This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2521442, IEEE Transactions on Medical Imaging

Multiple-Instance Learning for Anomaly Detection in Digital Mammography

Gwenolé Quellec, Mathieu Lamard, Michel Cozic, Gouenou Coatrieux, Guy Cazuguel

***Abstract*—This paper describes a computer-aided detection and diagnosis system for breast cancer, the most common form of cancer among women, using mammography. The system relies on the Multiple-Instance Learning (MIL) paradigm, which has proven useful for medical decision support in previous works from our team. In the proposed framework, breasts are first partitioned adaptively into regions. Then, features derived from the detection of lesions (masses and microcalcifications) as well as textural features, are extracted from each region and combined in order to classify mammography examinations as "normal" or "abnormal". Whenever an abnormal examination record is detected, the regions that induced that automated diagnosis can be highlighted. Two strategies are evaluated to define this anomaly detector. In a first scenario, manual segmentations of lesions are used to train an SVM that assigns an anomaly index to each region; local anomaly indices are then combined into a global anomaly index. In a second scenario, the local and global anomaly detectors are trained simultaneously, without manual segmentations, using various MIL algorithms (DD, APR, mi-SVM, MI-SVM and MILBoost). Experiments on the DDSM dataset show that the second approach, which is only weakly-supervised, surprisingly outperforms the first approach, even though it is strongly-supervised. This suggests that anomaly detectors can be advantageously trained on large medical image archives, without the need for manual segmentation.**

***Index Terms*—mammography, anomaly detection, multiple-instance learning, texture analysis**

I. Introduction

BREAST cancer is the most common form of cancer among women worldwide. Currently, mammography is the most effective tool for early detection of breast cancer, which increases the survivability of patients. In order to help radiologists interpret mammograms, many computer-aided detection and diagnosis systems have been developed in recent years [1]. The purpose of Computer-Aided Detection (CADe) systems is to detect suspicious areas within the breast that the radiologist should look at. These systems usually involve detecting candidate lesions: masses or microcalcifications (MCs). The purpose of Computer-Aided Diagnosis (CADx) systems, which sometimes rely on CADe systems, is to help assess the risk of cancer based on the analysis of the detected (and possibly segmented) lesions. These systems usually involve characterizing the suspected lesions and building a classifier based on these characterizations [2].

Several publicly available mammography datasets can be used to train and evaluate CADe and CADx systems. The largest such dataset is the Digital Database for Screening Mammography (DDSM) [3]. In this dataset, mammography images come with a manual segmentation of the findings (e.g. masses and MCs), in addition to the radiologist's report, which contains a description of each finding and a risk assessment for the patient. Using manual segmentations for supervision is expected to produce better CADe and CADx systems than simply using the radiologists' reports. But in their daily practice, radiologists usually do not manually segment findings in images. So relying on manual segmentations prevents us from training those systems in large, real-live datasets. Training on very large datasets clearly has an advantage: as the size of the dataset increases, the algorithm learns a greater variety of lesions, so it is expected to generalize better. But different training paradigms are needed to cope with the lack of manual segmentations.

We reviewed the recent literature about CADe and CADx systems relying on the DDSM dataset for training and testing. First, all CADe systems rely on the manual segmentation of findings. This is true for the detection [4], [5], [6], [7], [8], [9], [10], [11], false positive reduction [12], [13], [14], [15], [16] and segmentation [17], [10] of masses. This is also true for the detection [6], [18] and segmentation [19], [20] of MCs. Most CADx systems, whether mass classification systems [21], [5], [22], [23], [24], [25], [26], [27], [28], MC classification systems [21], [26], [29], or malignant / non-malignant classification systems in general [30], also rely on those manual segmentations. A few CADx systems, however, have been proposed for the classification of images or examination records as a whole, using global labels for training and testing. One system estimates breast tissue density in images as a whole [31]. Three others estimate the risk assessment assigned to images as a whole [32] or to examination records as a whole [33], [34]. However, these systems, which rely on the content-based image retrieval paradigm [35], work as black boxes: they cannot indicate which parts of images induced the automatic diagnosis. In this paper, the goal is to recognize abnormal examination records, using global labels only, and to highlight abnormal regions whenever an abnormal

G. Quellec, M. Lamard, G. Coatrieux and G. Cazuguel is with Inserm, UMR 1101, Brest, F-29200 France (e-mail: gwenole.quellec@inserm.fr)
M. Lamard is with Univ Bretagne Occidentale, Brest, F-29200 France
M. Cozic is with Medecom, Plougastel-Daoulas, F-29470 France
G. Coatrieux and G. Cazuguel are with INSTITUT TELECOM; TELECOM Bretagne; UEB; Dpt ITI, Brest, F-29200 France

examination record is detected.

The proposed solution relies on Multiple-Instance learning (MIL) [36]. MIL is a useful paradigm to solve image analysis problems where visual features are extracted locally in images, typically within image patches, but the labels used for supervision are assigned to images as a whole or even examination records as a whole. In recent studies, we have shown the relevance of MIL for both medical image and video analysis applications [37], [38]. Although the number of publications about MIL for medical image analysis is limited, the target is broad: pulmonary embolism and colon cancer detection [39], histopathology-based cancer diagnosis [40], [41], [42], [43], diabetic retinopathy diagnosis [37], [44], [40], breast ultrasound image classification [45], colitis detection [46], liver cancer recognition [47], dementia classification [48], [49], glandular tissue classification [50], tuberculosis diagnosis [51] and gastric cancer diagnosis [52]. Finally, MIL has also been applied to mammography images: MIL was used either for mass detection [53], [54] or for MC detection [55]. But MIL has never been used for anomaly detection in general, both at global and local level The advantage of MIL, over strongly-supervised CAD systems, is that clinicians are not required to manually segment mammograms in order to collect the ground truth required for training: examination reports are enough for supervision. The implication is huge: because manual segmentation is not required, millions of mammograms could potentially be used for training in order to better capture the variability of mammograms.

In order to apply these algorithms to mammograms efficiently, each image must be divided into regions and a feature vector should be extracted from each region. An adaptive region definition framework is presented in section II and the investigated features are presented in section III. It should be noted that in the particular case of mammography screening, the class label is not assigned to single images, but rather to mammography examinations, consisting of four images (one craniocaudal view and one mediolateral-oblique view of each breast). So a bag of instances groups together regions from four images: a specific solution is proposed for this scenario. The proposed MIL framework is presented in section IV and experiments in the DDSM dataset are reported in section V.

II. Partitioning the Breast into Regions

To apply MIL algorithms, images are usually partitioned into rectangular regions or "patches". Because a large portion of mammograms is irrelevant (black background, labels, etc.), previous MIL-based systems for mammograms only defined regions inside the breast area (§II-B). For mass detection, regions with arbitrary shapes were considered: each region is a candidate mass [53]. This approach is not suitable for MC cluster detection: regions would be disconnected objects and may be too small to extract texture features for instance. Therefore, Li et al. defined rectangular regions within the breast area [55]. Because no specific lesion type is targeted in this paper, a general region definition is needed as well. It should be noted that, depending on the features that are extracted from each region, rectangular regions may not be suitable for mammography images. Typically, local shape features or texture features would be inconsistent inside regions intersecting the breast edges. One solution would be to consider regions strictly included inside the breast area, but then some lesions may be missed. Instead, we propose to define regions with adaptive shapes that fit the shape of the breast (§II-B). These regions are also defined with respect to the nipple (§II-C) so that a relevant region coordinate system can be defined for anomaly detection (§III-C).

A. Preprocessing Images (see Fig. 1 (a))

Before processing images, the first step is to map the gray levels in input images to optical densities [3], which are then encoded on 12 bits. Therefore, it is possible to process images acquired with various scanners, and possibly digitized with various digitizers. For faster breast segmentation and nipple detection, images are downsampled by a factor of 4 (the original definition is used for feature extraction in section III).

B. Segmenting the Breast

To detect the breast area, bright areas need to be detected. But this is not enough: because image intensity decreases near the breast edges, these edges are usually dark. Nevertheless, breast edges are sharp, so detecting sharp areas as well ensures that breast are not under-segmented. Bright and sharp image areas are detected using a threshold δ_I on image intensities and a threshold δ_D on intensity derivatives (derivative filter radius: 8 pixels): $\delta_I = \delta_D = 1200$ (30% of the intensity range). To make sure noise is not enhanced with the derivative filter, the image is smoothed beforehand using a median filter (radius: 20 pixels). Holes, which may appear between sharp and bright areas in the resulting binary image are filled. Finally, the breast is defined as the largest connected component.

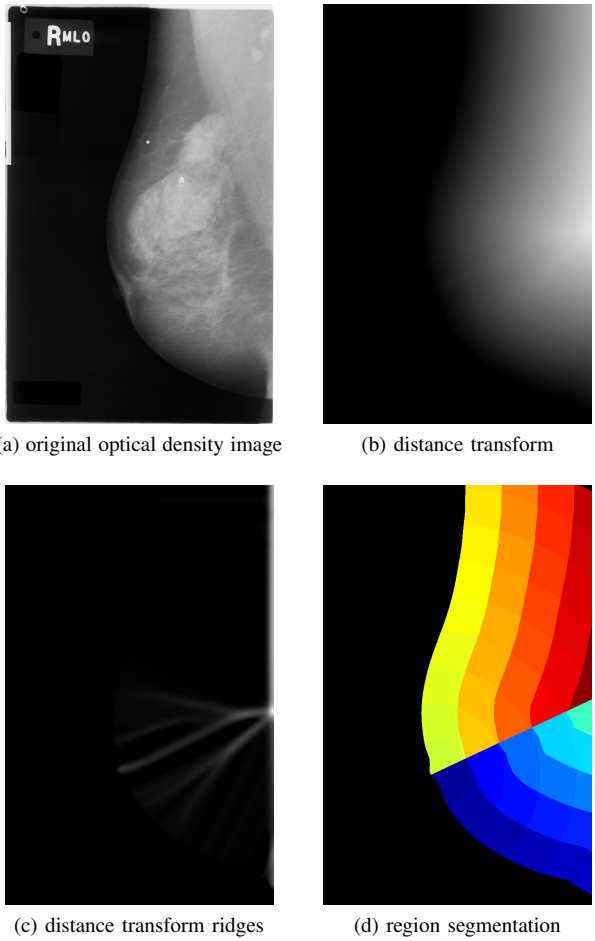
C. Finding the Nipple (see Fig. 1 (b)-(c))

Once the breast is segmented, the distance from each pixel in the breast to the nearest pixel on the breast edges (excluding the image border itself), is computed using the Maurer distance transform [56] (see Fig. 1 (b)). Ridges of this distance map are then detected using a Laplacian filter (radius: 4 pixels — see Fig. 1 (c)). It can be observed that the main ridge runs orthogonally to the breast edges and intersects those edges very close to the nipple. Bright ridge pixels are then fitted with a line, referred to as the main axis. A robust estimation of the nipple is given by the intersection of this line and of the breast edges.

D. Defining Regions (see Fig. 1 (d))

Image pixels are grouped together in regions with respect to their distance to the breast edges and to their distance to the nipple. The distance transform defined above is used

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2521442, IEEE Transactions on Medical Imaging



(a) original optical density image (b) distance transform

(c) distance transform ridges (d) region segmentation

Fig. 1: Breast segmentation and region definition

to partition the breast into uniform intervals of distances to the breast edges. The breast is also partitioned into uniform intervals of Euclidean distances to the nipple; the same interval width Δ is used for both distances (parameter optimization is presented in section V-D). Finally, the breast is partitioned into two halves, based on the main axis. Regions are defined as intersections of these partitions.

III. Feature Extraction

Three types of features are extracted from each region $\mathcal{R}$ defined above. Following the CADe literature, features derived from the detection of masses and MCs are extracted (§III-A). Following the CADx literature, texture features are also extracted from $\mathcal{R}$ (§III-B). Indeed, recent CADx algorithms predominantly focus on texture for the classification of lesions, patches, images or examination records. Texture features that were used for mammography CADx include Gabor wavelets [5], [26], derivatives of Gaussians [22], discrete wavelets and their derivatives (curvelets, contourlets) [33], [34], [26], [28], filters derived from a principal or independent component analysis [5], singular-value decomposition [31], gray-level difference matrices [26], [27], gray-level co-ocurrence matrices and Haralick features [32], [26], [30], gray-level run-length matrices [26], [27] and local binary pattern histograms [27], [28]. Shape features are popular as well [23], [24], [25] but unlike textural features, which can be extracted directly from arbitrary regions, shape features require a fine segmentation of suspected lesions, which limits the generality of the proposed region-based approach. Finally, the coordinates of region $\mathcal{R}$ are also used as descriptors (§III-C).

A. Features Derived from Lesion Detection

1) Mass Detection: a recent topographic approach was used for mass detection [57]. An isocontour map is built for each mammogram using N_{mass} topographic levels. In this map, a salient region (e.g. a mass) forms a dense quasi-concentric pattern of contours. The structure of that mammogram is analyzed using an inclusion tree that is a hierarchical representation of the enclosure relationships between contours. The saliency of a region enclosed by a contour, corresponding to one vertex in the tree, is defined as the minimum nesting depth, i.e. the minimum distance from that vertex to a descendant leaf. A local mass probability $p_{mass}(x)$ is defined at each pixel location p: it is defined as the maximal minimum nesting depth computed for contours enclosing that pixel (see Fig. 2 (a)-(b)).

2) Microcalcification Detection: a novel solution was used for MC detection (see Fig. 2 (c)-(d)). The proposed algorithm also looks for salient objects, but at a smaller scale and with more subtle intensity variations. The solution relies on the Fast Radial Symmetry Transform (FRST) [58], a detector initially proposed for face detection. As the name suggest, this detector uses local image gradients to locate points of high radial symmetry, i.e. small roundish objects. A local radial symmetry index $p_{MC}(x)$ is computed at each pixel location x: this index is used as a local MC probability. It combines local gradients computed in one or several circular neighborhoods centered on that location; let $\mathcal{S}_{MC}$ denote the set of neighborhood radii. To our best knowledge, this detector has never been applied to MC detection before.

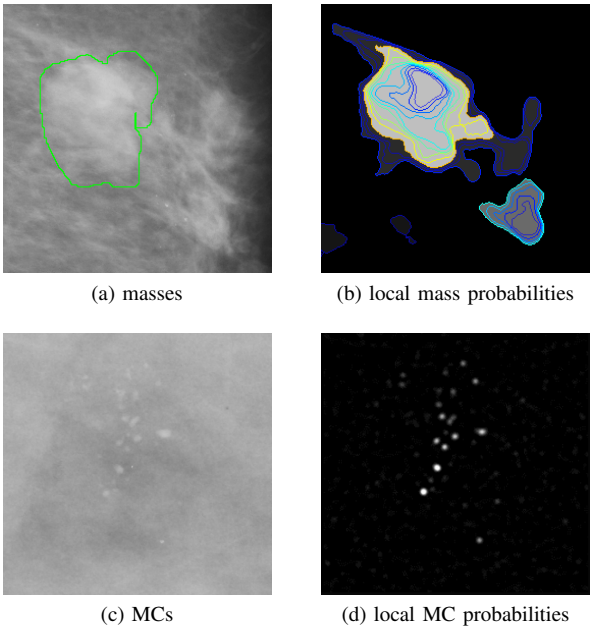
3) Feature vector: for each region $\mathcal{R}$ and each type of lesions l, the complementary cumulative distribution function $\overline{F}_l$ of the local lesion probabilities $(p_l(x))_{x\in\mathcal{R}}$ is computed:

$$\overline{F}_l(X) = \frac{1}{|\mathcal{R}|}\sum_{x\in\mathcal{R}} \mathbf{1}_{p_l(x)>X} \quad (1)$$

The feature vector describing $\mathcal{R}$ is defined as the concatenation of $\overline{F}_{mass}$ and $\overline{F}_{MC}$, discretized on L_{mass} and L_{MC} levels, respectively.

B. Texture Features

Texture features from one of the following four well-known families are extracted from region $\mathcal{R}$: gray-level co-occurrence matrices (GLCM) [59], the Haralick features based on those matrices (Haralick) [59], gray-level run-length matrices (GLRLM) [60] or histograms of local binary patterns (LBPH) [61]. To compute GLCM and GLRLM matrices, image intensities are first quantized on



(a) masses (b) local mass probabilities

(c) MCs (d) local MC probabilities

Fig. 2: Local lesion probabilities. In Fig. a, the manual segmentation from DDSM is in green. In Fig. b and d, lesion probabilities are proportional to gray level intensities. In Fig. b, the minimum nesting depth is represented by the color of isocontours: warmer colors indicate larger depths.

$Q_{texture}$ levels. Then, features are extracted at multiple scales and concatenated; let $\mathcal{S}_{texture}$ denote the set of subsampling factors. To avoid the curse of dimensionality, the resulting feature vector is further compressed by a principal component analysis. The feature vector describing $\mathcal{R}$ is defined as the first $L_{texture}$ principal components.

C. Region Coordinates

The region coordinates are two-dimensional: the first dimension derives from the distance to the breast edges, the second dimension derives from the distance to the nipple (§II-D).

To summarize, region $\mathcal{R}$ is described by a K-dimensional feature vector, $K = L_{mass} + L_{MC} + L_{texture} + 2$. To facilitate classification, each feature is normalized (zero mean, unit standard deviation) and the normalized feature vectors are fed to various MIL-based classifiers.

IV. Multiple-Instance Learning (MIL)

Multiple-instance learning generalizes supervised learning. In supervised learning, each instance is described by a feature vector and is associated with a class label. In MIL, each instance is also described by a feature vector. But in that case, the class label is not associated with an instance: it is associated with a *bags of instances*, containing an arbitrary large number of instances [36]. A bag is negative (i.e. "normal") if and only if all its instances are negative. It is positive (i.e. "abnormal") if and only if at least one of its instances is positive. In image processing, this representation is useful when images are divided into regions and a feature vector is extracted from each region: each feature vector is regarded as an instance and each image (or each examination record) is regarded as a bag of instances. If no manual segmentation is available for images, then only the class labels of images (or of examination records) are known. From an MIL point of view, only the class labels of bags of instances are known.

A. Implemented MIL Algorithms

1) Categories of MIL Algorithms: many algorithms exist for solving MIL problems [36], [40]. These algorithms can be classified into two main categories. In a first category, a global anomaly index is assigned to each bag of instances, but no local anomaly index can be assigned to each instance individually. The reason is that these algorithms rely on a global distance between distributions of instances. Among others, this category includes Citation k-NN [62] and mi-Graph [63]. In a second category, local anomaly indices are defined first and the global anomaly index results from these local indices. This paper focuses on the second category: in the proposed framework, we would like to highlight abnormal regions whenever a mammography examination is suspected to be abnormal.

2) Selected Algorithms in the Second Category: five algorithms in the second category are evaluated in this paper: Diverse Density (DD) [64] and Axis-Parallel Rectangles (APR) [65], which were the first MIL solutions, mi-SVM and MI-SVM [66], which generalize the popular support-vector machines (SVM) [67] in the context of MIL, and MILBoost [68], [69], which generalizes the popular ensemble classifiers in that context. More MIL algorithms exist, but most of them are variations on these algorithms [36], [40].

3) notation: Let B_i denote the i^{th} bag of instances in a training dataset and let $y_i \in \{-1, 1\}$ denote its label. Let $\boldsymbol{B}_{ij}$ denote the j^{th} instance in bag B_i. Let I_k denote the k^{th} feature of instance $\boldsymbol{I}$, $k = 1..K$.

B. Diverse Density (DD)

Diverse Density (DD) is a criterion used to search for one area of the feature space where, ideally, all positive bags have at least one instance but no negative bags have instances [64]. This area is defined as the neighborhood of a *key instance* $\widehat{\boldsymbol{I}} = (\widehat{I}_k)_{k=1..K}$. The optimal location of $\widehat{\boldsymbol{I}}$ in feature space, as well as the feature weights $\boldsymbol{\sigma} = (\sigma_k)_{k=1..K}$ defining the optimal neighborhood, are found by maximizing $DD(\widehat{\boldsymbol{I}}, \boldsymbol{\sigma})$:

$$\left\{\begin{array}{rcl} DD(\widehat{\boldsymbol{I}}, \boldsymbol{\sigma}) & = & \prod_{i, y_i=1} Pr(+|B_i, \widehat{\boldsymbol{I}}, \boldsymbol{\sigma}) \\ & \times & \prod_{i, y_i=-1} Pr(-|B_i, \widehat{\boldsymbol{I}}, \boldsymbol{\sigma}) \\ Pr(+|B, \widehat{\boldsymbol{I}}, \boldsymbol{\sigma}) & = & 1 - \prod_j \left[1 - a'(\boldsymbol{B}_j, \widehat{\boldsymbol{I}}, \boldsymbol{\sigma})\right] \\ Pr(-|B, \widehat{\boldsymbol{I}}, \boldsymbol{\sigma}) & = & \prod_j \left[1 - a'(\boldsymbol{B}_j, \widehat{\boldsymbol{I}}, \boldsymbol{\sigma})\right] \\ a'(\boldsymbol{I}, \widehat{\boldsymbol{I}}, \boldsymbol{\sigma}) & = & e^{-\sum_{k=1}^{K} \left[\sigma_k^2 \left(I_k - \widehat{I}_k\right)^2\right]} \end{array}\right. \quad (2)$$

The optimal values for $\widehat{\boldsymbol{I}}$ and $\boldsymbol{\sigma}$, namely $\widehat{\boldsymbol{I}}^*$ and $\boldsymbol{\sigma}^*$, are found by a gradient ascent [64]. In this paper, N_{DD} gradient ascents are performed: each ascent is initialized by a randomly selected instance within the positive bags and by unit weights; the solution maximizing DD is retained. The anomaly index of instance $\boldsymbol{I}$ is defined as $a'(\boldsymbol{I}, \widehat{\boldsymbol{I}}^*, \boldsymbol{\sigma}^*)$. The anomaly index of bag B is defined as $Pr(+|B, \widehat{\boldsymbol{I}}^*, \boldsymbol{\sigma}^*)$, namely the noisy-OR of the instance anomaly indices.

C. Axis-Parallel Rectangles (APR)

The second algorithm looks for the smallest axis-parallel rectangle (APR) in feature space that contains no instances from negative bags but, ideally, contains at least one instance from each positive bag [65]. Three heuristics are proposed to define this APR. Experiments by the authors have proven the superiority of the "inside-out" heuristic, which starts from a seed point and then iteratively grows an APR from it to include more positive instances. Kernel density estimation is used to improve generalization while updating the bounds along each feature dimension: these probability density functions are used to define a continuous anomaly index $a(\boldsymbol{I})$ for each instance $\boldsymbol{I}$. The anomaly index of bag B is defined as $\max_j a(\boldsymbol{B}_j)$.

D. Multiple-Instance Support Vector Machines

Two generalizations of the well-known strongly-supervised SVM algorithm have been proposed by Andrews et al. [66]. In a strongly-supervised setting, the hyperplane maximizing the separation between positive and negative instances in feature space (orthogonally to that hyperplane) is searched for. Using the so-called kernel trick, this hyperplane can also be searched for in a higher-dimensional feature space mapped by some kernel function ϕ (the popular Gaussian kernel was used).

1) mi-SVM: in the first weakly-supervised generalization, a joint optimization is performed: 1) a label $y_{ij} \in \{-1, 1\}$ is assigned to each instance $\boldsymbol{B}_{ij}$ within each bag B_i and 2) the optimal hyperplane w.r.t. these assignments is searched for [66]:

$$\begin{cases} \min\limits_{y_{ij}} \min\limits_{\boldsymbol{w},b,\xi_{ij}} \dfrac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum\limits_{i,j}\xi_{ij} \\ s.t. \quad \forall i,j : y_{ij}a''(\boldsymbol{B}_{ij},\boldsymbol{w},b) \geq 1-\xi_{ij}, \xi_{ij} \geq 0 \\ s.t. \quad \forall i : \max\limits_j y_{ij} = y_i \\ \qquad a''(\boldsymbol{I},\boldsymbol{w},b) = \left(\boldsymbol{w}^T\phi\left(\boldsymbol{I}\right)+b\right) \end{cases} \tag{3}$$

These two optimization tasks are performed alternately until convergence. Initially, all instances from positive bags are assigned a positive label: an initial strongly-supervised SVM is trained. To cope with class imbalance, each class C ('negative' or 'positive') is weighted by $1/|C|$. Then, instances classified negative by this initial classifier ($a''(\boldsymbol{I},\boldsymbol{w},b) < 0$) are assigned a negative label. Following the MIL definition, if all instances in some positive bag B are assigned a negative label, the label of the instance maximizing $a''(\boldsymbol{B}_j,\boldsymbol{w},b)$ is switched to positive. Then, a second strongly-supervised SVM is trained using the new assignments, and so on until convergence.

2) MI-SVM: in the second generalization, the goal is to maximize the separation between positive and negative bags (and not instances) in feature space. In that purpose, the algorithm focuses on the "most positive" instance in positive bags and the "least negative" instance in negative bags. The optimal hyperplane is defined as follows:

$$\begin{cases} \min\limits_{\boldsymbol{w},b,\xi_i} \dfrac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum\limits_i \xi_i \\ s.t. \quad \forall i : y_i \max_j a''(\boldsymbol{B}_{ij},\boldsymbol{w},b) \geq 1-\xi_i, \xi_i \geq 0 \end{cases} \tag{4}$$

This algorithm also relies on two-step iterations. Because only one instance per bag matters, this second generalization of SVM is expected to be less suitable for detecting abnormal regions, but more suitable for detecting abnormal examinations as a whole.

In mi-SVM and MI-SVM, the anomaly index of instance $\boldsymbol{I}$ is defined as $a''(\boldsymbol{I},\boldsymbol{w}^*,b^*)$, where $(\boldsymbol{w}^*,b^*)$ denotes the optimal hyperplane. The anomaly index of bag B is defined as $\max_j a''(\boldsymbol{B}_j,\boldsymbol{w}^*,b^*)$.

E. MILBoost

MILBoost generalizes boosting, a well-known ensemble learning framework [68]: in boosting, a *strong* instance classifier $\boldsymbol{h}$ is built by linearly combining multiple *weak* instance classifiers. In this paper, a weak classifier h_l derives from a single feature: $h_l(\boldsymbol{I}) = \text{sign}(\delta_l I_{f_l} - b_l)$, with $\delta_l \in \{-1, 1\}$, $f_l \in \{1, ..., K\}$ and $b_l \in \mathbb{R}$. Weak classifiers are iteratively added to the strong classifier. At each iteration, a weight is assigned to each instance: this weight depends on the bag's classification error using the current strong classifier. Then, one weak classifier h_k is trained per feature $f_k \in \{1, ..., K\}$ w.r.t. the instance weights. Finally, the weak classifier minimizing $\mathcal{L}(\boldsymbol{h}\oplus h_k, \boldsymbol{\sigma}\oplus\sigma_k)$, where σ_k is obtained by a line search and $\oplus$ denotes the concatenation operator, is added to the strong classifier [70]:

$$\begin{cases} \mathcal{L}(\boldsymbol{h},\boldsymbol{\sigma}) &= -\sum\limits_{i,y_i=1}\log p(B_i,\boldsymbol{h},\boldsymbol{\sigma}) \\ &\quad -\sum\limits_{i,y_i=-1}\log(1-p(B_i,\boldsymbol{h},\boldsymbol{\sigma})) \\ p(B,\boldsymbol{h},\boldsymbol{\sigma}) &= \underset{j}{\text{softmax}}\left(a'''(\boldsymbol{B}_j,\boldsymbol{h},\boldsymbol{\sigma})\right) \\ a'''(\boldsymbol{I},\boldsymbol{h},\boldsymbol{\sigma}) &= \frac{1}{1+\exp\left(-2\sum_l \sigma_l h_l(\boldsymbol{I})\right)} \end{cases} \tag{5}$$

softmax denotes any soft (i.e. differentiable) approximation of the maximum function. Note that the noisy-OR function could be used, like in DD, but the generalized mean GM_r is recommended when the number of instances per bag is high [69]: $GM_r(\boldsymbol{v}) = (\sum_j v_j^r / \sum_j 1)^{\frac{1}{r}}$. The anomaly index of instance $\boldsymbol{I}$ is defined as $a'''(\boldsymbol{I},\boldsymbol{h}^*,\boldsymbol{\sigma}^*)$, where $(\boldsymbol{h}^*,\boldsymbol{\sigma}^*)$ denotes the optimal strong classifier. The anomaly index of bag B is defined as $p(B,\boldsymbol{h}^*,\boldsymbol{\sigma}^*)$.

F. Note on Anomaly Index Aggregators

It can be noted that different aggregators are used to combine instance anomaly indices into bag anomaly indices: the noisy-OR softmax function in DD, the exact maximum in APR, mi-SVM and MI-SVM, and the generalized mean

softmax function in MILBoost. Although those three aggregators could be used indifferently for testing, they cannot be used indifferently for training. We decided to use the same aggregator for training and testing. These aggregators are referred to as 'generalized max aggregators' and denoted by gmax.

G. Mammography-Specific Anomaly Index Aggregators

These generalized max aggregators combine local anomaly scores computed for all regions of all four images inside an examination record, regardless of which image they come from. The fact that two images represent the same breast under two different views (CC = craniocaudal and MLO = mediolateral-oblique) is ignored. However, anomalies are often visible in both views, so if one anomaly is only detected in one view, chances are that it is a false alarm. Consequently, an improved global anomaly score $A(B)$ is proposed for B, the bag associated with the examination record:

$$A(B) = \underset{l\in\{left,right\}}{\mathrm{gmax}} \; \underset{v\in\{CC,MLO\}}{\mathrm{gmin}} \; \underset{j}{\mathrm{gmax}}\, a(B_j^{lv}) \qquad (6)$$

where $a(B_j^{lv})$ is the local anomaly index of the j^{th} region in the $v \in \{CC, MLO\}$ view of the $l \in \{left, right\}$ breast. If gmax is the $\max$ function, gmin denotes the $\min$ function. If gmax is a $\mathrm{softmax}$ function, gmin denotes the $\mathrm{softmin}$ function associated with gmax: $1 - \mathrm{gmax}$ in the case of noisy-OR, GM_{-r} in the case of the generalized mean GM_r. The proposed aggregators are referred to as 'two-views aggregators'. For improved performance, these novel two-views aggregators are included into the learning procedure of each MIL algorithm, in place of the gmax function (see equations 2, 3, 4 and 5).

V. Experiments on the DDSM Dataset

These five MIL algorithms have been applied to anomaly detection in the Digital Database for Screening Mammography (DDSM), the largest publicly available mammography dataset, and compared to a strongly-supervised framework.

A. The DDSM Dataset

The Massachusetts General Hospital, the University of South Florida and the Sandia National laboratories have collected, for research purposes, a dataset of mammography examinations, called DDSM [3]. It contains mammography screening examinations conducted between October of 1988 and February of 1999 at four institutions: Massachusetts General Hospital (MGH), Wake Forest University School of Medicine (WFU), Sacred Heart Hospital (SH) and Washington University of St. Louis Medical Center (WU). Each examination record includes two images of each breast, corresponding to craniocaudal and mediolateral-oblique views, along with contextual data (patient age, acquisition and digitization dates, digitizer, etc.) and the radiologist's report. Three digitizers were used: DBA, HOWTEK, LUMISYS. Additionally, each mammogram containing lesions comes with a manual segmentation roughly delineating each finding (a mass or a MC cluster) [3].

1) Normal versus Abnormal Examination Records: for the purpose of this study, examination records were divided into two groups: 'abnormal' examination records, which correspond to examination records with at least one benign or one cancerous finding, and 'normal' examination records, which correspond to examination records with no findings. Statistics are reported in table I.

TABLE I: Statistics about the DDSM dataset

diagnosis	DBA	HOWTEK	LUMISYS
normal	430	183	82
abnormal	97	966	721
total	527	1149	803

2) Training and Test Subsets: for performance evaluation, this dataset needs to be partitioned into a training subset and a test subset: for each digitizer (DGA, HOWTEK, LUMISYS), one half of the examination records is assigned to the training subset, the other half is assigned to the test subset. Ten random training/test subset partitions were performed. The first partition was used to report optimal parameters and plot performance curves. All ten partitions were used for statistical comparisons between methods.

B. Performance of Mass and Microcalcification Detection

Manual segmentations from the DDSM training subset were used to evaluate the performance of mass and MC detection, as well as the performance of anomaly detection at local level (§V-C). In the DDSM dataset, a manually-segmented region of interest corresponds to one finding: either a mass or an MC Cluster (MCC). Let 'mass ROI' and 'MCC ROI' denote these regions of interest, respectively.

1) Mass Detection: Because each mass ROI contains a single mass, mass detection can be evaluated by a free-response ROC analysis: for various thresholds on the minimum nesting depth (§III-A), the percentage of mass ROIs intersecting a detection (i.e. the sensitivity) and the number of detections intersecting no mass ROI (i.e. the number of false positives) per pathological mammogram were determined. Here, an 'abnormal mammogram' refers to a mammogram from an 'abnormal' examination record. This experiment was performed for various numbers N_{mass} of topographic levels.

2) Microcalcification Detection: An MCC ROI often contains several MCs, so a different evaluation protocol was used to evaluate MC detection. A probability of presence was determined for each MCC ROI in the training subset. This probability was defined as the mean local probability inside the ROI (§III-A and Fig. 2 (d)). For comparison, the mean local probability inside the breast area (§II-B) but outside any MCC ROIs was computed for each abnormal mammogram in the training subset. An ROC curve was obtained by varying a threshold on the mean local probability. This experiment was performed for various sets $\mathcal{S}_{MC}$ of filter radii.

Results obtained using our own implementations of the mass and MC detectors are reported in Fig. 3.

C. Criteria for the Evaluation of Anomaly Detection

For anomaly detection, performance was assessed separately for each digitizer [9]. The reason is that the distribution of normal and abnormal samples is very different from one digitizer to another (see table I): if the system was able to recognize the specific digitization characteristics of each device, which may be possible since texture features are used, then a trivial classifier could be '$\{\text{DBA}\} \rightarrow$ normal, $\{\text{HOWTEK, LUMISYS}\} \rightarrow$ abnormal' (sensitivity: 61.9%, specificity: 94.6%). So, one anomaly detector was trained for each combination of MIL algorithm, digitizer and texture feature family (see table II).

At global level, the bag-level anomaly indices are evaluated against the anomaly scores assigned to mammography examinations as a whole (§V-A1). By varying a threshold on the bag-level anomaly indices, sensitivity and specificity are determined and a ROC curve is obtained (see Fig. 4 (a)-(c)). Performance at global level is defined as the area under this ROC curve. At local level, the instance-level indices are evaluated against manual segmentations. By varying a threshold on the instance-level indices, a curve is obtained that shows the average fraction of the breast areas above this threshold (in all four mammograms) versus the sensitivity, i.e. the average fraction of abnormal examination records in which all mass and MCC ROIs are above this threshold (see Fig. 4 (d)-(f)). Performance at local level is defined as the area under this curve. Ideally, all mass and MCC ROIs should be above the threshold while, on average, only a small fraction of the breast areas is above it.

D. Parameter Optimization

Several feature extraction parameters need to be tuned, regardless of the classifier used: Δ, the region size (§II-D), N_{mass} and L_{mass}, the mass detection parameters (§III-A), $\mathcal{S}_{MC}$ and L_{MC}, the MC detection parameters (§III-A), $Q_{texture}$, $\mathcal{S}_{texture}$ and $L_{texture}$, the texture description parameters (§III-B). For each training/test subset partition, parameters were tuned by two-fold cross validation to maximize the performance of anomaly detection at global level, averaged across algorithms and digitizers, in the training subset. The following parameter values were found optimal in the first training/test subset partition: Δ = 300 pixels, N_{mass} = 40, L_{mass} = 12 levels, $\mathcal{S}_{MC} = \{3, 4\}$, L_{MC} = 8 levels, $Q_{texture}$ = 24 levels, $\mathcal{S}_{texture} = \{2, 4\}$, $L_{texture}$ = 32 components.

Each classifier also has undetermined parameters: the number N_{DD} of gradient ascents for DD (§IV-B), the kernel density parameter in APR (§IV-C), the penalty parameter C and the scale parameter of kernel ϕ for mi-SVM, MI-SVM and SVM (§IV-D), the order r of the $\mathrm{softmax}$ function for MILBoost (§IV-E). For each combination of training/test subset partition, digitizer and texture family, the classifier parameters were tuned by two-fold cross validation to maximize the performance of anomaly detection at global level in the training subset.

E. Baseline Methods

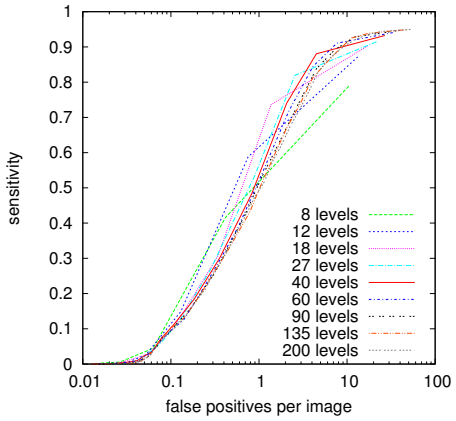
1) Strongly-Supervised Version of this Framework: to assess the relevance of MIL, the standard, strongly-supervised SVM classifier [67] was also evaluated. Of course, this classifier was not trained using global labels: it was trained using manual segmentations. To train this classifier, a given instance was assigned a positive label if and only if the corresponding region intersects a mass ROI or an MCC ROI (§V-B). The class imbalance problem was managed like in multiple-instance SVM (§IV-D). Instance and bag anomaly indices were also defined as in multiple-instance SVM.

2) Other Mammography CADe Systems: the proposed framework was also compared to two mammography CAD systems: 1) a mass detection and false positive reduction system by Kim et al. [27] and 2) a MC and MC cluster detector by Oliver et al. [71]. In Kim's system, candidate masses are segmented using a topographic approach. A heterogeneous feature vector (texture, shape, intensity, spiculation) is then extracted from each candidate. Next,

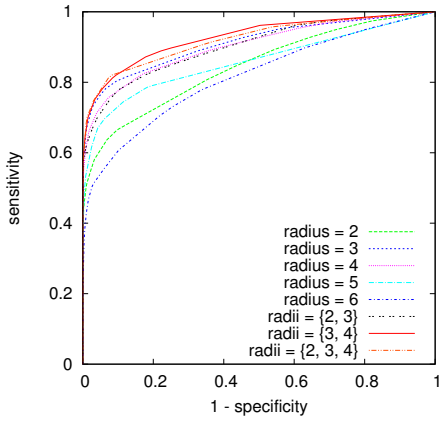
TABLE II: Anomaly detection performance — average area under the curve and standard error

supervision	algorithm	level	aggregator	no texture	Haralick	GLCM	GLRLM	LBPH
strong	SVM	global		0.740 ± 0.009	0.745 ± 0.007	**0.758 ± 0.008**	0.749 ± 0.008	0.741 ± 0.007
		local		0.898 ± 0.008	0.912 ± 0.008	**0.917 ± 0.007**	0.914 ± 0.008	0.901 ± 0.009
weak	DD	global	gmax	0.761 ± 0.007	0.770 ± 0.008	0.785 ± 0.007	**0.791 ± 0.007**	0.757 ± 0.008
		global	two-views	0.773 ± 0.006	0.786 ± 0.008	**0.804 ± 0.007**	0.803 ± 0.007	0.770 ± 0.008
		local	two-views	0.905 ± 0.008	0.917 ± 0.007	0.916 ± 0.008	**0.925 ± 0.007**	0.899 ± 0.008
	APR	global	gmax	**0.712 ± 0.010**	0.696 ± 0.009	0.707 ± 0.008	0.703 ± 0.011	0.704 ± 0.009
		global	two-views	**0.721 ± 0.009**	0.703 ± 0.010	0.713 ± 0.007	0.702 ± 0.012	0.712 ± 0.010
		local	two-views	0.820 ± 0.009	0.812 ± 0.011	0.792 ± 0.008	0.793 ± 0.006	**0.821 ± 0.011**
	mi-SVM	global	gmax	0.751 ± 0.008	0.760 ± 0.009	0.764 ± 0.008	**0.765 ± 0.007**	0.752 ± 0.007
		global	two-views	0.769 ± 0.007	0.775 ± 0.007	**0.784 ± 0.008**	0.783 ± 0.007	0.763 ± 0.008
		local	two-views	0.912 ± 0.008	0.915 ± 0.008	0.930 ± 0.007	**0.936 ± 0.007**	0.909 ± 0.008
	MI-SVM	global	gmax	0.759 ± 0.009	0.765 ± 0.007	**0.772 ± 0.007**	0.769 ± 0.007	0.765 ± 0.008
		global	two-views	0.777 ± 0.008	0.780 ± 0.008	**0.800 ± 0.007**	0.785 ± 0.009	0.779 ± 0.007
		local	two-views	0.896 ± 0.006	**0.912 ± 0.008**	0.908 ± 0.008	0.907 ± 0.008	0.904 ± 0.008
	MILBoost	global	gmax	0.723 ± 0.010	0.716 ± 0.009	**0.742 ± 0.008**	0.742 ± 0.010	0.733 ± 0.007
		global	two-views	0.741 ± 0.008	0.730 ± 0.008	0.755 ± 0.008	**0.759 ± 0.007**	0.743 ± 0.008
		local	two-views	0.815 ± 0.011	**0.837 ± 0.009**	0.835 ± 0.011	0.832 ± 0.010	0.825 ± 0.010

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2521442, IEEE Transactions on Medical Imaging



(a) mass detection



(b) MC detection

Fig. 3: Performance of lesion detection in the DDSM training subset. Fig. (a) reports the detection performance of masses against manual segmentations, for different numbers of topographic levels. Fig. (b) reports the ability to discriminate manually segmented MCC ROIs from the rest of the breast area, for different sets of filter radii.

each candidate is classified as 'mass' or 'normal tissue' using a Sparse Representation Classifier (SRC). In Oliver's system, local features are extracted using a bank of filters. Next, a boosted classifier (Gentleboost) is used to identify individual MCs. Finally, MC clusters are identified through a local integration of MC probabilities.

F. Performance of Anomaly Detection

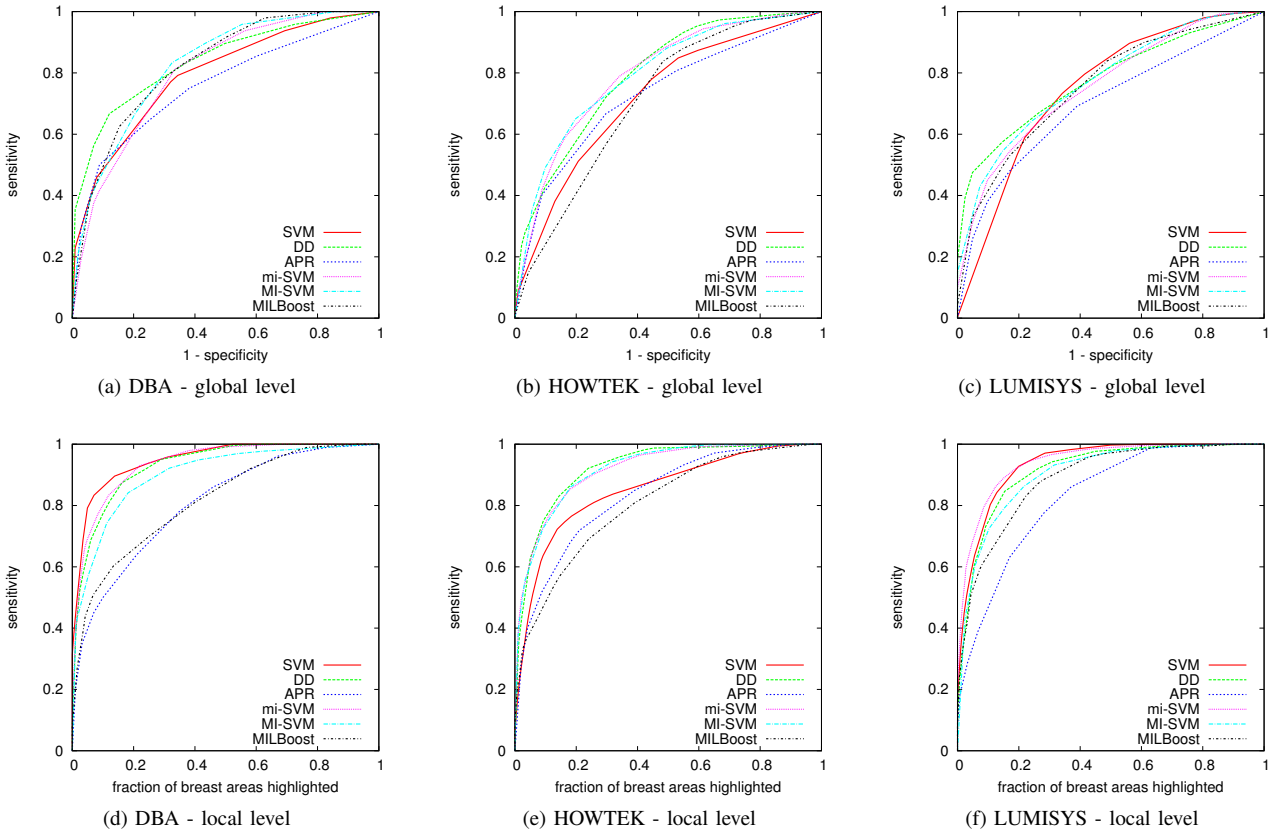
Performance has been evaluated at global and local level, using either the gmax aggregators or the two-views aggregators. In each experiment, a single texture family was used, together with the lesion detection features and the region coordinates. For each training/test subset partition, one score was obtained per digitizer (§V-C) and the average $\bar{s}$ of these three scores was computed. Table II reports the mean of $\bar{s}$, as well as its standard error, over the ten partitions. For the first partition, the best global ROC curve and the corresponding local performance curve (§V-C) are reported in Fig. 4. Comparisons with Oliver et al. and Kim et al. are reported in table III. For a fair comparison with Kim et al., which only detect masses, all examination records containing MCC ROIs have been discarded during the comparison. Similarly, all examination records containing mass ROIs have been discarded during the comparison with Oliver et al. During those comparisons, the two-views aggregators were used with the best texture feature families (bolded in table II). Detection examples, obtained with mi-SVM and GLRLM features, are given in Fig. 5.

VI. Discussion

A novel computer-aided detection and diagnosis system for mammography examinations was presented in this paper. A novel strategy for partitioning the breast adaptively into regions was proposed. In each region, features derived from the detection of masses and microcalcifications (MCs), as well as texture features, are extracted. Then, a Multiple-Instance Learning (MIL) algorithm (DD, APR, mi-SVM, MI-SVM or MILBoost) is used to recognize abnormal regions and map local anomaly labels to a global label ("normal examination record" or "abnormal examination record") using a proposed mammography-specific aggregator. Supervision does not rely on manual segmentations, but simply on global labels assigned to examination records as a whole. This weakly-supervised framework is compared to a strongly-supervised framework where a standard SVM classifier is used to map each feature vector to a local label, and ultimately a global label; in that framework, supervision relies on manual segmentations of lesions. These two frameworks are compared in the Digital Database for Screening Mammography (DDSM).

At global level, DD clearly is the best MIL classifier for anomaly detection, followed by MI-SVM and mi-SVM (see table II). These three MIL algorithms outperform the strongly-supervised SVM classifier ($p < 10^{-5}$, $p < 10^{-5}$ and $p = 9.98 \times 10^{-4}$, respectively). Because the DDSM dataset is rather large, we did expect MIL algorithms to reach the performance of a strongly-supervised algorithm, which is a very positive result in itself: it means that we do not need experts to manually segment images, a task that they usually do not perform in their daily practice. However, we did not expect these algorithms to outperform the strongly-supervised algorithm. One reason why MIL algorithms outperform a strongly-supervised classifier at global level may be that the general aspect of images is affected somehow by cancer. Typically, the texture of breast tissues may change even before MCs and masses appear. By design, these early modifications of tissues cannot be captured by a strongly-supervised classifier. MIL algorithms, which are not constrained by expert segmentations, can find other ways to recognize pathological images. We have shown that the proposed two-views aggregator, used both for training and testing, outperforms the standard generalized maximum (e.g. $p = 0.00665$ using DD and GLCM). At local level, mi-SVM outperforms DD, but not

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2521442, IEEE Transactions on Medical Imaging



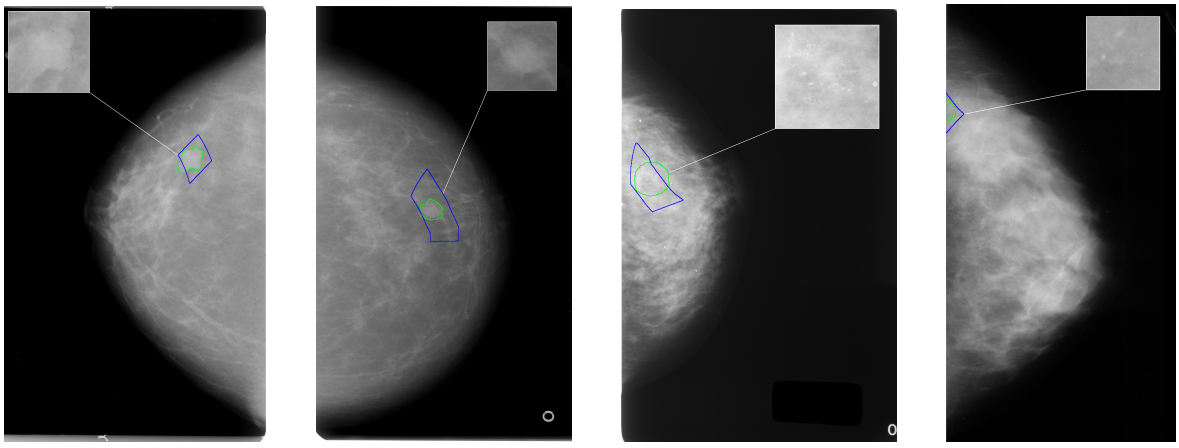
(a) DBA - global level (b) HOWTEK - global level (c) LUMISYS - global level

(d) DBA - local level (e) HOWTEK - local level (f) LUMISYS - local level

Fig. 4: Anomaly detection performance using the two-views aggregators

TABLE III: Mass and MCC detection performance — average area under the curve and standard error

mass detection	texture features	DD	APR	mi-SVM	MI-SVM	MILBoost	Kim et al. [27]
	global level	**0.809 ± 0.007**	0.723 ± 0.008	0.790 ± 0.007	0.806 ± 0.008	0.758 ± 0.007	0.795 ± 0.008
	local level	0.932 ± 0.007	0.835 ± 0.010	**0.943 ± 0.006**	0.924 ± 0.007	0.835 ± 0.009	0.948 ± 0.007
MCC detection	texture features	DD	APR	mi-SVM	MI-SVM	MILBoost	Oliver et al. [71]
	global level	**0.800 ± 0.007**	0.723 ± 0.008	0.779 ± 0.008	0.798 ± 0.008	0.749 ± 0.007	0.782 ± 0.007
	local level	0.912 ± 0.009	0.803 ± 0.011	**0.923 ± 0.008**	0.907 ± 0.008	0.830 ± 0.008	0.905 ± 0.007



(a) segmented mass (b) segmented mass (c) segmented MCC (d) segmented MCC

Fig. 5: Anomaly detection examples using the mi-SVM detector and GLRLM features. The most abnormal region (in Fig. a and d) or regions (in Fig. b and c) are in blue; manual segmentations from DDSM are in green. Simply delineating regions may not be optimal in a GUI: local probability maps (see Fig. 2) may also be used for improved visualization.

significantly ($p = 0.121$). More importantly, it outperforms strongly-supervised SVM ($p = 7.45 \times 10^{-3}$). Because we compared SVM to mi-SVM and MI-SVM, the direct MIL generalizations of SVM, using the same SVM engine (libSVM[1]), the superiority of MIL over strong supervision cannot be attributed to algorithmic choices.

When the system is trained to focus on a single type of anomalies, masses or MCs, comparisons with other frameworks are possible (see table III). The proposed framework is comparable to Kim et al.'s method for mass detection: it is slightly better at global level ($p = 0.0519$ using DD) and slightly worse at local level ($p = 0.428$ using mi-SVM). The advantages over Kim et al.'s framework are generality in terms or target lesion and, more importantly, weak supervision. The proposed framework outperforms Oliver et al.'s method for MCC detection both at global and local level ($p = 0.0124$ using DD and $p = 0.0263$ using mi-SVM, respectively).

Regarding texture, the optimal family of texture features changes a lot from one digitizer or one classifier to another. Although there is no clear winner, the GLCM and GLRLM features outperform Haralick and LBPH features in most scenarios. Of course, it is possible to combine several families of texture features to push performance further. Other improvements are possible: in particular, for simplicity, the proposed framework was limited to a binary risk assessment ("normal" versus "abnormal"). This is because MIL classifiers, and classifiers in general, are usually designed for two class problems. But the proposed framework may be generalized to multiple class problems (e.g. "normal" versus "benign" versus "cancer").

To conclude, in line with our works on MIL for medical decision support, we have reported a computer-aided detection and diagnosis system for mammography screening. This study shows that MIL can help detect abnormal mammography examination records and locate anomalies within those records, which may be useful for attention focusing during mammography screening.

VII. Acknowledgments

This work was supported in part by a grant from the Brittany Region, through the ID2Santé program (DéCaRICA), and in part by a grant from the French *Agence Nationale de la Recherche* (ANR, LabCom SePEMeD).

References

[1] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. Saripan, A. R. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clin Imaging*, vol. 37, no. 3, pp. 420–6, May–Jun 2013.

[2] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K. H. Ng, "Computer-aided breast cancer detection using mammograms: a review," *IEEE Rev Biomed Eng*, vol. 6, pp. 77–98, 2013.

[3] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Proc IWDM*, M. J. Yaffe, Ed., 2001, pp. 212–8.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[4] M. A. Mazurowski, P. A. Habas, J. M. Zurada, and G. D. Tourassi, "Decision optimization of case-based computer-aided decision systems using genetic algorithms with application to mammography," *Phys Med Biol*, vol. 53, no. 4, pp. 895–908, Feb. 2008.

[5] D. D. Costa, L. F. Campos, and A. K. Barros, "Classification of breast tissue in mammograms using efficient coding," *Biomed Eng Online*, vol. 10, no. 1, p. 55, Jun. 2011.

[6] A. Hapfelmeier and A. Horsch, "Image feature evaluation in two new mammography CAD prototypes," *Int J Comput Assist Radiol Surg*, vol. 6, no. 6, pp. 721–35, Nov. 2011.

[7] M. Bator and M. Nieniewski, "Detection of cancerous masses in mammograms by template matching: optimization of template brightness distribution by means of evolutionary algorithm," *J Digit Imaging*, vol. 25, no. 1, pp. 162–72, Feb. 2012.

[8] N. Gargouri, A. Dammak Masmoudi, D. Sellami Masmoudi, and R. Abid, "A new GLLD operator for mass detection in digital mammograms," *Int J Biomed Imaging*, vol. 2012, p. 765649, 2012.

[9] A. García-Manso, C. J. García-Orellana, H. González-Velasco, R. Gallardo-Caballero, and M. Macías-Macías, "Consistent performance measurement of a system to detect masses in mammograms based on blind feature extraction," *Biomed Eng Online*, vol. 12, p. 2, 2013.

[10] D. C. Pereira, R. P. Ramos, and M. Z. do Nascimento, "Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm," *Comput Methods Programs Biomed*, vol. 114, no. 1, pp. 88–101, Apr. 2014.

[11] F. Soares Sérvulo de Oliveira, A. Oseas de Carvalho Filho, A. Corrêa Silva, A. Cardoso de Paiva, and M. Gattass, "Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and SVM," *Comput Biol Med*, vol. 57, pp. 42–53, Feb. 2015.

[12] X. Lladó, A. Oliver, J. Freixenet, R. Martí, and J. Martí, "A textural approach for mass false positive reduction in mammography," *Comput Med Imaging Graph*, vol. 33, no. 6, pp. 415–22, Sep. 2009.

[13] M. Hussain, "False positive reduction using Gabor feature subset selection," in *Proc ICISA*, Jun. 2013, pp. 1–5.

[14] G. Braz Junior, S. Vieira da Rocha, M. Gattass, A. Corrêa Silva, and A. Cardoso de Paiva, "A mass classification using spatial diversity approaches in mammography images for false positive reduction," *Expert Syst Appl*, vol. 40, no. 18, pp. 7534–43, Dec. 2013.

[15] I. Zyout, J. Czajkowska, and M. Grzegorzek, "Multi-scale textural feature extraction and particle swarm optimization based model selection for false positive reduction in mammography," *Comput Med Imaging Graph*, Feb. 2015.

[16] C.-C. Jen and S.-S. Yu, "Automatic detection of abnormal mammograms in mammographic images," *Expert Syst Appl*, vol. 42, no. 6, pp. 3048–55, Apr. 2015.

[17] J. Liu, J. Chen, X. Liu, L. Chun, J. Tang, and Y. Deng, "Mass segmentation using a combined method for cancer detection," *BMC Syst Biol*, vol. 5 Suppl 3, p. S6, 2011.

[18] R. Gallardo-Caballero, C. J. García-Orellana, A. García-Manso, H. M. González-Velasco, and M. Macías-Macías, "Independent component analysis to detect clustered microcalcification breast cancers," *Sci World J*, vol. 2012, p. 540457, 2012.

[19] N. S. Arikidis, S. Skiadopoulos, A. Karahaliou, E. Likaki, G. Panayiotakis, and L. Costaridou, "B-spline active rays segmentation of microcalcifications in mammography," *Med Phys*, vol. 35, no. 11, pp. 5161–71, Nov. 2008.

[20] I. I. Andreadis, G. M. Spyrou, and K. S. Nikita, "A CADx scheme for mammography empowered with topological information from clustered microcalcifications' atlases," *IEEE J Biomed Health Inform*, vol. 19, no. 1, pp. 166–73, Jan. 2015.

[21] S. Yoon and S. Kim, "AdaBoost-based multiple SVM-RFE for classification of mammograms in DDSM," *BMC Med Inform Decis Mak*, vol. 9 Suppl 1, p. S1, 2009.

[22] S. K. Biswas and D. P. Mukherjee, "Recognizing architectural distortion in mammogram: a multiscale texture modeling approach with GMM," *IEEE Trans Biomed Eng*, vol. 58, no. 7, pp. 2023–30, Jul. 2011.

[23] Y. Zhang, N. Tomuro, J. Furst, and D. S. Raicu, "Building an ensemble system for diagnosing masses in mammograms," *Int J Comput Assist Radiol Surg*, vol. 7, no. 2, pp. 323–9, Mar. 2012.

[24] C.-H. Wei, S. Y. Chen, and X. Liu, "Mammogram retrieval on similar mass lesions," *Comput Methods Programs Biomed*, vol. 106, no. 3, pp. 234–48, Jun. 2012.

[25] A. Vadivel and B. Surendiran, "A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories," *Comput Biol Med*, vol. 43, no. 4, pp. 259–67, May 2013.

[26] D. C. Moura and M. A. Guevara López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *Int J Comput Assist Radiol Surg*, vol. 8, no. 4, pp. 561–74, Jul. 2013.

[27] D. H. Kim, S. H. Lee, and Y. M. Ro, "Mass type-specific sparse representation for mass classification in computer-aided detection on mammograms," *Biomed Eng Online*, vol. 12 Suppl 1, p. S3, 2013.

[28] Y. A. Reyad, M. A. Berbar, and M. Hussain, "Comparison of statistical, LBP, and multi-resolution analysis features for breast mass classification," *J Med Syst*, vol. 38, no. 9, p. 100, Sep. 2014.

[29] X.-S. Zhang, "A new approach for clustered MCs classification with sparse features learning and TWSVM," *Sci World J*, vol. 2014, p. 970287, 2014.

[30] S. Sharma and P. Khanna, "Computer-aided diagnosis of malignant mammograms using Zernike moments and SVM," *J Digit Imaging*, vol. 28, no. 1, pp. 77–90, Feb. 2015.

[31] J. E. E. de Oliveira, A. de Albuquerque Araújo, and T. M. Deserno, "Content-based image retrieval applied to BI-RADS tissue classification in screening mammography," *World J Radiol*, vol. 3, no. 1, pp. 24–31, Jan. 2011.

[32] R. Nithya and B. Santhi, "Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer," *Int J Comput Appl*, vol. 28, no. 6, pp. 21–5, Aug. 2011.

[33] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Wavelet optimization for content-based image retrieval in medical databases," *Med Image Anal*, vol. 14, no. 2, pp. 227–41, Apr 2010.

[34] ——, "Adaptive nonseparable wavelet transform via lifting and its application to content-based image retrieval," *IEEE Trans Image Process*, vol. 19, no. 1, pp. 25–35, Jan. 2010.

[35] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications - clinical benefits and future directions," *Int J Med Inform*, vol. 73, no. 1, pp. 1–23, Feb 2004.

[36] J. Amores, "Multiple instance classification: review, taxonomy and comparative study," *Artif Intell*, vol. 201, pp. 81–105, Aug 2013.

[37] G. Quellec, M. Lamard, M. D. Abràmoff, E. Decencière, B. Lay, A. Erginay, B. Cochener, and G. Cazuguel, "A multiple-instance learning framework for diabetic retinopathy screening," *Med Image Anal*, vol. 16, no. 6, pp. 1228–40, Aug 2012.

[38] G. Quellec, M. Lamard, B. Cochener, and G. Cazuguel, "Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials," *IEEE Trans Med Imaging*, vol. 34, no. 4, pp. 877–87, Apr 2015.

[39] M. M. Dundar, G. Fung, B. Krishnapuram, and R. B. Rao, "Multiple-instance learning algorithms for computer-aided detection," *IEEE Trans Biomed Eng*, vol. 55, no. 3, pp. 1015–21, Mar. 2008.

[40] M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: A benchmarking study," *Comput Med Imaging Graph*, vol. 42, pp. 44–50, Jun 2015.

[41] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Med Image Anal*, vol. 18, no. 3, pp. 591–604, Apr. 2014.

[42] M. Kandemir, C. Zhang, and F. A. Hamprecht, "Empowering multiple instance histopathology cancer diagnosis by cell graphs," in *Proc MICCAI*, vol. 17, no. Pt 2, 2014, pp. 228–35.

[43] Y. Xu, J. Zhang, E. I.-C. Chang, M. Lai, and Z. Tu, "Context-constrained multiple instance learning for histopathology image segmentation," in *Proc MICCAI*, vol. 15, no. Pt 3, 2012, pp. 623–30.

[44] R. Venkatesan, P. Chandakkar, B. Li, and H. K. Li, "Classification of diabetic retinopathy images using multi-class multiple-instance learning based on color correlogram features," in *Proc IEEE EMBS*, vol. 2012, 2012, pp. 1462–5.

[45] J. Ding, H. D. Cheng, J. Huang, J. Liu, and Y. Zhang, "Breast ultrasound image classification based on multiple-instance learning," *J Digit Imaging*, vol. 25, no. 5, pp. 620–7, Oct. 2012.

[46] M. T. McCann, R. Bhagavatula, M. C. Fickus, J. A. Ozolek, and J. Kovacevic, "Automated colitis detection from endoscopic biopsies as a tissue screening tool in diagnostic pathology," in *Proc ICIP*, vol. 2012, 2012, pp. 2809–12.

[47] H. Jiang, R. Zheng, D. Yi, and D. Zhao, "A novel multi-instance learning approach for liver cancer recognition on abdominal CT images based on CPSO-SVM and IO," *Comput Math Methods Med*, vol. 2013, p. 434969, 2013.

[48] T. Tong, R. Wolz, Q. Gao, J. V. Hajnal, and D. Rueckert, "Multiple instance learning for classification of dementia in brain MRI," in *Proc MICCAI*, vol. 16, no. Pt 2, 2013, pp. 599–606.

[49] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, D. Rueckert, and Alzheimers Disease Neuroimaging Initiative, "Multiple instance learning for classification of dementia in brain MRI," *Med Image Anal*, vol. 18, no. 5, pp. 808–18, Jul. 2014.

[50] J. C. Azar, M. Simonsson, E. Bengtsson, and A. Hast, "Automated classification of glandular tissue by statistical proximity sampling," *Int J Biomed Imaging*, vol. 2015, p. 943104, 2015.

[51] J. Melendez, B. van Ginneken, P. Maduskar, R. H. H. M. Philipsen, K. Reither, M. Breuninger, I. M. O. Adetifa, R. Maane, H. Ayles, and C. I. Sánchez, "A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays," *IEEE Trans Med Imaging*, vol. 34, no. 1, pp. 179–92, Jan. 2015.

[52] C. Li, C. Shi, H. Zhang, Y. Chen, and S. Zhang, "Multiple instance learning for computer aided detection and diagnosis of gastric cancer with dual-energy CT imaging," *J Biomed Inform*, pp. 358–68, Aug. 2015.

[53] B. Krishnapuram, J. Stoeckel, V. Raykar, B. Rao, P. Bamberger, E. Ratner, N. Merlet, I. Stainvas, M. Abramov, and A. Manevitch, "Multiple-instance learning improves CAD detection of masses in digital mammography," in *Proc IWDM*, vol. 5116, 2008, pp. 350–357.

[54] P. Lu, W. Liu, W. Xu, L. Li, B. Zheng, J. Zhang, and L. Zhang, "Multi-instance learning for mass retrieval in digitized mammograms," in *Proc SPIE Medical Imaging*, vol. 8315, 2012, p. 831523.

[55] C. Li, K. M. Lam, L. Zhang, C. Hui, and S. Zhang, "Mammogram microcalcification cluster detection by locating key instances in a multi-instance learning framework," in *Proc IEEE ICSPCC*, 2012, pp. 175–9.

[56] C. R. Maurer Jr., R. Qi, and V. Raghavan, "A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions," *IEEE Trans Pattern Anal Mach Intell*, vol. 25, no. 2, pp. 265–70, Feb 2003.

[57] B. W. Hong and B. S. Sohn, "Segmentation of regions of interest in mammograms in a topographic approach," *IEEE Trans Inf Technol Biomed*, vol. 14, no. 1, pp. 129–39, Jan 2010.

[58] G. Loy and A. Zelinsky, "Fast radial symmetry for detecting points of interest," *IEEE Trans Pattern Anal Mach Intell*, vol. 25, no. 8, pp. 959–73, Aug 2003.

[59] R. M. Haralick, K. Shanmugam, and I. D. Dinstein, "Textural features for image classification," *IEEE Trans Syst Man Cybern*, vol. SMC-3, no. 6, pp. 610–21, Nov 1973.

[60] M. M. Galloway, "Texture analysis using gray level run lengths," *Comput Graph Image Process*, vol. 4, no. 2, pp. 172–9, Jun 1975.

[61] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit*, vol. 29, no. 1, pp. 51–9, Jan 1996.

[62] J. Wang and J.-D. Zucker, "Solving the multiple-instance problem: a lazy learning approach," in *Proc ICML*, 2000, pp. 1119–25.

[63] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-I.I.D. samples," in *Proc ICML*, 2009, pp. 1249–56.

[64] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc NIPS*, vol. 10. MIT Press, 1998, pp. 570–6.

[65] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif Intell*, vol. 89, no. 12, pp. 31–71, Jan 1997.

[66] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc NIPS*, vol. 15. MIT Press, 2003, pp. 561–8.

[67] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–97, 1995.

[68] P. Viola, J. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Proc NIPS*, vol. 18. MIT Press, 2006, pp. 1417–24.

[69] B. Babenko, P. Dollár, Z. Tu, and S. Belongie, "Simultaneous learning and alignment: multi-instance and multi-pose learning," in *Proc ECCV*, 2008.

[70] Y. Xu, J. Zhang, E. Chang, M. Lai, and Z. Tu, "Contexts-constrained multiple instance learning for histopathology image analysis," in *Proc MICCAI*, vol. 7512, 2012, pp. 623–30.

[71] A. Oliver, A. Torrent, X. Llad, M. Tortajada, L. Tortajada, M. Sents, J. Freixenet, and R. Zwiggelaar, "Automatic microcalcification and cluster detection for digital and digitised mammograms," *Knowledge-Based Systems*, vol. 28, pp. 68–75, Apr. 2012.