Diagnosis and Detection of Breast Cancer using Deep Multiple Instance Learning

Pedro Diogo
pedro.m.diogo@tecnico.ulisboa.pt
Instituto Superior Técnico
Lisboa, Portugal

Supervisor: Carlos Santiago
Supervisor: Jacinto Nascimento

ABSTRACT

The detection and classification of breast lesions in the early stages of its development may increase patients' chance of survival as well as the number of effective treatment options. With the intent of improving the radiologists' workflow in their effectiveness and efficiency, Computer-Aided Diagnosis or Detection systems have been emerging alongside with Deep Learning. Challenges such as data insufficiency and lack of local annotations provided by experts are the main practical issues when applying these systems in medical imaging. To handle these issues, this work proposes an autonomous system that takes advantage of deep convolutional features for image analysis and the Multiple Instance Learning framework for labeling a set of slices within volumes and/or a set of patches within slices. The ultimate goal is to achieve classification based on the whole MRI and based on the slices, where the former will permit to assess the slices that triggered the classification, and the latter will make possible the visual explanation of the proposed diagnosis through the localization of the lesion in the image.

KEYWORDS

Breast Cancer, Medical Imaging, Convolutional Neural Networks, Multiple Instance Learning (MIL), Magnetic Resonance Imaging (MRI)

1 INTRODUCTION

In agreement with the World Health Organization (WHO), by the end of 2020, Breast Cancer (BC) was consider the world's most predominant cancer since there were 7.8 million women alive that contracted this type of cancer in the past 5 years. Nonetheless, early detection of BC can significantly improve the outcomes of its treatment, reducing the mortality related to it [15].

Mammography screening has been confirmed as the most effective method to produce significant reductions in mortality rate of BC in women [6]. Nonetheless, Magnetic Resonance Imaging (MRI) has been providing better results in women with dense breast tissue [2]. For instance, some studies have shown that MRI is recommended along with a yearly mammography for some women with high risk for BC [13, 20] mainly due to his high sensitivity [19]. However, a larger sensitivity could also reveal things that turn out not to be cancer (false positive findings), leading to unnecessary biopsies which not only cause patient anxiety and morbidity, but also increase the money spent on health-care. Therefore, to avoid this situation, improvements in screening and discovering other ways to complement the reviews of the radiologists are truly important.

One way to meet this challenge is through Computer-Aided Detection or Diagnosis (CAD) systems since, nowadays, they have been considered as a second clinical opinion, improving the radiologist performance when used in the right way, not to decide but to counsel[9, 10]. At the same time, some studies have shown that CAD systems increase the risk of false positives [7], and this is why they cannot and should not replace a complete evaluation by the radiologist. The problem is, since it is standard for MRI screening to take several volumes for each patient, the accumulation of radiologists' scans increases and so does the complexity of their interpretation. Consequently, this can lead to a decrease of performance due to their exhaustion/fatigue.

Recent enhancements in Deep Learning (DL) methodologies have demonstrated revolutionary changes in radiology, making artificial intelligence and human-computer interaction advance with big strides, especially with the usage of Convolution Neural Networks (CNNs) [3, 11, 21]. Moreover, researchers found that the combination of expert radiologists and CAD systems outperform both individual performance [22]. However, despite the growth in DL, these models are dependent on massive sets of hand-labeled training data. These hand-labeled training sets are expensive and time-consuming to create, especially when domain expertise is required. However, deep architectures with a weak label approach can move past the constraint of data unavailability [23]. That said, it is of great importance to achieve performant CAD models through weak label classification as it could have a positive impact on future employments in medical facilities and DL research.

This work will be focusing in CAD systems applied to the MRI screening modality, aiming to differentiate malignant from not malignant lesions in BC. By means of a weakly supervised learning approach, it will be possible to obtain Volume-wise classification, extract the slices in which the lesion was found and, finally, detect the lesions within the slices chosen.

2 OBJECTIVES

The purpose of this work is the development of an autonomous system capable of providing the diagnosis, the approximate slices containing the malignant lesion and, within each slice selected, the region of the breast where the malignancy is found. This system was designed taking into account the large number of slices within an MRI volume and how consuming their examination can be for the radiologists.

To accomplish the defined aim of this research, the following objectives are established:

(1) Implement a model that predicts whether an MRI scan is malignant or not; Additionally, the model should output the MRI slices where the lesion is most noticeable.
(2) Implement a model that, given an MRI slice, classifies and localizes the lesion within that slice.

The classification task and the regions of the image that would justify this classification are going to be achieved through a Multiple Instance Learning (MIL) architecture. The output will be distinguished between two classes: malignant or not malignant.

3 MULTIPLE INSTANCE LEARNING

MIL is proposed as a weakly supervised learning strategy that deals with collections of instances arranged in sets, called bags, where there's only a label assigned for the entire bag instead of individuals labels for each instance. In computer vision problems, these bags are usually treated as images and the instances as patches. The MIL assumption corresponds to the typical binary problem in which a bag is positive if at least one instance in that bag is positive, and the bag is negative if all the instances are negative. Let Y be the single binary label of a bag X, defined as a set of instances, $X = (x_1, x_2, ..., x_N)$, where N is not necessarily equal among different bags. Each instance x_n corresponds to a label y_n, that remains unknown during the training phase. Finally, the label of the bag Y can be summarized as follow:

$$Y = \begin{cases} 1 & \text{, if } \exists y_n : y_n = 1, \\ 0 & \text{, otherwise.} \end{cases} \tag{1}$$

Or even in a more compact way:

$$Y = \max_n \{y_n\}. \tag{2}$$

4 STATE OF THE ART

Early detection of BC can significantly improve the outcomes of its treatment, reducing the mortality related to it [15]. Since different imaging modalities provide complementary information regarding lesions, it is important that the workflow for radiologists involves the analysis of these modalities, such as mammography, Ultrasound (US) and MRI. Although the combination of these modalities may increase the accuracy of the diagnostic, this can overwhelm radiologists. Therefore, several CAD systems using different breast imaging techniques have been developed for the detection and diagnosis of breast masses. However, CAD systems for BC related to MRI are still limited. In general, the existing approaches usually address the problem by a three-stage system: (i) identification of possible malignant Region of Interest (ROI) by a candidate generator, (ii) computation of descriptive features for each candidate, and (iii) labeling of each candidate (e.g., as benign or malignant) by a classifier. The main problem of these systems is, before the classification procedure, they either rely on manually malignant regions annotated by experienced radiologists [1, 8, 14] or they build an algorithm just for the ROI detection and selection [16]. Thus, if only global labels were attributed for the whole image, they could not indicate which parts of images induced the automatic diagnosis neither highlight abnormal regions in the image whenever an abnormal examination instance is detected.

In order to identify regions of the image that justify the ground truth label, MIL was proposed and approaches around it have been explored to extract features from patches obtained from the entire image without the need of lesion segmentation. MIL has been used in BC, specially in mammography images, although a few studies have already explored their potential in Ultrasound [5]. Due to the emergence of deep features, some studies have been combining MIL with deep neural networks. For instance, W. Zhu et al. [24] used a pooling function that involved ranking instances with the goal of performing end-to-end mass classification for the whole mammogram. In their approach, since each spatial location is a single instance associated with a score that is correlated with the existence of a malignant finding, they do not need an automated lesion detection stage, even though they can detect lesions as a side effect of their approach. Conversely, Sarath et al. [18] proposed a two-stage MIL framework where a localization network (CNN) is trained in the first stage to extract local candidate patches in the mammograms and, in the second stage, a MIL strategy is employed to obtain a global image-level feature representation from the extracted image patches to classify the mammograms as benign or malignant. Note that the purpose of the localization network in the first stage is not to get an accurate semantic segmentation but to obtain an approximate localization of the masses in terms of bounding boxes so that the second stage does not have to deal with irrelevant patches from the entire image.

Despite the advantages above-mentioned related to MIL-based CNNs, these approaches have limitations since they (1) rely on a fixed amount of patches (instances) to assign a classification to the whole image and (2) they do not explore the potentiality of overlapped patches. With that being said, and given the scarcity of MIL studies applied to the MRI modality in breast cancer, this work will aim to counter the shortcomings mentioned by adaptively learning the number of instances needed to classify the whole MRI and by performing classification at two levels: volume-level and slice-level. This first part is specially important in order to avoid misclassification of some instances.

5 PROPOSAL APPROACH

This work proposes and combines MIL with Deep Learning in order to achieve classification, slice-selection and patch-selection. To accomplish such a system, two different models have to be considered: while the first one will classify the MRI volume as a whole and extract the slices that triggered the classification, the second one will be fed with those slices and perform classification in each slice and extract the patches that triggered the classification. Therefore, the first model will be called from now on Volume-wise model, and the second one Slice-wise model. Nonetheless, both models share similarities: (1) the reliance on the MIL approach and (2) the way of extracting deep features from the images. Figure 1 illustrates the overview scheme of the system.

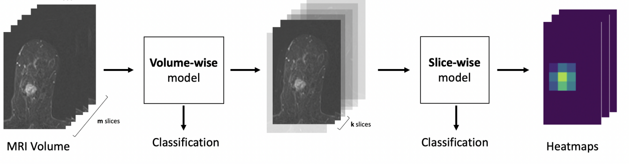


Figure 1: Deep MIL system overview

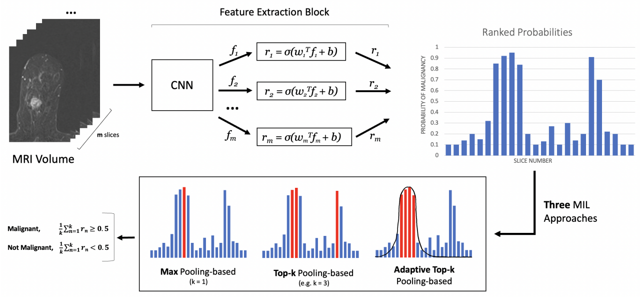

Feature Extraction Block
MRI Volume
m slices
CNN
f_1
f_2
f_m
$r_1 = \sigma(w_1^T f_1 + b)$
$r_2 = \sigma(w_2^T f_2 + b)$
$r_m = \sigma(w_m^T f_m + b)$
r_1
r_2
r_m
Ranked Probabilities
PROBABILITY OF MALIGNANCY
SLICE NUMBER
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Three MIL Approaches
Max Pooling-based (k = 1)
Top-k Pooling-based (e.g. k = 3)
Adaptive Top-k Pooling-based
Malignant, $\frac{1}{k}\sum_{n=1}^{k} r_n \geq 0.5$
Not Malignant, $\frac{1}{k}\sum_{n=1}^{k} r_n < 0.5$

Figure 2: Volume-wise model overview

5.1 Volume-wise classification

As referred before, the Volume-wise model performs **classification** and **slice-selection** in MRI volumes. In other words, this model diagnoses MRI volumes and selects the slices that contributed the most for that diagnosis. For that purpose, this model is based on the assumption that a lesion in an MRI volume typically remains (approximately) in the same spatial localization during a few continuous slices. Consequently, exploring and comparing different manners of selecting those slices will be the main focus of this model.

In terms of the MIL parameters, this model defines the whole MRI volume as the bag and the slices within that volume as the instances. Following the overview scheme present in Figure 2, the first step is to extract the most relevant features from the slices in the volumes. For a given volume B containing a set of slices $(I_1, I_2, ..., I_m)$, where m is the number of slices inside that volume, through the usage of a CNN, it is possible to acquire features for all those slices. Thus, after multiple convolutional layers and max pooling layers, a feature map f_i that represents deep CNN features can be obtained for each I_i. Then, since the goal of this work is to predict whether or not a slice contains a malignant mass, this is a typical standard binary classification problem. Therefore, a logistic regression can be used for classification with the weights shared across all values of f with a sigmoid activation function, whose output represents the probability of a slice being malignant. Formally, the malignant probability of a slice I_i can be given by:

$$r_i = \sigma(w^\top f_i + b) \tag{3}$$

where w corresponds to the weights in the logistic regression and b is the bias. From the combination of all r_i, a general $\boldsymbol{r}$ can be defined as a one-dimensional vector, $\boldsymbol{r} = (r_1, r_2, ..., r_m)$, corresponding to all slices in a volume B.

Once the malignant probabilities are obtained, three different MIL approaches to combine multiple instances (slices, in this case) can be explored: (1) the Max pooling-based MIL that only takes the largest element from the ranking layer; (2) the Top-k pooling-based MIL, which consists on grabbing the first k largest probabilities; and (3) the Adaptive Top-k Pooling-based MIL that adaptively selects the optimal number of slices for classification.

- **Max Pooling-based MIL:** Considering the general MIL assumption defined in Section 3, if each image (a slice or a patch, depending on the model) I_i of B is treated as an instance, the whole image classification problem can be seen as a standard multiple instance task. Hence, positive bags are expected to have, at least, one r_i close to 1 and negative bags with all values of $\boldsymbol{r}$ close to 0. Consequently, the malignant probability of a bag B, can be translated by taking the maximum over the $\boldsymbol{r}$ vector

$$p(y = 1|I, \theta) = \max\{r_1, r_2, ..., r_m\} \tag{4}$$

where θ represents the parameters of the CNN. The downside of this approach is that it only relies on a single instance to classify a bag, which is not optimal for a model that operates at a volume-level since, certainly, exists more than one image within a volume containing a lesion.
- **Top-*K* Pooling-based MIL:** In this case, after ranking the malignant probabilities $\boldsymbol{r} = (r_1, r_2, ..., r_m)$ for all the instances in the bag, a sort operation can be applied in descending order

$$\{r'_1, r'_2, ..., r'_m\} = sort(\{r_1, r_2, ..., r_m\}) \tag{5}$$

where $\{r'_1, r'_2, ..., r'_m\}$ corresponds to the descending ranked $\boldsymbol{r}$. This approach is particularly good for exploiting information from other instances, instead of only considering the instance with the highest malignant probability, r'_1. In fact, if the first k instances with the largest malignant probabilities are considered, the general MIL assumption is no longer adopted, since now the assumption is that each element of $\{r'_1, r'_2, ..., r'_k\}$ should be consistent with the label of the bag, while the remaining instances should be labelled as negative. The final malignant probability of the whole bag can be translated as

$$p(y = 1|I, \theta) = \frac{r'_1 + r'_2 + ... + r'_k}{k} \tag{6}$$

where θ represents the parameters of the CNN and $k > 1$. The disadvantage of this method is that a general hyper-parameter k is hard to estimate since it can vary from case to case. In the experiments made the k was chosen in an arbitrary manner, which is not optimal. Thus, an adaptive way to estimate the hyper-parameter k is preferred.
- **Adaptive Top-*K* Pooling-based MIL:** From a medical perspective, every lesion in an MRI volume typically comprises a few continuous slices. That said, this approach was designed only taking into account the Volume-wise model as it enforces choosing continuous instances inside a bag. Thus, after ranking the malignant probabilities $\boldsymbol{r} = (r_1, r_2, ..., r_m)$ and normalize them so that the sum of all the values were equal to one, a suitable approach to estimate the hyper-parameter k would be to fit a Gaussian distribution to its probability curve. This way, the expected value from the Gaussian distribution, μ, would give an idea of the lesion's center position inside the volume, and the standard deviation, σ, the rough amount of slices that the lesion occupies in the volume. Formally, we assume that the position of the lesion, X, is a random variable with Gaussian distribution, $X \sim \mathcal{N}(\mu, \sigma^2)$, in which the probability density function, $p(y_n = 1|I_n, \theta)$, represents the probability of a slice, in position x_n, being in

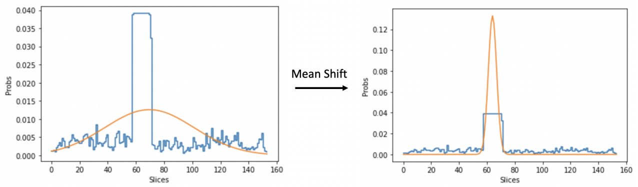


Figure 3: Example of a Gaussian distribution estimation before and after the Mean Shift application

conformity with the lesion. The **mean** (expected value) and the **standard deviation** are given by

$$\mu = E[X] = \sum_{n=0}^{N} x_n p(y_n = 1 | I_n, \theta) \tag{7}$$

$$\sigma = \sqrt{E[(X-\mu)^2]} = \sqrt{\sum_{n=0}^{N} (x_n - \mu)^2 p(y_n = 1 | I_n, \theta)} \tag{8}$$

where N is the last slice present in a volume. The final malignant probability of the bag is given by Equation 6. In theory, this parameters estimation would result in a Gaussian distribution perfectly fitted to the curve probability. However, in practice, this is not so simple as the probabilities far from the peak are not close to zero as they should be (left graph from Figure 3). This leads to the conclusion that the mean and standard deviation estimations are not noise robust. Therefore, in order to address this problem, a variation of the **Mean Shift** [4] algorithm is going to be implemented. This technique is particularly good since assigns a lower weight to data samples (x - slice, y - probability) far from the peak, enforcing the Gaussian estimation to shift towards the mean in an iterative way. Moreover, as illustrated in Figure 3, with this algorithm, it is possible to ensure that the standard deviation of the Gaussian is being shrunk (or the opposite) in each step by establishing acceptable limits to its value. These 'acceptable limits' represent the minimum and maximum number of slices in which a lesion can be found.
The mean and standard deviation updates are given by

$$\mu = E[X] = \sum_{n=0}^{N} x_n p(y_n = 1 | I_n, \theta) w_n \tag{9}$$

$$\sigma = \sqrt{\sum_{n=0}^{N} (x_n - \mu)^2 p(y_n = 1 | I_n, \theta) w_n} \tag{10}$$

where w is the probability of each slice according to the previous Gaussian distribution estimation. Once the Mean shift algorithm finishes its estimation of the new mean and standard deviation, the amount of slices that contains the lesion can be calculated.

By observing the graphs in figure 2, it is very clear that the Max and Top-k Pooling-based approaches select slices without concerning whether they are continuous, unlike the Adaptive Top-k Pooling-based approach. Nonetheless, based on those slices, a binary classifier can be achieved by choosing a threshold of 0.5 and classifying inputs with probability greater than 0.5 as malignant and smaller as not malignant. Given that we are dealing with a binary classification problem, the loss function used for training the model will be the binary cross-entropy:

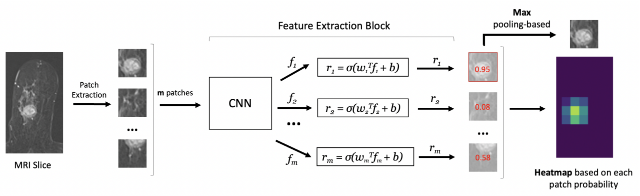


Figure 4: Slice-wise model overview

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} y_n \log(p(y_n | I_n, \theta)) + (1 - y_n) \log(1 - p(y_n | I_n, \theta)) \tag{11}$$

where N is the total number of MRI volumes, $y_n \in \{0, 1\}$ is the ground truth label and $p(y_n | I_n, \theta)$ is the predicted probability of the slice be malignant ($y_n = 1$) or not malignant ($y_n = 0$).

5.2 Slice-wise classification

The purpose of the Slice-wise model is to detect the lesions within the slices chosen by the Volume-wise model. For that to happen, and based on Figure 4, the first step is to obtain patches from each of the input slices. Only then, the process of getting features from all the patches begins. This process is exactly the same one as in the Volume-wise model. In fact, for both models, the "Feature Extraction Block" is identical, but in this case patches from the slices are used as input instead of slices from the volumes. Thus, once the probabilities of the patches are obtained through logistic regression, it is possible to classify the slice itself by using the Max Pooling-based approach, with the patches corresponding to the instances. Note that this model just needs to rely on the Max Pooling-based strategy because the lesion could be too small and only visible on a single patch. Therefore, this model follows the general MIL assumption that, if a slice has a lesion, at least, one patch contains it. Additionally, since each patch has a probability of being malignant, a heat map can be computed based on those probabilities with the same size as the input slices. The implementation of the heat map will be further explained in the next section. In order to train this model, similar to the Volume-wise model, the binary cross-entropy function (Equation 11) will be used.

6 IMPLEMENTATION

6.1 Experimental Setup

6.1.1 ***Dataset.*** This work proposes a new private dataset for training its models. This private dataset contains a compilation of several MRI scans with a Breast Imaging - Reporting and Data System (BI-RADS) classification for each one of them. Other expert annotations are unavailable for this dataset. Although each MRI scan comprises different sequences, for this work, only the one that gives a clearer view of the lesions was selected, which corresponds to the Dynamic Contrast Enhanced Subtraction (DCE sub) sequence. It is proven that this technique is accurate for detection of subtle lesions, since it can remove high-intensity signal from background fat, ending up improving lesion conspicuity and definition [12].

One of the main characteristics of this dataset is a strong class imbalance, with the majority of the MRI scans being classified as BI-RADS 1 and 5. This occurs since only the exams with strong suspicion of malignancy (observed in the mammography) are pursued for MRI. That said, the focus of this work was to solve the Normal vs Malignant problem, which corresponds to {1} vs {4, 5} in terms of BI-RADS.

The dataset used contained **164 MRI scans**. In order to train and validate the model, 134 MRI scans (71 malignants and 63 normals) were collected from that dataset. The technique used to split the data was the random sampling, which divided the data into training and validation sets in an 80%-20% ratio, respectively. Afterwards, the remaining data (30 MRI scans) was used as a test set to evaluate the performance of the models in their final version.

6.1.2 *Dataset Pre-Processing*.

Pre-processing procedures were part of this work in the hope that the model could extract the most relevant features, leading to a better performance in classification. The pre-processing made involved image normalization, cropping the image, resizing it and apply a grayscale contrast enhancement. A common task when preparing datasets for training DL models is to normalize and standardize the data, which means that all the samples should be centered and scaled according to the mean and to the standard deviation of the dataset. Then, in order to remove the chest area, all MRI volumes were cropped in terms of height so that the model could only focus on the area of interest (i.e. the breast). However, since every patient has different physical characteristics, removing the chest zone resulted on having image volumes with different sizes in terms of height. Therefore, all volumes were resized to the same dimensions. The size of the volumes ended up with 192 × 128 pixels.

Once the image volumes were cropped and resized, enhancement on each image's contrast was employed through Contrast Limited Adaptive Histogram Equalization (CLAHE) [25]. This technique partitions the images into contextual regions, called titles, and then applies the histogram equalization to each one of them. This way, the distribution of used gray values becomes more balanced and thus hidden features of the image are more visible.

Once the pre-processing at image-level was made, the first fifteen and last ten slices were removed from the volumes since those were volumes where the breasts were composing and fading, respectively. Even with this reduction, each volume ended up with its slices still ranging from 106 to 170.

6.2 Architecture

For the overall performance of a system, the computational and power efficiency of the CNN architecture is something to take into account. For this reason, the **MobileNetV2** [17] was chosen as the target state-of-the-art CNN for this work. The MobileNetV2, when compared with other CNNs, is an architecture that has a relatively small model size and very low memory requirements, which is essential for this work as it operates on volume-level instead of image-level.

Hyper-parameter Spefication	
Optimizer	Adam
Loss Function	Binary Cross Entropy
Number of Epochs	50
Batch Size	4
Learning Rate	lr = 1e-3

Table 1: MobileNetV2 Hyper-parameter Specification for the Volume-wise model

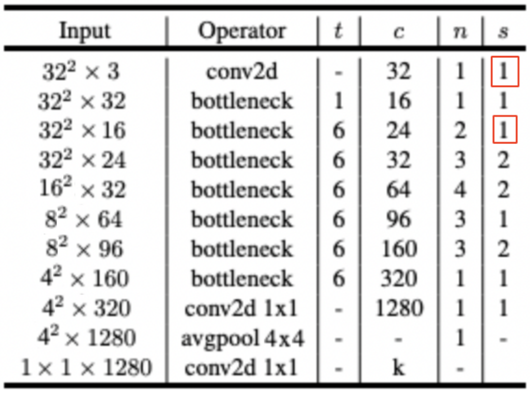
6.2.1 *Volume-wise description*.

In order to accomplish the classification and the slice-selection on the MRI volumes, as abovementioned, three MIL implementation strategies were defined. Each of them used the same MobileNetV2 architecture, which corresponds to the original configuration. The defined hyper-parameters for the network are demonstrated in Table 1. It is worth mentioning that the Batch Size number needed to be low due to the fact that each sample (bag) aggregates m images all at once, where m is the number of slices within a volume. In fact, decreasing the Batch Size number was still not enough as calculating unnecessary gradients for all those images can quickly consume all the GPU memory. Therefore, since each of the three MIL approaches only selects a certain number of slices per volume, in the training phase, the network just needs to calculate the gradients for the selected slices rather than all of them. This way, it is guaranteed that the GPU is not occupied with irrelevant information regarding the calculations. It should be noted that, for each MRI volume, the slices predicted by the Adaptive Top-k strategy were stored in a JSON file along with their respective probabilities of malignancy. This was done so that the Slice-wise model could train its model relying on the Volume-wise model.

6.2.2 *Slice-wise description*.

The Slice-wise model was implemented based on the slices outputted from the Volume-wise model. In other words, this means that the Adaptive Top-k Pooling-based approach was the only one used to extract the interesting slices from the volumes in order to train the Slice-wise model. This decision was made based on the fact that a continuous amount of slices adapted to each volume is more reliable than an arbitrary k, at least in a medical perspective. However, relying on this approach to chose the slices resulted in an unequal distribution of the input data for this model. In fact, the Adaptive Top-k Pooling-based does not work so well for negative (not malignant) cases due to the probabilities being all closer to 0, which most certainly will not follow a Gaussian Distribution. Hence, when the Gaussian distribution was not fitted as desired, most of the negative volumes reached the maximum limit of slices that was previously established by the Volume-wise model, causing a data unbalanced issue for the input data. Note that, since the Max and Top-k Pooling-based approaches have a previously known value for the hyper-parameter k, the input data for this model would be perfectly balanced. Nevertheless, the input slices were gathered in three different ways: (1) by choosing the original interval of slices from the JSON file even though that would make the input data not balanced, (2) by selecting a sub-interval from the interval of slices in the JSON file and (3) by relying on the probabilities in the JSON file to select the slices that were going to be used to train the model. Note that this last technique was implemented to refine the training input data rather than making it more balanced.

Once the input data was collected, the next challenge was to partition each slice into overlapped patches. The size of each patch was 32 × 32 pixels, and the overlapped step was half of the patch size, i.e., 16 pixels. Remembering that the size of each slice was previously defined as 192 × 128 pixels, this means that all the bags for this model ended up with the exact same amount of instances (patches). Furthermore, the MobileNetv2 architecture had to be adapted from its original form to be able to receive 32 × 32 patches. As illustrated in Figure 5, the first and the third layer were changed from stride 2 to stride 1 so that the dimension of the patches was not reduced too early in the first layers.



Input	Operator	t	c	n	s
$32^2 \times 3$	conv2d	-	32	1	1
$32^2 \times 32$	bottleneck	1	16	1	1
$32^2 \times 16$	bottleneck	6	24	2	1
$32^2 \times 24$	bottleneck	6	32	3	2
$16^2 \times 32$	bottleneck	6	64	4	2
$8^2 \times 64$	bottleneck	6	96	3	1
$8^2 \times 96$	bottleneck	6	160	3	2
$4^2 \times 160$	bottleneck	6	320	1	1
$4^2 \times 320$	conv2d 1x1	-	1280	1	1
$4^2 \times 1280$	avgpool 4x4	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	

Figure 5: Overall architecture of MobileNetV2 for the Slice-wise model (based on [17])

Once these modifications were made, the network was in conditions to be trained based on each of the abovementioned strategies for the input slices. The defined hyper-parameters for the network are identical to those shown in Table 1, with the only difference on the Batch Size, which was raised from 4 to 8.

The last step concerning the implementation of this model was the heat maps construction. To accomplish the heat maps, for each pixel i of the image, the probability of that region has a lesion, P_i, is given by averaging the probabilities, p_n, of the N_i patches that contributed for that region:

$$P_i = \frac{1}{N_i} \sum_{n=1}^{N_i} p_n \tag{12}$$

7 RESULTS

7.1 Volume-wise model experiments

The experiments made for this model aimed to compare the Adaptive Top-k against the Max and Top-k Pooling-based approaches. In that sense, the Volume-wise model was trained and validated with different choices for the hyper-parameter k. As shown in Figure 6, for this validation set, the Top-10 simulation outperformed the Adaptive Top-k Pooling-based approach. However, the accuracy started to decline with the increase in the hyper-parameter k. This behavior was expected since, for every volume, there is a limited number of slices where a malignant lesion can be found.

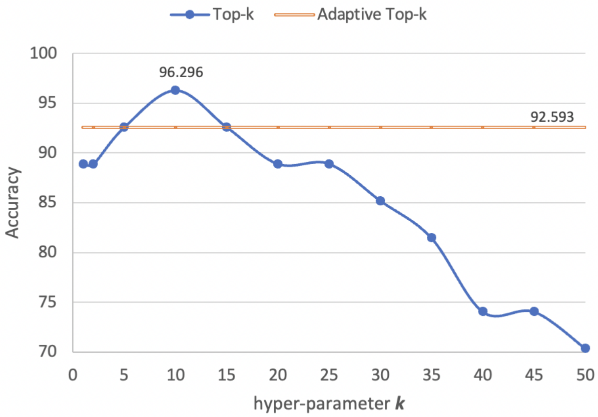


Figure 6: Accuracy comparison between the Top-k and the Adaptive Top-k Pooling-based approach

Although the Top-10 simulation seems to be preferable in terms of classification, it does not enforces a continuous selection of slices as the Adaptive Top-k does. Therefore, in order to fully assess the Top-10 simulation, the slices chosen for each of the malignant cases in the classification process were analysed. Hence, after sorting the 10 selected slices, two metrics were extracted: (1) the largest continuous sub-interval (2) and the number of discontinuities between the 10 chosen slices. From 17 positive (malignant) cases in the validation set, the mean of continuous slices chosen by the Top-10 simulation was 5.9 slices, with none of the cases reaching the full continuity. Beyond that, the mean number of discontinuities were 2.7 per case. This means that, despite the Top-10 simulation reached a higher accuracy, it is not a trustworthy model when it comes to slice-selection. Therefore, we chose the Adaptive Top-k strategy to extract the slices for the second model.

7.2 Slice-wise model experiments

As above-stated, the Slice-wise model relied on the JSON file provided by the Volume-wise model to extract the relevant slices for its training and validation phase. In order to make the data more balanced and/or avoid misclassified slices from the previous model, three different strategies were employed to the input slices used for training. In that sense, the validation set was used to assess the behaviour of the model when trained with those different strategies. It should be noted that, unlike the training set, this set ended up being balanced. In total, 221 positive slices (that contains a malignant lesion) and 232 negative slices were selected by the Adaptive Top-k Pooling-based approach from the former model. From the experiments made, despite all the results being very similar, the approach that made use of the malignant probabilities in the training phase seem to slightly outperform the other ones, reaching an accuracy of 84.3%. The results are stated in Figure 7. Note that the "Data unbalanced" strategy corresponds to the one that used the unfiltered slices from the volume-wise model and the "Data balanced" strategy the one that used a sub-interval of slices only for the negative MRI cases.

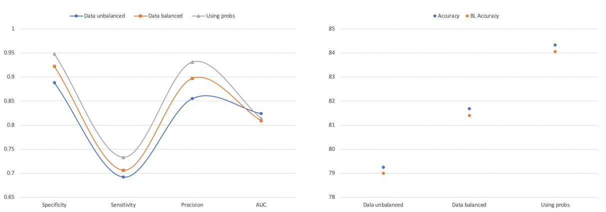


Figure 7: Validation results for the Slice-wise model

Although the approach that made use of the probabilities is the one that performs better when classifying a slice, other metrics have to be considered in order to fully understand whether the heat maps produced for each slice are in conformity with the lesion position or not. However, as mentioned before, the dataset used does not contain any annotations in terms of object localization within the image, which prevents determining to which extent the lesion location is accurately predicted.

7.3 Final experiments

The final system was composed by the Volume-wise model with the Adaptive Top-k Pooling-based approach and by the Slice-wise model with the strategy that exploited the malignant probabilities from the Volume-wise model to select the slices for the training phase. The test set used contained 30 MRI volumes, where 18 were diagnosed as malignant and 12 as normal. The evaluation process started by giving those volumes to the Volume-wise model so that the chosen slices were given as input to the Slice-wise model in a later stage. From the 30 volumes processed, the first model outputted 166 positive slices (that contains a malignant lesion) and 254 negative slices, meaning that the Slice-wise model was evaluated with $166 + 254 = 420$ slices. The classification results for both models are expressed in Table 2. Comparing the results, the performance of the Top-10 turned out not to be so outstanding as the Adaptive Top-k approach. This lead to the conclusion that a general hyper-parameter k optimal for a dataset may not be as optimal for another different dataset. That said, relying on a fixed amount of slices to classify future volumes is clearly not the best option, enforcing the idea that the Adaptive Top-k Pooling-based strategy is the most convenient approach as its the one capable of finding an optimal number of continuous slices adapted to each volume.

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 2: Classification results for the final versions of the models

In terms of lesion localization, Figure 8 presents four malignant slices with their respective heat maps. As mentioned before, neither the slices selected by the Volume-wise nor the heatmps can be truly assessed since there is no access to annotations regarding the location of the lesions. Therefore, some of those slices were surely misclassified as the first model did not reach an accuracy of 100%. Even if it did, there were no ground truth slices annotated to compare and confirm that selection. In the end, even operating with uncertainty on the data, the Slice-wise model was still capable of achieving positive accuracy results and a proper detection of malignant lesions, giving evidences that the slices facilitated by the Volume-wise model were indeed in conformity with the malignant lesion.

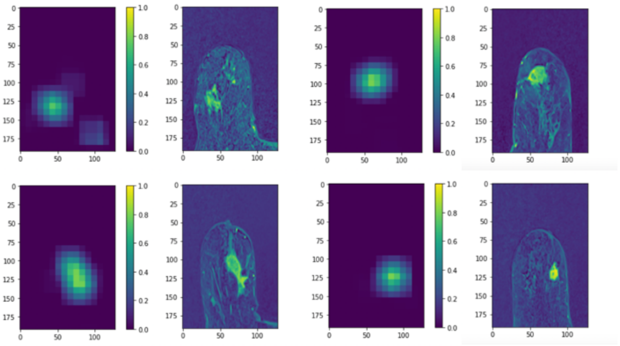


Figure 8: Detection results for four malignant predictions by the Slice-wise model

8 CONCLUSION

Breast cancer is the most common form of cancer affecting women. Its early detection has been proven to be highly beneficial when found in its earliest and most treatable stages. Due to its sensitivity, MRI screening has been used along side with mammograms to screen women who are at a high risk of having BC. However, as screenings increases, the time spent in their analyses also increases, which could overwhelm radiologists. Therefore, with the intention of helping radiologists in their workflow, a Deep MIL system is proposed in this work.

One of the problems with the MIL approaches is that the number of instances selected to classify a bag is fixed and not adapted to each case. However, this work proposed a method that adaptively selects a continuous amount of slices to classify an MRI volume. 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.

Beyond volume-wise classification and slice-selection, another objective established for this work was to perform slice-wise classification and lesion detection within the slices. Even though the dataset only had weak-labels at a volume-level instead of a slice-level, this part of the work was still possible due to the previous extraction of slices by the former model. However, as expected, the performance of this model was not so outstanding as the former one in terms of classification. We do not consider this as a problem since, from a medical perspective, the volume-wise classification is the one that truly matters. Beyond that, it was still possible to highlight the abnormal regions of the slices through heat maps, meaning that the radiologists could also reconfirm the position of the lesion within the slices when making their final judgment.

9 FUTURE WORK

Due to the positive results achieved, this thesis can serve as a starting point for other works that may want to explore the MIL framework.

Remembering that this work only used the DCE sub sequence, one of the possibilities to extend it would be to explore and compare the behavior of the proposed system with different MRI sequences as input. Once this work is done, it also could be enlarged to another type of BC screening modality, such as the Mammography or even the Ultrasound. This way, it would be possible not only to conclude whether the DCE sub sequence is indeed the most reliable sequence but also to compare the performance of the different screenings used in the BC field.

Another possibility to extend this work would be by adding benign cases to the dataset, with the purpose of distinguishing Severe cases (malign) from Mild cases (no lesion or benign). In practical terms, this is the same as establishing a binary classifier prepared to discriminate volumes with BI-RADS {1,2,3} from {4,5}.

Finally, we believe that the progression of the dataset used is also of great importance. Despite the results obtained, it would be worthwhile to understand the behavior of the models proposed when trained and evaluated with more data. Furthermore, adding more annotations to the data regarding the location and size of the lesions within the volumes would also be beneficial for future work. With this type of additional information it will be possible not only to compare if the slices selected are indeed the slices that justifies the classification but also to truly assess the lesion detection results.

REFERENCES

[1] Shannon C. Agner, Jun Xu, Hussain Fatakdawala, Shridar Ganesan, Anant Madabhushi, Sarah Englander, Mark Rosen, Kathleen Thomas, Mitchell Schnall, Michael Feldman, and John Tomaszewski. 2009. Segmentation and classification of triple negative breast cancers using DCE-MRI. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. https://doi.org/10.1109/isbi.2009.5193283

[2] N. Biglia, V.E. Bounous, L. Martincich, E. Panuccio, V. Liberale, L. Ottino, R. Ponzone, and P. Sismondi. 2011. Role of MRI (magnetic resonance imaging) versus conventional imaging for breast cancer presurgical staging in young women or with dense breast. *European Journal of Surgical Oncology (EJSO)* 37, 3 (March 2011), 199–204. https://doi.org/10.1016/j.ejso.2010.12.011

[3] Jeremy R Burt, Neslisah Torosdagli, Naji Khosravan, Harish RaviPrakash, Aliasghar Mortazi, Fiona Tissavirasingham, Sarfaraz Hussein, and Ulas Bagci. 2018. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. *The British Journal of Radiology* (April 2018), 20170545. https://doi.org/10.1259/bjr.20170545

[4] D. Comaniciu and P. Meer. 2002. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 5 (May 2002), 603–619. https://doi.org/10.1109/34.1000236

[5] Jianrui Ding, H. D. Cheng, Jianhua Huang, Jiafeng Liu, and Yingtao Zhang. 2012. Breast Ultrasound Image Classification Based on Multiple-Instance Learning. *Journal of Digital Imaging* 25, 5 (June 2012), 620–627. https://doi.org/10.1007/s10278-012-9499-x

[6] Stephen W. Duffy, László Tabár, Amy Ming-Fang Yen, Peter B. Dean, Robert A. Smith, Håkan Jonsson, Sven Törnberg, Sam Li-Sheng Chen, Sherry Yueh-Hsia Chiu, Jean Ching-Yuan Fann, May Mei-Sheng Ku, Wendy Yi-Ying Wu, Chen-Yang Hsu, Yu-Ching Chen, Gunilla Svane, Edward Azavedo, Helene Grundström, Per Sundén, Karin Leifland, Ewa Frodis, Joakim Ramos, Birgitta Epstein, Anders Åkerlund, Ann Sundbom, Pål Bordás, Hans Wallin, Leena Starck, Annika Björkgren, Stina Carlson, Irma Fredriksson, Johan Ahlgren, Daniel Öhman, Lars Holmberg, and Tony Hsiu-Hsi Chen. 2020. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549, 091 women. *Cancer* 126, 13 (May 2020), 2971–2979. https://doi.org/10.1002/cncr.32859

[7] Joshua J. Fenton, Guibo Xing, Joann G. Elmore, Heejung Bang, Steven L. Chen, Karen K. Lindfors, and Laura-Mae Baldwin. 2013. Short-Term Outcomes of Screening Mammography Using Computer-Aided Detection. *Annals of Internal Medicine* 158, 8 (April 2013), 580. https://doi.org/10.7326/0003-4819-158-8-201304160-00002

[8] Albert Gubern-Mérida, Robert Martí, Jaime Melendez, Jakob L. Hauth, Ritse M. Mann, Nico Karssemeijer, and Bram Platel. 2015. Automated localization of breast cancer in DCE-MRI. *Medical Image Analysis* 20, 1 (Feb. 2015), 265–274. https://doi.org/10.1016/j.media.2014.12.001

[9] Lubomir Hadjiiski, Heang-Ping Chan, Berkman Sahiner, Mark A. Helvie, Marilyn A. Roubidoux, Caroline Blane, Chintana Paramagul, Nicholas Petrick, Janet Bailey, Katherine Klein, Michelle Foster, Stephanie Patterson, Dorit Adler, Alexis Nees, and Joseph Shen. 2004. Improvement in Radiologists' Characterization of Malignant and Benign Breast Masses on Serial Mammograms with Computer-aided Diagnosis: An ROC Study. *Radiology* 233, 1 (Oct. 2004), 255–265. https://doi.org/10.1148/radiol.2331030432

[10] Lubomir Hadjiiski, Berkman Sahiner, and Heang-Ping Chan. 2006. Advances in computer-aided diagnosis for breast cancer. *Current Opinion in Obstetrics & Gynecology* 18, 1 (Feb. 2006), 64–70. https://doi.org/10.1097/01.gco.0000192965.29449.da

[11] P. Herent, B. Schmauch, P. Jehanno, O. Dehaene, C. Saillard, C. Balleyguier, J. Arfi-Rouche, and S. Jégou. 2019. Detection and characterization of MRI breast lesions using deep learning. *Diagnostic and Interventional Imaging* 100, 4 (April 2019), 219–225. https://doi.org/10.1016/j.diii.2019.02.008

[12] Vivian S Lee, Mark A Flyer, Jeffrey C Weinreb, Glenn A Krinsky, and Neil M Rofsky. 1996. Image subtraction in gadolinium-enhanced MR imaging. *AJR. American journal of roentgenology* 167, 6 (1996), 1427–1432.

[13] Constance D. Lehman, Jeffrey D. Blume, Paul Weatherall, David Thickman, Nola Hylton, Ellen Warner, Etta Pisano, Stuart J. Schnitt, Constantine Gatsonis, and Mitchell Schnall and. 2005. Screening women at high risk for breast cancer with mammography and magnetic resonance imaging. *Cancer* 103, 9 (2005), 1898–1905. https://doi.org/10.1002/cncr.20971

[14] Lina Arbash Meinel, Alan H. Stolpen, Kevin S. Berbaum, Laurie L. Fajardo, and Joseph M. Reinhardt. 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, 1 (Jan. 2007), 89–95. https://doi.org/10.1002/jmri.20794

[15] N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, and L. von Karsa. 2008. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. *Annals of Oncology* 19, 4 (April 2008), 614–622. https://doi.org/10.1093/annonc/mdm481

[16] Reza Rasti, Mohammad Teshnehlab, and Son Lam Phung. 2017. Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks. *Pattern Recognition* 72 (Dec. 2017), 381–390. https://doi.org/10.1016/j.patcog.2017.08.004

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. (2018). https://doi.org/10.48550/ARXIV.1801.04381

[18] Chandra K. Sarath, Arunava Chakravarty, Nirmalya Ghosh, Tandra Sarkar, Ramanathan Sethuraman, and Debdoot Sheet. 2020. 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)*. IEEE. https://doi.org/10.1109/embc44109.2020.9176427

[19] Francesco Sardanelli, Gian M. Giuseppetti, Pietro Panizza, Massimo Bazzocchi, Alfonso Fausto, Giovanni Simonetti, Vincenzo Lattanzio, and Alessandro Del Maschio. 2004. Sensitivity of MRI Versus Mammography for Detecting Foci of Multifocal, Multicentric Breast Cancer in Fatty and Dense Breasts Using the Whole-Breast Pathologic Examination as a Gold Standard. *American Journal of Roentgenology* 183, 4 (Oct. 2004), 1149–1157. https://doi.org/10.2214/ajr.183.4.1831149

[20] D. Saslow, C. Boetes, W. Burke, S. Harms, M. O. Leach, C. D. Lehman, E. Morris, E. Pisano, M. Schnall, S. Sener, R. A. Smith, E. Warner, M. Yaffe, K. S. Andrews, and C. A. Russell and. 2007. American Cancer Society Guidelines for Breast Screening with MRI as an Adjunct to Mammography. *CA: A Cancer Journal for Clinicians* 57, 2 (March 2007), 75–89. https://doi.org/10.3322/canjclin.57.2.75

[21] Meiyin Wu and Li Chen. 2015. Image recognition based on deep learning. In *2015 Chinese Automation Congress (CAC)*. IEEE. https://doi.org/10.1109/cac.2015.7382560

[22] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanislaw Jastrzebski, Thibault Fevry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. 2020. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging* 39, 4 (April 2020), 1184–1194. https://doi.org/10.1109/tmi.2019.2945514

[23] Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (Aug. 2017), 44–53. https://doi.org/10.1093/nsr/nwx106

[24] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. 2017. Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Springer International Publishing, 603–611. https://doi.org/10.1007/978-3-319-66179-7_69

[25] Karel Zuiderveld. 1994. Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems*. Elsevier, 474–485. https://doi.org/10.1016/b978-0-12-336156-1.50061-6