85	0.82	-
[23]	Weak	0.86	0.84	0.91	0.69
[24]	Weak	-	**0.98**	**0.96**	0.97
Ours	Weak	**0.96**	0.97	0.94	**1.00**

Table 2. Comparison with other state-of-the-art approaches for BC classification in MRI.

5.2. Lesion Localization Through Heatmaps

Lesion localizations were obtained by creating heatmaps from the malignancy probabilities predicted by the slice-wise model. Fig. 7 presents four malignant slices with their respective heat maps. While the heatmaps cannot be quantitatively assessed due to the lack of annotations, the examples in Fig. 7 clearly show that these heatmaps highlight the regions where the lesions are located. Therefore, even if the resolution of the heatmaps is not very high, they can provide radiologists with regions of interest where they should focus their attention during screening. These results also validate that a MIL approach is capable of learning which instances (patches) justify the decision of the model, even without guidance from GT annotations about the location of the lesions. Additionally, the proposed adaptive top-k pooling approach also provides a localization of the lesions along the third dimension of the MRI data, further helping radiologists improve their workflow when analyzing this type of data.

5.3. Comparison With State of the Art

The classification results of both models in the test set are expressed in Table 1. Comparing the results, the performance of the volume-wise model with proposed adaptive top-k approach achieved the best results (96.67%), outperforming both the top-10 approach and the slice-wise predictions. This further confirms that a pre-defined value for k is not guaranteed to lead to the best results across different datasets. On the other hand, it reinforces that adapting the number of selected slices to each volume is the most reliable strategy, as it allows finding an optimal number of consecutive slices for each case.

The proposed approach was also compared with the other state-of-the-art approaches. The results, shown in Table 2, demonstrate that the proposed approach achieves competitive results against the state of the art, including approaches that relied on GT annotations about the location of the lesions during training (strong supervision). Despite these results, it is important to note these works were trained and tested on different datasets, which prevents better benchmarking.

6. CONCLUSION

This work proposed a framework for the classification of BC in 3D MRI data. The framework consists of two deep MIL models trained in a weakly-supervised approach, which overcomes the need for annotations about the location of the lesions. The first model processes the entire MRI volume and adaptively selects a continuous amount of slices that correspond to the possible presence of the lesion, which it then uses to make the final prediction. Since some of the MRI volumes have more than one hundred slices, this accomplishment could be very helpful for radiologists as it excludes irrelevant slices within those volumes. The second model performs a slice-wise analysis that classifies and identifies relevant regions in the image where the lesion is present through heatmaps. Ours results show that the proposed approach outperforms other state-of-the-art approaches in classification accuracy, while simultaneously allowing the identification of regions within the MRI volume containing the lesion.

Acknowledgement: This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) project 2022.04485.PTDC and UIDB/50009/2020.

7. REFERENCES

[1] JR Parikh et al, “What causes the most stress in breast radiology practice? a survey of members of the society of breast imaging,” *Journal of breast imaging*, vol. 3, no. 3, pp. 332–342, 2021.

[2] NIR Yassin et al, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Computer methods and programs in biomedicine*, vol. 156, pp. 25–45, 2018.

[3] FM Calisto et al, "Breastscreening-ai: Evaluating medical intelligent agents for human-ai interactions," *Artificial Intelligence in Medicine*, vol. 127, pp. 102285, 2022.

[4] C Wang et al, "Knowledge distillation to ensemble global and interpretable prototype-based mammogram classification models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 14–24.

[5] J Kim et al, "Weakly-supervised deep learning for ultrasound diagnosis of breast cancer," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[6] R Dar et al, "Breast cancer detection using deep learning: datasets, methods, and challenges ahead," *Computers in Biology and Medicine*, p. 106073, 2022.

[7] N Cho et al, "Breast cancer screening with mammography plus ultrasonography or magnetic resonance imaging in women 50 years or younger at diagnosis and treated with breast conservation therapy," *JAMA Oncology*, vol. 3, no. 11, pp. 1495, Nov. 2017.

[8] R Rasti et al, "Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks," *Pattern Recognition*, vol. 72, pp. 381–390, Dec. 2017.

[9] A Gubern-Mérida et al, "Automated localization of breast cancer in DCE-MRI," *Medical Image Analysis*, vol. 20, no. 1, pp. 265–274, Feb. 2015.

[10] LA Meinel et al, "Breast MRI lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system," *Journal of Magnetic Resonance Imaging*, vol. 25, no. 1, pp. 89–95, Jan. 2007.

[11] SC Agner et al, "Segmentation and classification of triple negative breast cancers using DCE-MRI," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. June 2009, IEEE.

[12] TG Dietterich et al, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[13] W Zhu et al, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pp. 603–611. Springer International Publishing, 2017.

[14] CK Sarath et al, "A two-stage multiple instance learning framework for the detection of breast cancer in mammograms," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. July 2020, IEEE.

[15] J Ding et al, "Breast ultrasound image classification based on multiple-instance learning," *Journal of Digital Imaging*, vol. 25, no. 5, pp. 620–627, June 2012.

[16] D Comaniciu and P Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.

[17] Z Jun et al, "Breast tumor segmentation in dce-mri using fully convolutional networks with an application in radiogenomics," in *Medical Imaging 2018: Computer-Aided Diagnosis*. SPIE, 2018, vol. 10575, pp. 192–196.

[18] VS Lee, "Image subtraction in gadolinium-enhanced mr imaging.," *AJR. American journal of roentgenology*, vol. 167, no. 6, pp. 1427–1432, 1996.

[19] M Sandler et al, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[20] DP Kingma and J Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] N Antropova et al, "Performance comparison of deep learning and segmentation-based radiomic methods in the task of distinguishing benign and malignant breast lesions on dce-mri," in *Medical imaging 2017: Computer-aided diagnosis*. SPIE, 2017, vol. 10134, pp. 369–373.

[22] H Zheng et al, "Small lesion classification in dynamic contrast enhancement mri for breast cancer early detection," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 876–884.

[23] J Zhou et al, "Weakly supervised 3d deep learning for breast cancer classification and localization of the lesions in mr images," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 4, pp. 1144–1151, 2019.

[24] AM Ibraheem et al, "Automatic mri breast tumor detection using discrete wavelet transform and support vector machines," in *2019 Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. IEEE, 2019, vol. 1, pp. 88–91.

CLASSIFICATION OF BREAST CANCER IN MRI WITH MULTIMODAL FUSION

Margarida Morais[†] *Francisco Maria Calisto*[†] *Carlos Santiago*[†]
Clara Aleluia[⋆] *Jacinto C. Nascimento*[†]

[†]Institute for Systems and Robotics, Instituto Superior Técnico, Portugal
[⋆]Hospital Prof. Doutor Fernando Fonseca, Imaging Services, Portugal

ABSTRACT

Magnetic resonance imaging (MRI) is the recommended imaging modality in the diagnosis of breast cancer. However, each MRI scan provides dozens of volumes for the radiologist to inspect, each providing its own set of information on the tissues being scanned. This paper proposes a multimodal framework that processes all the available MRI data in order to reach a diagnosis, instead of relying on a single volume, mimicking the radiologists' workflow. The framework comprises a 3D convolutional neural network for each modality, whose predictions are then combined using a late fusion strategy based on Dempster-Shafer theory. Results highlight the most relevant modalities required to obtain accurate diagnosis, in agreement with clinical practice. They also show that combining multiple modalities leads to better overall results than their individual counterparts, achieving promising results against state of the art.

Index Terms— Breast Cancer, Magnetic Resonance Imaging, 3D Convolutional Neural Networks, Late Fusion

1. INTRODUCTION

Magnetic resonance imaging (MRI) is the recommended imaging modality when screening for breast cancer in women with higher-than-average risk of developing breast cancer [1, 2]. This is due to its high sensitivity rate, which allows for a better detection of lesions and their diagnosis, especially on patients with high breast density [3, 4].

Radiologists typically analyze multiple MRI volumes, obtained with different acquisition sequences, that highlight specific features and tissues in the breast. For instance, Dynamic Contrast Enhanced (DCE), Dynamic Contrast Enhanced Subtraction (DCEsub), T1 weighted (T1), T2 weighted (T2), and T2 fat saturated (T2fatsat) are among the most commonly used to diagnose breast cancer. But analyzing all this (3D) information is tremendously demanding and requires an experienced radiologist or even a second reader's opinion [5, 6].

Numerous computer-aided diagnosis (CAD) systems have been developed to assist medical practitioners with image interpretation [7]. These systems can decrease the number of errors made by radiologists, as shown in recent studies [8]. However, few works have addressed the classification of breast cancer in MRI, which requires dealing with 3D data. Instead, most approaches focus on alternative 2D imaging modalities, such as mammography [9].

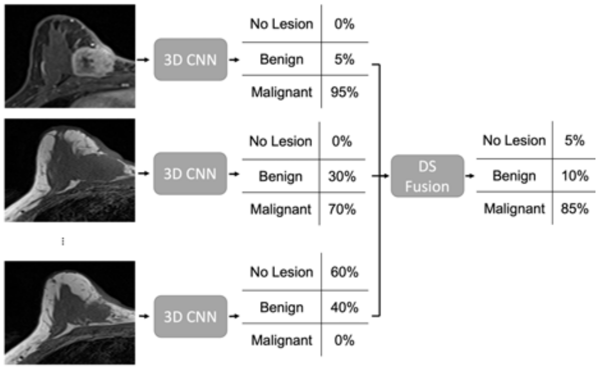

Fig. 1. Overview of the proposed multimodal fusion.

In this work, we propose a CAD system, illustrated in Fig. 1, that analyzes multiple MRI volumes and provides a diagnosis by combining all the available 3D information, mimicking the radiologists' workflow. The system relies on an ensemble of 3D convolutional neural networks (CNNs), and the final classification is given by a late-fusion strategy based on Dempster-Shafer's theory (DST) [10]. We show that this approach leads to promising results, which are competitive with other state-of-the-art works that require prior additional information about the location of lesions. Additionally, our best performing individual classifiers correspond to the MRI volumes that radiologists often analyze more thoroughly, thus suggesting that our framework automatically identifies which volumes are the most relevant for breast cancer diagnosis.

2. PROPOSED FRAMEWORK

This section provides an overview of the proposal (Sec. 2.1), describing how the training is performed (Sec. 2.2), followed by the fusion (Sec. 2.3) as a way to combine the different sources of information (*i.e.* different modality MRI volumes).

2.1. Overview

In this work we propose to leverage the 3D nature of MRI, by using a 3D CNN to extract volumetric features from the data. In our proposal, the left and right breasts volumes are obtained with an initial pre-processing that allows to obtain two volumes from the entire MRI volume. This allows us to classify each breast individually. Then, a model is trained specifically for each type of volume on a multi-class classification task. A description of the training procedure is given in Sec. 2.2. Each of these models then assigns a probability to each class and performs breast cancer diagnosis based on a single input modality, as shown in Fig. 1 (left).

This single-modal approach offers only a very limited view of the entire available data. Therefore, the final diagnosis for the patient is obtained by combining the individual results using a late fusion approach, based on the Dempster-Shafer's theory of evidence [10]. This second step is shown in Fig. 1 (right) and discussed in Section 2.3.

2.2. Model Training

Each model is trained to classify volumes obtained with a specific MRI sequence (*e.g.*, T1, T2, etc). Three classes are considered: no lesions (normal), benign, and malignant. Since large datasets with all this information are scarce and classes are often imbalanced, instead of using the conventional cross entropy loss for the multi-class problem, we propose to combine it with the sample-weighting approach LOW [11]. This loss function overcomes the drawbacks of small and imbalanced datasets, especially when class distribution is long-tailed.

LOW estimates the weight of each training sample in each step of gradient descent, in order to determine its contribution in the training process. By doing this, the model focuses on different samples during training, preventing it from overfitting the predominant classes. Specifically, given the predicted probability of the correct class, $\widehat{y}_j$, for each sample $j = 1, \ldots, M$ in a batch, the final loss function is given by

$$L_{\text{LOW}} = -\frac{1}{M}\sum_{j=1}^{M} w_j \log \widehat{y}_j \quad , \tag{1}$$

where w_j is the weight assigned to the j-th sample. These weights are obtained by maximizing the norm of the gradient of the cross-entropy (CE) loss, under the constraint that weights should be close to 1 (standard CE) and that they must be positive and have an average value of 1. This leads to the following optimization problem:

$$\begin{aligned} \arg\max_{\mathbf{w}} \quad & \mathbf{w}^\top \boldsymbol{\nabla} - \lambda \left\|\mathbf{w} - \mathbf{1}\right\|^2 \\ \text{subject to} \quad & \mathbf{w} \geq 0 \\ & \mathbf{w}^\top \mathbf{1} = M \end{aligned} \tag{2}$$

where

$$\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_M \end{bmatrix} \qquad \boldsymbol{\nabla} = \begin{bmatrix} \left\|\nabla \ell_{\text{CE}}\left(\widehat{y}_1\right)\right\|^2 \\ \vdots \\ \left\|\nabla \ell_{\text{CE}}\left(\widehat{y}_M\right)\right\|^2 \end{bmatrix} \tag{3}$$

and λ is hyperparameter (see [11] for more details).

2.3. Model Fusion

For the purpose of combining the information from the different models, a late-fusion strategy based on the DST was used. DST allows the combination of different classification predictions, while taking into account the uncertainty associated with each classifier.

To measure the uncertainty, we compute the positive predictive values (PPV) of each class in a validation set and multiplying it by the probability for that class. The remainder of the value is attributed to the unknown decision. Formally, let the predicted class probabilities given by the i-th classifier be $[p^i_1, \ldots, p^i_C]$, where C is the number of classes. Denoting the PPV of class c and classifier i as U^i_c, the predictions with uncertainty are given by $\tilde{p}^i_c = U^i_c p^i_c$, where $1 - \sum_{c=1}^{C} \tilde{p}^i_c$ is the probability of an unknown decision. Then, DST computes the final prediction, $\hat{p}_c$, from N classifiers according to the combination rule

$$\hat{p}_c = \frac{1}{1-K}\prod_{i=1}^{N} \tilde{p}^i_c \quad , \tag{4}$$

where K is a normalization coefficient that represents the conflict factor between the different classifiers. For illustration purposes, in a simple case with $N = 2$, K would be

$$K = \sum_{c=1}^{C}\sum_{k\neq c} \tilde{p}^1_c \tilde{p}^2_k \tag{5}$$

For further details, see [12].

3. EXPERIMENTAL SETUP

In this work, we use a private breast cancer dataset, containing the MRI scans from patients of Hospital Fernando Fonseca, Portugal. The dataset contains the MRI scans of 124 patients, with an average age of 61.3, in a total of 620 volumes from 5 different acquisition sequences: DCE2 (second instant post-contrast), DCE2sub (second instant post-contrast), T1, T2, and T2 fatsat.

The dataset was preprocessed to remove background regions and to separate each volume in two, one containing the left breast and the other containing the right breast (248 individual breasts). For each case, a senior radiologist provided the ground truth label for each breast according to the classes: 'No Lesion', 'Benign Lesion', and 'Malign Lesion', with the class composition of 93, 39, and 116 cases, respectively. The available data was then divided into train, validation, and test sets. The train and validation sets were then used to train multiple classifiers and hyperparameter tuning with 5-fold cross-validation.

Each model was trained with one of the available MRI volumes as to obtain the performance of the classifier when making a diagnosis with a single modality.

Each classifier is based on a 3D CNN that was trained with the Adam optimizer [13] ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for 100 epochs with a batch size of 30. The initial learning rate was set to 10^{-3}, with decay to 10^{-4} after 50% of the epochs, and to 10^{-5} for 75% of the epochs.

For the purpose of evaluating the performance of the classifiers, we use the following statistical metrics: area under the curve (AUC), balanced accuracy, sensitivity, specificity and precision.

4. RESULTS

This section shows several experiments validating our proposed approach. Three ablation studies were performed: 1) to justify our choice of loss function; 2) to compare the proposed 3D CNN against a state-of-art 3D architecture based on Resnet; and 3) to compare the performance of the proposed DST-based fusion module against the standard fusion approach. Additionally, we also analyze the performance of the individual classifiers and different combinations of data for the fusion approach. Finally, we compare our approach against other state-of-the-art MRI classification systems.

4.1. Loss Function

We evaluated the impact of the LOW loss in the proposed framework by comparing the performance of the DCE2sub classifier using different loss functions: 1) the standard cross entropy (CE) used in multiclass problems; 2) weighted CE, where each class weight is inversely proportional to the class frequency in the dataset (*i.e.*, more prevalent classes have a lower weight); and 3) LOW, which assigns specific weights to each sample, as described in Section 2.2.

Fig. 2 shows the performance of the model trained with each loss function using three different metrics. The results shows that the standard CE leads to a model with significantly lower performance compared to both the weighted CE and LOW. Additionally, training the model with LOW leads to the best performances across all three metrics. This shows that LOW is capable of dealing with the class imbalance in the training set and forces the model to classify all classes correctly, and not just the predominant classes.

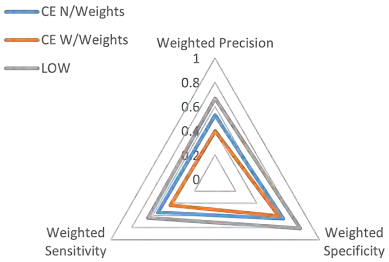

Fig. 2. Performance of the 3D CNN classifier trained with three different loss functions: CE, weighted CE, and LOW.

Model	BalAcc	AUC	Spe	Sen	Pre
3D Resnet	34.2 (1.3)	55.3 (2.8)	54.9 (3.2)	47.3 (0.3)	36.2 (13.9)
Ours	49.2 (3.6)	73.2 (3.8)	73.5 (2.3)	58.6 (2.9)	55.7 (5.3)

Table 1. Performance of the proposed 3D CNN architecture against 3D Resnet using 5-fold cross-validation.

4.2. Model Architecture

To find suitable classifiers, several 3D CNN architectures were evaluated using 5-fold cross validation on the DCE2sub data. The main limitation was the amount of volumes in the training set in relation to the high dimensionality of the data. Consequently, most larger models suffered from severe overfitting. To illustrate one example, we compare the average performance across all five folds between the final 3D CNN model and a state-of-the-art 3D Resnet [14]. The results, shown in Table 1, demonstrate a huge performance decrease when using 3D Resnet as the architecture for the classifier. An analysis of the loss and accuracy curves during training showed that the larger model was underperforming in the validation sets, indicating that more samples would be required to achieve generalization.

4.3. Individual Classifier Analysis

Comparing the performance of the individual classifiers, we see that, from Tab. 2, the best balanced accuracy was obtained using the DCE2sub volume. These results are consistent with our initial expectations, since it corresponds to acquisition sequence that radiologists most rely on, when searching for malignant breast cancer lesions in their clinical workflow [15]. It is also interesting to note that the classifiers trained with the other volumes achieve very similar and considerably lower performances, suggesting that they do not provide sufficient information on their own. However, radiologists analyze several volumes to reach their decision. Therefore, combining their information in a fusion strategy is critical to have a more reliable diagnosis.

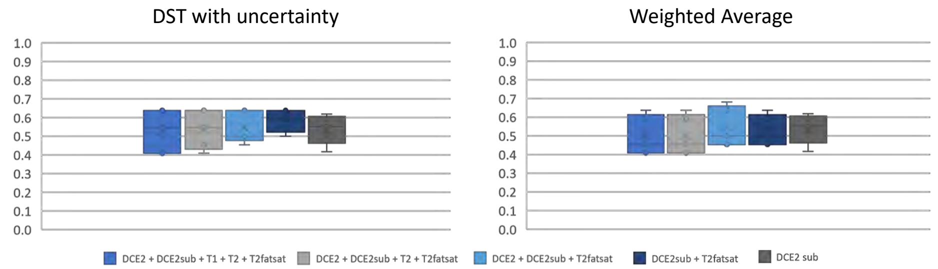

Fig. 3. Comparison between the weighted accuracy for DST with (left) and without (right) uncertainty on DCE2sub.

	DCE2	DCE2sub	T1	T2	T2fatsat
BalAcc	36 (6)	**54 (7)**	39 (3)	36 (4)	38 (3)

Table 2. Comparison of the individual classifier performances using the balanced accuracy metric.

4.4. Fusion Strategies

Two different approaches were compared for the late-fusion process: 1) a standard fusion approach that computes a weighted average of the predicted class probabilities, where the contribution of each individual classifier is proportional to its performance in the validation set; and 2) using the DST, which estimates the uncertainty associated with each class prediction and for each model, uses this information to combine the outputs into a final decision.

Additionally, we assessed the impact of each additional volume type to the final performance by evaluating several combinations of multimodal data. Specifically, we started by combining all the available modalities, and gradually removed the worst performing individual classifier (Table 2).

The results are shown in Fig. 3, where the average balanced accuracy using the proposed DST-based approach is plotted on the left, and the traditional fusion approach on the right. The plots show that the proposed approach better overall performances across all combinations, thus validating that taking the uncertainty of the models into account leads to better results. It is also interesting to note that the combination that leads to higher balanced accuracy and lower variation is DCE2sub and T2fatsat, with an accuracy of 58.2($\pm$5)%. This might suggest that these two volumes contain complementary information that help identify the necessary features to distinguish between benign and malignant lesions, or simply better identify the presence of more challenging lesions.

4.5. Comparison with State of the Art

Table 3 compares the best results obtained using the proposed approach with other state-of-the-art works. Due to the lack of benchmarks with multiple MRI volumes per patient, a rigorous comparison with the state of the art is not possible. In fact, the dataset used in this work is substantially smaller than the datasets used by other works in the literature. Nonetheless, the AUC obtained with our framework achieves promising results, especially when compared to other works with similar-sized datasets like [16]. Additionally, our work does not rely on bounding box annotations, which are often used by other works to limit the size of the volume that has to be processed [17]. Finally, most state-of-the-art works only address the problem of distinguishing between benign and malignant lesions. This subproblem is significantly easier since it does not required dealing with the often challenging task of discriminating between benign lesions and no lesions at all.

Reference	**Dataset Size**			**AUC**
	N	B	M	
Dalmiş et al. [16]	-	208	368	85.2%
Zhou et al.[18]	-	506	1031	85.9%
Li et al. [17]	-	66	77	80.1%
Our work: 3D CNN	85	34	107	77.8%

Table 3. Test set results obtained with the proposed approach and comparison with the state of the art. N is “No Lesion”, B is “Benign Lesion” and M is “Malignant Lesion”.

5. CONCLUSION

This work presented a new framework to classify MRI for breast diagnosis. The framework is base on a combination of multiple 3D CNN classifiers, each targeting a specific acquisition sequence, which are then combined using a late fusion approach based on Dempster-Shafer theory. This framework allows the information from the multiple modalities available to be jointly used to obtain the final decision. Results showed that the proposed approach is competitive with other state-of-the-art works. By combining the diagnosis from different input volumes, we obtain an increase in the performance of the final decision, compared to individual predictions. Additionally, an analysis of the performance of the individual classifiers highlights a hierarchy of relevant MRI sequences to the diagnosis, in accordance with the feedback from radiologists.

Acknowledgement: This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) project 2022.04485.PTDC and UIDB/50009/2020.

6. REFERENCES

[1] Nariya Cho et al, “Breast cancer screening with mammography plus ultrasonography or magnetic resonance imaging in women 50 years or younger at diagnosis and treated with breast conservation therapy,” *JAMA Oncology*, vol. 3, no. 11, pp. 1495, Nov. 2017.

[2] Christiane K. Kuhl, Kevin Strobel, Heribert Bieling, Claudia Leutner, Hans H. Schild, and Simone Schrading, “Supplemental breast MR imaging screening of women with average risk of breast cancer,” *Radiology*, vol. 283, no. 2, pp. 361–370, May 2017.

[3] Dorothy A Sippo, Kristine S Burk, Sarah F Mercaldo, Geoffrey M Rutledge, Christine Edmonds, Zoe Guan, Kevin S Hughes, and Constance D Lehman, “Performance of screening breast MRI across women with different elevated breast cancer risk indications,” *Radiology*, vol. 292, no. 1, pp. 51–59, July 2019.

[4] Debra L Monticciolo, Mary S Newell, Linda Moy, Bethany Niell, Barbara Monsees, and Edward A Sickles, “Breast cancer screening in women at higher-than-average risk: Recommendations from the ACR,” *J. Am. Coll. Radiol.*, vol. 15, no. 3, pp. 408–414, Mar. 2018.

[5] Emmanuelle Bouic Pages, Ingrid Millet, Denis Hoa, Fernanda Curros Doyon, and Patrice Taourel, “Undiagnosed breast cancer at MR imaging: Analysis of causes,” *Radiology*, vol. 264, no. 1, pp. 40–50, July 2012.

[6] N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, and L. von Karsa, “European guidelines for quality assurance in breast cancer screening and diagnosis. fourth edition—summary document,” *Annals of Oncology*, vol. 19, no. 4, pp. 614–622, Apr. 2008.

[7] Nisreen IR Yassin, Shaimaa Omran, Enas MF El Houby, and Hemat Allam, “Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review,” *Computer methods and programs in biomedicine*, vol. 156, pp. 25–45, 2018.

[8] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento, “Breastscreening-ai: Evaluating medical intelligent agents for human-ai interactions,” *Artificial Intelligence in Medicine*, vol. 127, pp. 102285, 2022.

[9] Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Al-garadi, Fariha Zulfiqar, Ghulam Raza, and Nor Aniza Azmi, “Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges,” *Artificial Intelligence Review*, vol. 53, no. 3, pp. 1655–1720, May 2019.

[10] Qi Chen, Amanda Whitbrook, Uwe Aickelin, and Chris Roadknight, “Data classification using the dempster–shafer method,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 26, no. 4, pp. 493–517, 2014.

[11] Carlos Santiago, Catarina Barata, Michele Sasdelli, Gustavo Carneiro, and Jacinto C Nascimento, “LOW: Training deep neural networks by learning optimal sample weights,” *Pattern Recognition.*, vol. 110, no. 107585, pp. 107585, Feb. 2021.

[12] Yuexiang Yang, Xing Pan, and Qingde Cui, “An evidence combination rule based on transferable belief model and application in reliability assessment with multi-source data,” *IEEE Access*, vol. 8, pp. 69096–69104, 2020.

[13] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[14] Nader Aldoj, Steffen Lukas, Marc Dewey, and Tobias Penzkofer, “Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network,” *Eur. Radiol.*, vol. 30, no. 2, pp. 1243–1253, Feb. 2020.

[15] V S Lee, M A Flyer, J C Weinreb, G A Krinsky, and N M Rofsky, “Image subtraction in gadolinium-enhanced MR imaging,” *AJR Am. J. Roentgenol.*, vol. 167, no. 6, pp. 1427–1432, Dec. 1996.

[16] Mehmet U. Dalmiş, Albert Gubern-Mérida, Suzan Vreemann, Peter Bult, Nico Karssemeijer, Ritse Mann, and Jonas Teuwen, “Artificial intelligence–based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, t2, and DWI,” *Investigative Radiology*, vol. 54, no. 6, pp. 325–332, June 2019.

[17] Jing Li, Ming Fan, Juan Zhang, and Lihua Li, “Discriminating between benign and malignant breast tumors using 3D convolutional neural network in dynamic contrast enhanced-MR images,” in *Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*, Tessa S Cook and Jianguo Zhang, Eds. Mar. 2017, SPIE.

[18] Juan Zhou, Lu-Yang Luo, Qi Dou, Hao Chen, Cheng Chen, Gong-Jie Li, Ze-Fei Jiang, and Pheng-Ann Heng, “Weakly supervised 3d deep learning for breast cancer classification and localization of the lesions in MR images,” *Journal of Magnetic Resonance Imaging*, vol. 50, no. 4, pp. 1144–1151, Mar. 2019.