Medical Image Analysis 51 (2019) 116–124

Contents lists available at ScienceDirect

Medical Image Analysis

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

A graph-based lesion characterization and deep embedding approach for improved computer-aided diagnosis of nonmass breast MRI lesions

Cristina Gallego-Ortiz[a,b,*], Anne L. Martel[a,b]

[a] Department of Medical Biophysics, University of Toronto, Canada
[b] Department of Imaging Research, Sunnybrook Research Institute, Toronto, Canada

ARTICLE INFO

Article history:
Received 3 May 2018
Revised 28 September 2018
Accepted 30 October 2018
Available online 2 November 2018

Keywords:
Computer-aided diagnosis (CAD)
Breast MRI
Nonmass enhancement
Dimensionality reduction
Embedded feature space

ABSTRACT

Nonmass-like enhancements are a common but diagnostically challenging finding in breast MRI. Nonmass-like lesions can be described as clusters of spatially and temporally inter-connected regions of enhancements, so they can be modeled as networks and their properties characterized via network-based connectivity. In this work, we represented nonmass lesions as graphs using a link formation energy model that favors linkages between regions of similar enhancement and closer spatial proximity. However, adding graph features to an existing computer-aided diagnosis (CAD) pipeline incurs an increase of feature space dimensionality, which poses additional challenges to traditional supervised machine learning techniques due to the inability to increase accordingly the number of training datasets. We propose the combination of unsupervised dimensionality reduction and embedded space clustering followed by a supervised classifier to improve the performance of a CAD system for nonmass-like lesions in breast MRI. Our work extends a previoulsy proposed framework for deep embedded unsupervised clustering (DEC) to embedding space classification, with the joint optimization of objective functions for DEC and supervised multi-layered perceptron (MLP) classification. The strength of the method lies in the ability to learn and further optimize an embedded feature representation of lower dimensionality that maximizes the diagnostic accuracy of a CAD lesion classifier to discriminate between benign and malignant lesions. We identified 792 nonmass-like enhancements (267 benign, 110 malignant and 415 unknown) in 411 patients undergoing breast MRI at our institution. The diagnostic performance of the proposed method was evaluated and compared to the performance of a conventional supervised MLP classifier in original feature space. A statistically significant increase in diagnostic area under the ROC curve (AUC) was achieved. Generalization AUC increased from 0.67 ± 0.08 to 0.81 ± 0.10 (21% increase, p-value$= 4.2 \times 10^{-8}$) with the proposed graph-based lesion characterization and deep embedding framework.

1. Introduction

The Breast Imaging Reporting and Data System (BI-RADS) lexicon (Morris et al., 2013) defines a nonmass-like enhancement in breast MR Imaging as an area of enhancement distinct from the surrounding parenchyma, that is not a space-occupying mass or a focus (<5-mm area of enhancement). A nonmass-like enhancement is further characterized by its specific distribution within the breast as well as its internal enhancement pattern. The distribution can be defined as focal, linear, segmental, regional, multi-regional, or diffuse, and internal enhancement patterns can be homogeneous, heterogeneous, clumped or clustered ring (Tozaki and Fukuda, 2006). Although a nonmass-like finding is most commonly due to benign fibrocystic changes, it can also be a sign of intraductal or diffuse breast cancer. In addition, healthy breast tissue enhancement known as background parenchymal enhancement (BPE) can be hard to distinguish from nonmass-like abnormal lesion enhancement. The analysis of quantitiative tissue architecture in biomedical imaging has been attempted before.

Accurate interpretation of nonmass-like enhancement findings remains a challenge in breast MR imaging (Lourenco et al., 2014; Shao et al., 2013) Several studies (Baltzer et al., 2010; Gutierrez et al., 2009; Tozaki and Fukuda, 2006) have correlated nonmass-like distribution and internal enhancement descriptors with pathological outcomes with conflicting results. A branching pattern of enhancement has been significantly associated with a higher probability of malignancy (Machida et al., 2015; Tozaki and Fukuda, 2006) while heterogeneous internal enhancement and clustered ring distribution have been associated with a high positive predictive value (PPV) for cancer

[*] Corresponding author.
E-mail address: c.gallego@mail.utoronto.ca (C. Gallego-Ortiz).

https://doi.org/10.1016/j.media.2018.10.011
1361-8415/© 2018 Elsevier B.V. All rights reserved.

(Tozaki and Fukuda, 2006). However, other studies have found that BI-RADS descriptors of nonmass-like lesions are not significant predictors of malignancy (Gutierrez et al., 2009). High inter-observer variability among radiologists has also been linked to controversial results (Giess et al., 2013). Computer-aided diagnosis (CAD) technologies that produce standardized descriptors of breast lesions have been proposed to aid interpretation. Recently, it was found that inter-reader and intra-reader agreement improved significantly when a semi-automatic software was used to evaluate background parenchymal enhancement (BPE) (Tagliafico et al., 2015). Similarly, CAD can help improve diagnostic specificity and decrease false-positive interpretations by identifying nonmass-like enhancements where follow-up surveillance is more appropriate than biopsy due to their low risk for upgrade to cancer (Bahl et al., 2018).

Unlike mass-like enhancing lesions, nonmass-like lesions have no well-defined boundaries, limiting their analysis with typical computer extracted morphology features. Instead, nonmass-like lesions can be described as clusters of spatially and temporally interconnected regions of enhancements (Thomassin-Naggara et al., 2009). These clusters can be modeled as networks and their properties characterized by graph-based connectivity features. Some original literature exists on the use of structural features based on the Voronoi Diagram (VD) and its subgraphs applied to diagnostic pathology images (Sudbø et al., 2000). This work found that structural features were able to distinguish structurally between normal and cancerous oral mucosa, and between good and poor outcome groups in prostatic and cervical carcinomas. Another more recent study proposed the use of the minimum spanning tree (MST) of graphs connecting epithelial nuclei for the diagnosis and prognosis of prostate cancer (Khan et al., 2017). We are not aware a formal graph modeling approach being applied to lesions of the breast in breast MRI CAD lesion characterization. Our work was motivated by previous work in histopathology and how it could be extended to medical imaging characterization.

Intuitively, enhancements following branching distributions will have associated graphs and connectivity metrics that differ from graphs representing linearly distributed enhancements. However, adding graph features to a CAD pipeline results in an increase of feature space dimensionality. The increase in feature space size poses a challenge: the curse of dimensionality occurs when the number of features becomes very large relative to the number of observations, and when further increasing the number of observations is not feasible. To avoid the curse of dimensionality reducing the size of the resulting feature space without losing representation ability is imperative.

The objectives of this study were twofold: First, to investigate if graph-based features are a suitable representation of nonmass-like lesions and to determine if they can lead to improved diagnostic performance of CAD for discrimination of benign and malignant breast lesions. Second, to establish whether the resulting high-dimensional feature space from graph-connectivity features combined with other standard DEC-MRI features could be effectively reduced to a lower-dimensional space and optimized for improved discrimination between benign and malignant nonmass-like lesions. Our approach uses Deep-Embedded Clustering or (DEC) (Xie et al., 2016); a clustering-based dimensionality reduction framework capable of learning an embedded feature space representation optimized for feature-based data discrimination. The intuition behind our proposed framework is that if the lower-dimensional representation improves the separation between lesion classes, then the diagnostic performance of CAD in embedded space naturally improves.

2. Materials

2.1. Breast MR imaging and nonmass-like enhancements

All women underwent dynamic contrast-enhanced (DCE)-MRI on a 1.5-T magnet (Signa, General Electric Medical Systems, Milwaukee, Wis) using a dedicated breast coil. The dynamic protocol comprised one pre-contrast and four post-contrast acquisitions using a bolus injection of 0.1 mmol/kg of Gadolinium contrast agent Gadovist ®injection at 2 cc/s, 20 s delay. To ensure that motion between breast DCE-MRI scans was corrected, a pre-processing pipeline that registered each post-contrast scan to the baseline or pre-contrast scan was used (Martel et al., 2007). The location reported by the radiologist at the time of evaluation was used to identify the scan slice and the region of interest (ROI) in which the nonmass-like enhancement appeared for further processing. Each ROI consisted of the 2D bounding box that encompassed the diameter of the lesion, or a cross-sectional length that was used as the diagonal of the bounding box. The lesion inside the identified ROI was further automatically segmented using the deep Artificial Neural Network (ANN) algorithm previously developed in our research group. The lesion segmentation algorithm was extensively validated in the work of a master's thesis in (Wu, 2016). It's important to note that the processing pipeline is fully automated but requires the initial selection of the ROI bounding box. Analyzing lesions previously found by the radiologist is the paradigm of computer-aided diagnostic (CADx) systems and our proposed methodology agrees well with this paradigm by providing supplemental information to support the differential diagnosis.

We retrospectively reviewed consecutive clinical breast MRI studies conducted at our institution between 2011 and 2015 to identify reported nonmass-like enhancement findings. Ground truth histopathology diagnosis was classified as benign or malignant based on histopathology procedures performed no more than three months after breast MRI imaging, or unknown if no procedure was performed. Between 2011 and 2015 we identified 306 women with 629 nonmass-like findings, in which a subset ($n = 214$, benign=166, malignant=48) had a corresponding subsequent histopathology assessment, but the majority ($n = 415$) did not warranted biopsy at breast MRI imaging. From these, follow up imaging was available in 345 lesions, and 86 enhancement findings were stable for up to 2-years. To increase the number of histopathology confirmed lesions, we included other 163 histopathology proven lesions (101 benign, 62 malignant) that appeared as nonmass-like enhancement in breast MRI conducted in our institution before 2011. Total breast lesions appearing as nonmass-like enhancements included in this study consisted of 792 nonmass-like enhancements (267 benign, 110 malignant and 415 unknown) identified in 411 women aged 46.7 ± 11 (mean ± std) years.

3. Methodology

3.1. A graph model for nonmass-like enhancements

The concept of a network (or graph) was considered suitable for representing nonmass-like enhancements. We can build graphs that represent inter-connected regions of enhancement as nodes and use links to associate spatially and temporally connected regions. While physiologically, only proximal tissues share the interstitial space where extravasation of contrast agent occurs, nonproximal tissues can simultaneously enhance in the case of diffuse or multi-regional lesions. Thus, graphs can be used to model different physiological aspects of DCE-MR.

Although quantitative DCE-MRI methods rely on the ability to measure the extent of contrast agent extravasation from the vascular space to the interstitial space through pharmacokinetic modeling, some researchers have proposed other empirical indirect measures with great success (Hylton, 2006; Jansen et al., 2008). Such an indirect measure is the relative enhancement. Relative enhancement is defined as the difference in lesion signal intensities between each post-contrast scan (after contrast injection) and the pre-contrast or baseline scan, and can be used to model the degree of contrast agent extravasation.

Generally speaking, a graph can be defined as a set of nodes N that represent areas of enhancement and a set of links L that represent relationships between proximal and distant areas of enhancement. For any two nodes n_i, $n_j \in N$, a relationship between them can be defined as $\{n_i, n_j\} \in L$. In this work, we assume that node relationships are non-reflexive and undirected since areas of enhancement are not related to themselves and the relationship between enhancements is independent of direction. This means that in the proposed enhancement network, for any $\{n_i, n_j\} \in L$, $\{n_i, n_j\} \equiv \{n_j, n_i\}$ and for any node $n_i \in N$, $\{n_i, n_i\} \notin L$. Furthermore, the higher the probability of lesion presence, the higher the likelihood that the region of enhancement contains nodes that belong to the network and the higher the similarity of enhancement between any pair of nodes, the higher the likelihood that the nodes are connected in the graph.

To formalize these notions into a network formation model, we use the probability output of a deep Artificial Neural Network (ANN) previously developed in Wu (2016), to classify overlapping tiles at the finding's slice as lesion or non-lesion. Then, we define a link energy function that quantifies enhancement similarity modified by proximity between nodes. Thus, the energy between any pair of nodes can be written as:

$$E\left(\{n_i, n_j\} \in L\right) = exp^{\frac{-D(n_i,n_j)}{G} * \frac{1}{RMSD(n_i,n_j)}}$$

where $D(n_i, n_j)$ is a measure of pixel Euclidean distance between node locations and $RMSD(n_i, n_j)$ or the root mean square distance is a measure of relative enhancement similarity between any two nodes n_i, $n_j \in N$. $RMSD$ is calculated by taking the square root average of the four time-point differences between two node locations. D modifies the energy equation by increasing the energy when nodes are proximal. Similarly, $RMSD$ increases the energy when nodes have similar relative enhancement, since $RMSD$ will be small. We take the inverse so that the energy increases when $RMSD$ is small. G is a growth factor that represents a sensitivity threshold for link formation. As G increases, an unconnected network will slowly grow until it is fully connected (see Fig 1). A growth factor of 10 was used for generating all graphs since it captured empirically a balanced relationship between the two terms of the energy link formation equation.

The result is a graph that favours linkages between similar regions of time-course relative enhancement, which are internally connected to each other based on spatial proximity. From the resulting graph, we can calculate features of graph-based connectivity to gain an increased characterization of nonmass-like clusters of spatially and temporally inter-connected regions. In summary, to obtain a nonmass-like enhancement network from DCE-MRI data we employed the following pipeline: First, each nonmass-like lesion was automatically segmented using a deep Artificial Neural Network (ANN) algorithm. Since the output of the ANN is a probability map, we use it to weight each post-contrast scan to obtain a weighted DCE-MRI by lesion probability (see Fig 1f). Second, nodes are extracted at the boundary of the region of enhancement (see Fig 1c) and at the center line using the topological skeleton of the lesion (see Fig 1d). The topological skeleton is a thin version of the lesion that is equidistant to its boundaries. Fig 1e shows the resulting nodes of the graph. Lastly, we apply the link formation model to all node pairs. Fig 1g–j illustrates the resulting network links by applying the link formation energy with different growth factors. The energy value between any pair of nodes is used to assign weights to each of the edges of the network and obtain a final weighted undirected graph representation of a nonmass-like lesion.

3.2. Characterizing nonmass-like lesions via network connectivity analysis

Representing a nonmass-like lesion as a graph has the advantage of allowing us to investigate several interesting connectivity properties. A graph is a representative way of specifying relationships. If two nodes are linked together then we say that they are neighbors. A path in a graph is defined by a collection of links that connect several nodes. Connectivity analysis is founded in these concepts. Graph connectivity metrics useful in mining features for nonmass-like lesions include measures such as the degree, clustering coefficient and measures of centrality such as closeness and betweenness, as well as distance measures between graph paths. Next, we describe how these metrics can help characterize nonmass-like lesions.

Degree denotes the distribution of the number of incident links per node. Nodes with high degree are located in denser areas of enhancement and reflect the most typical relative enhancement in the lesion. The average degree distribution of node neighborhoods provides information about the shape of the network and an example is shown in Fig 2a. Betweenness is a measure of the importance of a particular node to its cluster. Specifically, betweenness quantifies the prominence of a node based on tightly-knit clusters of nodes linked to this node but not to each other. One way to formalize the role of a prominent node is to observe that the largest connected component would break apart into other distinct components if the given node was removed. Therefore, high betweenness means that a node is important to the structure of the network or is important in connecting two or more enhancing regions into one group. An example of node betweenness is shown in Fig 2b. Closeness of a node is the reciprocal of the sum of the shortest paths from the node to all other nodes. Higher values of closeness indicate higher centrality, and hence more spread networks tend to exhibit lower closeness values. Nodes with high closeness values are shown in Fig 2c as red nodes. The clustering coefficient is a basic measure for local density or connections in a network. An example of clustering coefficient values across nodes is illustrated in Fig 2d, with higher clustering coefficient nodes color-coded in red.

In this study we mined a total of 20 connectivity metrics and since each metric consists of a distribution, we extracted summary statistics from the univariate and cumulative distributions, namely the mean, standard deviation, minimum, maximum as well as 25, 50 and 75 percentiles. Also skewness, kurtosis, and variance of the univariate distribution, for a total of 17 summary statistics. As a consequence the resulting dimensionality of graph features is 17 times the number of connectivity metrics, for a total of 20 x 17 or 340 graph features.

3.3. Reducing the dimensionality of feature space

In a typical CAD processing pipeline other features are extracted directly from imaging data to characterize the kinetic behavior of contrast agent, the morphology and texture of enhancing regions. Therefore, we added 197 additional imaging features (34 dynamic, 19 morphological, 44 texture, 20 dispersion, and 80 single-enhancement features), proposed in a previous study (Gallego-Ortiz and Martel, 2016). As a result, the final feature space consisted of 537 total features (340 graph-based and 197 image-based

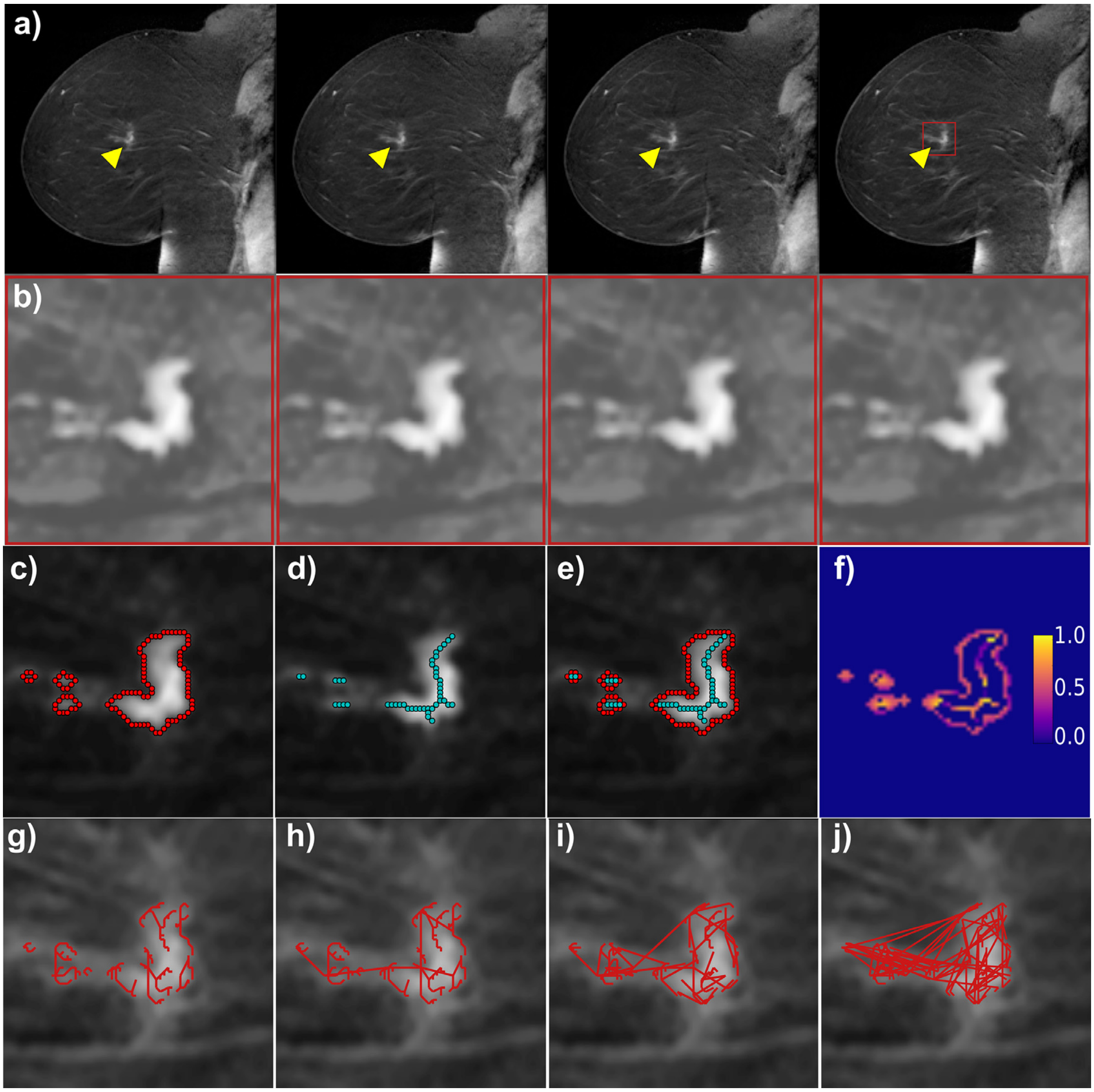

Fig. 1. Clinical breast DEC-MRI saggital fat-suppressed acquisitions at a) 2.5, 4.2, 5.9 and 7.5 min post-contrast injection that have been motioned corrected and subtracted from the baseline pre-contrast scan. Arrow indicates nonmass-like enhancement finding in the clinical radiology report. b) zoom-in into the region of interest (ROI) (red box), c) nodes extracted at the boundary, d) nodes extracted at the center line, e) Resulting graph edges, by combining nodes in c) and d). f) ANN probability of lesion detection at the node candidate locations. g) to j) resulting network links by applying the link formation energy with different growth factors (i.e) $G = [0.01, 1, 10, 100]$ respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

features). Relative to the total number of datasets available, the final feature space is considered a high dimensional feature space.

With ground truth histopathology labels, training a classifier to discriminate between malignant nonmass-like enhancements versus benign or normal enhancement is relatively straightforward, since dimensionality reduction can be accomplished via supervised feature selection. Feature selection mainly removes low discriminative features and selects only the most discriminative features to train the classifier. Without labels or partial labels (some labeled and unlabeled data) dimensionality reduction is more difficult. Unsupervised learning can be used to map high-dimensional data spaces to low-dimensional spaces in an attempt to reduce the number of dimensions while maximizing representability. The very broad field that addresses this problem is known as manifold learning.

Manifold learning can reduce data dimensionality, however, the additional challenge in this study is to learn a suitable lower-dimensional data representation based on the features that are most discriminative of malignant and benign non-mass lesions. More formally, given an original high-dimensional feature set $x_1, \ldots, x_l$ of l points in $\mathbb{R}^{HD}$, the goal is to learn a representation $z_1, \ldots, z_l$ of the original transformed l points in $\mathbb{R}^{LD}$, or embedded space, such that each z_i sufficiently describes or represents the relevant properties found in the original data point x_i informative of the malignancy discrimination task.

In this study, we applied an embedding framework that simultaneously learns feature space representations via unsupervised clustering, and subsequently performs supervised classification in embedded space. We used Deep-Embedded Clustering or (DEC) (Xie et al., 2016) combined with a multi-layered perceptron (MLP) classifier. DEC is optimal for clustering data when ground-truth cluster labels are not necessarily available. One of the contributions of this work is the extension of Xie et al. deep embedded unsupervised clustering (DEC) to an embedding space classification method, with the joint optimization of objective functions for DEC and supervised multi-layered perceptron (MLP). In the next section we explained how we applied DEC to the original high-dimensional data to learn an embedded feature representation that

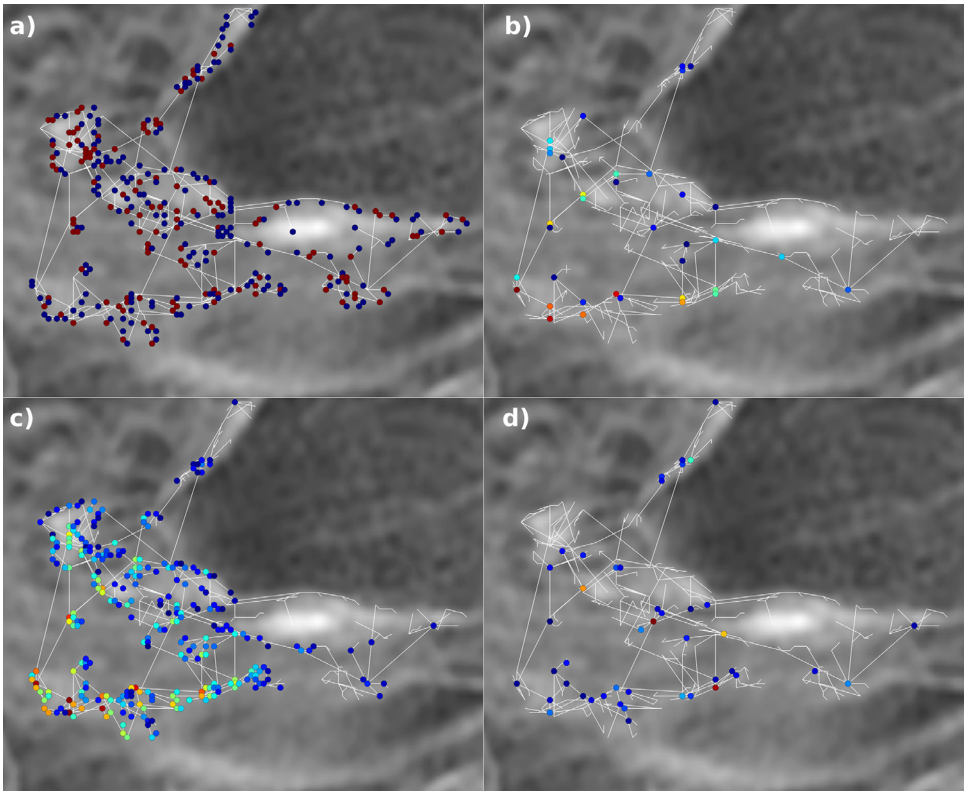

Fig. 2. Graphical representation of node connectivity metrics. a) average degree of node neighborhoods, b) Betweenness, c) closeness centrality, d) clustering coefficient. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

maximizes diagnostic performance of a MLP classifier trained in embedded space to discriminate between malignant and benign non-mass lesions.

3.4. Deep-embedding framework via deep stochastic auto-Encoders (SAE)

The mapping between original high-dimensional and embedded low-dimensional feature space is parameterized using a deep neural network known as Stacked Autoencoder (SAE).

Autoencoding, generally speaking, is an unsupervised data compression and decompression framework that can be learned automatically from data and is an effective way to learn hierarchical representations (Vincent et al., 2008). An autoencoder consists of mainly three components: an encoding function, a decoding function, and a distance function that measures the amount of information loss between the compressed and the decompressed representation of the data. Both encoding and decoding functions are differentiable with respect to the distance function, so the parameters of the encoder and decoder can be optimized to minimize the reconstruction loss using stochastic gradient descent (SGD). In DEC, two stacked autoencoders encode the high-dimensional original data to an embedded feature space of lower dimensionality. We used the same network architecture and training parameters for the stacked autoencoders proposed in the original DEC publication (Xie et al., 2016).

3.4.1. Stochastic auto-Encoder (SAE) architecture

We employed a commonly used architecture for a deep network previously investigated for parametric dimensionality reduction Maaten (2009). The network dimesions consists on $d-500-500-2000-k$, where d represents the input dimensionality, and the subsequent layers correspond to three fully connected layers with 500, 500 and 2000 nodes respectively. Each layer is a denoising autoencoder trained to reconstruct the previous layer's output. The process of training each SAE consist of tunning the parameters of each encoder and decoder such that:

$$\hat{x} \sim Dropout(x)$$

$$h = g_1(W_1\hat{x} + b_1)$$

$$\hat{h} \sim Dropout(h)$$

$$y = g_2(W_2\hat{h} + b_2)$$

Dropout is a stochastic transformation used to randomly set a portion of the input dimensions (20%) to zero as a strategy to avoid overfitting. g_1 and g_2 are the activation functions of the encoding and decoding layers respectively, which consists on rectified linear units (RELUs), and $\theta = \{W_1, b_1, W_2, b_2\}$ are model parameters. Training proceeds in a greedy layer-wise manner by minimizing the least-squares loss $\|x-y\|_2^2$ of each SAE. To form a deep autoencoder, all encoder layers were concatenated followed by all decoder layers in reverse layer-wise training order. Then, the final deep autoencoder was fine-tuned to minimize the reconstruction loss. After fine-tuning the deep autoencoder, the decoding layers were discarded and only the two stacked encoding layers (see Fig 3a) were used as the initial mapping between original data and an embedded space of lower dimensionality also referred here as embedded feature space Z.

3.5. DEC embedded clustering and lesion classification via a Multi-layer Perceptron (MLP) classifier

Following the SAE tuning, DEC performs clustering in embedded space using an algorithm that iteratively refines the deep au-

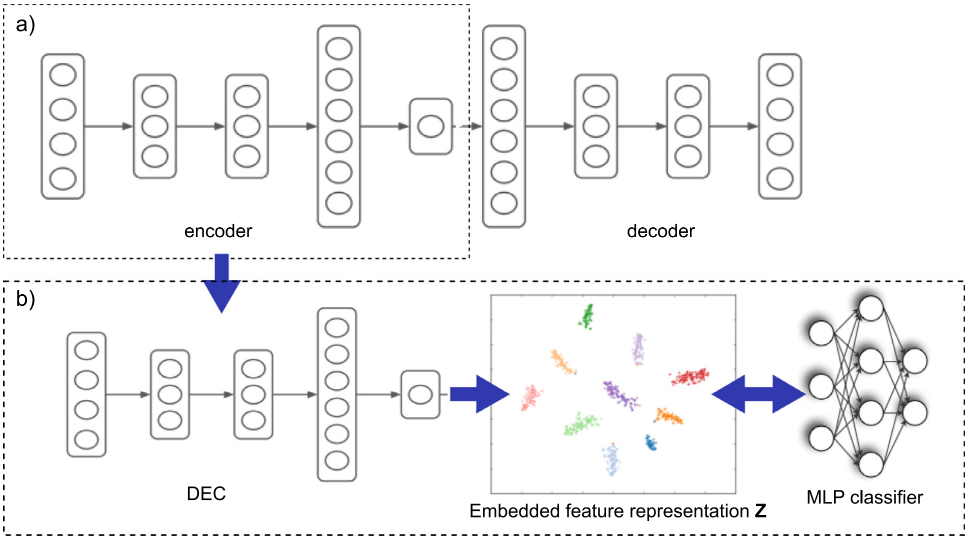

Fig. 3. Schematic illustration of DEC training and optimization in embedded feature space. a) DEC Stacked autoencoder layers representation. b) Network structure of DEC clustering and MLP classification in embedded space.

toencoder parameters and cluster membership assignments. DEC initializes a set of cluster centroids by performing standard k-means on embedded space to obtain k initial centroids μ_j (see Fig 3b). Following k-means initialization, a Students t-distribution is computed from high confidence cluster-assignments, referred as soft-assignments. Soft-assignments can be interpreted as the probability of assigning sample i to cluster j, or q_{ij}. DEC defines its minimization objective without the need for labeled data but with the help of an auxiliary target distribution p_{ij} defined as follows:

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_j'}$$

where $f_j = \sum_i q_{ij}$ are soft cluster assignment frequencies. Intuitively, the purpose of the target distribution is to improve cluster purity by putting more emphasis on data points assigned to each cluster with high confidence. Then, DEC optimization proceeds by minimizing the KullbackLeibler (KL) divergence loss between soft assigned and target distributions as follows:

$$L_{DEC} = \mathbf{KL}(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

In this work, in addition to minimizing the KL divergence loss, we added a fully connected multi-layer perceptron (MLP) classifier to the clustering output in embedded space. To the best of our knowledge the DEC pipeline has not been extended to supervised classification approaches in embedded space. It is important to note that this is the only supervised stage in the framework, but the lesion classifier uses the embedded representation learned with unlabeled data to best discriminate between benign and malignant non-mass lesions. More formally, for training we combined the KL minimization with the MLP negative log-likelihood, or the log-probability of classifying the correct label in the labeled data, defined as:

$$L_{MLP} = \mathbf{NLL}(\theta, \mathbf{D}_{labeled}) = -\sum_i log\Big(P(Y = y^{(i)} | z^{(i)}, \theta)\Big)$$

The combined loss $L_{DEC} + L_{MLP}$ was minimized using mini-batch stochastic gradient descend. Iteratively clusters start to emerge by self-learning from high confidence cluster-assignments using the auxiliary target distribution in DEC (see schematic in Fig 3b). The MLP parameters θ were learned from the embedded feature space Z and the cluster-assignment distribution p_i of each labeled dataset, guided by the clusters that optimize the separation between labeled classes.

3.5.1. Fully-connected MLP architecture

The MLP architecture consisted on the $\mathbb{R}^{LD} - 128 - 32 - 2$, where $\mathbb{R}^{LD}$ represents the lower dimensionality embedded representation produced by DEC, and the subsequent layers correspond to two fully connected hidden layers with 128 and 32 nodes respectively, and an output layer. The two hidden layers have Relu activation functions and the output layer has a softmax activation function that produces a probabilistic output for each class needed to calculate the negative log-likelihood. Our implementation is based on the open-source DEC implementation in MXNet (Xie et al., 2016) and other MXNet python packages (Chen et al., 2015).

3.6. Framework validation

We applied the proposed DEC+MLP pipeline to the entire set of 792 cases, including 415 unlabeled cases, used to train the SAE component and DEC embedding. However, CAD lesion classification performance of the MLP classifier in embedded space could only be assessed using the 377 labeled cases.

To avoid sampling bias, datasets were split into training and testing sets. The training set consisted of a random sample of 80% ($n = 298$) of the labeled cases and the total unlabeled cases ($n = 415$) for a total of 713 datasets. The testing set consisted of the remaining of the labeled cases ($n = 79$) that were held-out of any of the training steps and used to assess the final CAD classifier performance on an independent sample. To assess an unbiased performance assessment we carried out 5-fold cross-validation using the training set, and compute the area under the ROC curve (AUC) per hyperparameter setting across all of the kth held-out validation sets. The same splits of training and held-out sets were used when comparing algorithms. We then retrain the models using the best performing hyper-parameter setting in all training data and assess generalization performance in the testing set. We conducted AUC difference hypothesis testing using the stratified bootstrap resampling method (Robin et al., 2011).

In order to study the contribution to the overall aim of improved CAD performance by each of the methodological aspects of the proposed DEC + MLP classifier, we evaluated the performance at each of the following steps of the proposed pipeline and compared it to a baseline classifier:

Baseline: Original space (no-reduction) with supervised feature selection. To establish a baseline for comparison we used the original space (no-reduction) performance of an MLP classifier trained with an optimal set of selected features. The choice if this baseline is appropriate as it reflects conventional frameworks previously proposed for lesion classification in CADx systems (Nahid and Kong, 2017). For the baseline classifier we used the same MLP architecture and feature selection was implemented using recursive feature elimination (RFE). In RFE, each input feature is ranked using its importance to the model. In this work, features were ranked accordingly to Random Forests importance ranking. If ($S_1 > S_2,$) is the sequence of ordered candidate feature subsets for the number of features to retain, at each iteration of feature selection, the S_i top-ranked predictors are retained and performance is assessed. A repeated 5-fold cross-validation resampling approach was used and the subset of candidate features S_i with the best resampled performance was determined as the optimal feature set. We then trained an MLP with this subset of original input space features.

Step 1: Performance gain with dimensionality reduction, or the SAE component. SAE outputs a lower-dimensional representation of the original feature space. We trained a SAE + MLP classifier in this encoded representation, at varying dimensionality reduction ratios. This resulted in low-dimensional embedded spaces of size 15x, 10x, 5x and 2x times smaller than the size of the original dimensional space.

Step 2: Performance gain with DEC KL divergence based clustering. This step further improves clustering centroids as well as the embedded representation. DEC utilizes the embedded unsupervised SAE representation to learn from high confidence cluster-assignments to modify the embedded feature space. Since we combined KL minimization with the MLP negative log-likelihood, or log-probability of classifying the correct label in a joint optimization, this becomes a semi-supervised component. Diagnostic performance of the proposed DEC + MLP classifier in embedded space was investigated at each of the reduction ratios as in step 1, and in addition we investigated the effect of the number of cluster centroids in a range of cluster centroids from 3 to 12.

Although the number of optimal clusters is typically unknown, it can be used as a trainable parameter. Broadly speaking there are three major lesion classes we are interested in discriminating: malignant, benign lesions and normal parenchymal tissue. Intuitively, as the embedded space representation improves the separation between classes, the diagnostic performance of CAD in embedded space naturally improves compared to the performance of a CAD classifier trained with the original high-dimensional features. Hence, in order to validate our proposed framework, the performance of the proposed DEC+MLP classifier was investigated at varying dimensionality reduction ratios and for a range of cluster centroids (from 3 to 12) and compared to the performances obtained in the original space (baseline) and with only the dimensionality reduction component (step 1).

Table 1
Performance gains at each of the pipeline steps and for a baseline fully-supervised classifier. a) MLP in original space with supervised feature selection, b) Step1: Immediately after the dimensionality reduction step, or the SAE component. c) Step 2: After semi-supervised component or DEC KL divergence based clustering. Results are summarized across different reduction ratios and for best performing number of centroids (in step 2). Results in parenthesis are shown for best performing number of centroids (k).

a) Baseline: Original space (no-reduction) with supervised feature selection
Data subsets

	AUC (*mean* ± *std*)
cv Train	0.71 ± 0.02
cv Validation	0.67 ± 0.08
held-out Test	0.71

b) Across dimensionality reduction ratios

	AUC (*mean* ± *std*)			
	× 15	× 10	× 5	× 2
cv Train	0.98 ± 0.01	0.91 ± 0.03	0.83 ± 0.06	0.77 ± 0.03
cv Validation	0.61 ± 0.07	0.60 ± 0.10	0.67 ± 0.12	0.70 ± 0.05
held-out Test	0.66	0.65	0.70	0.74

c) Results shown for best performing number of centroids (k)

	AUC (*mean* ± *std*)			
	$\times 15(k=6)$	$\times 10(k=5)$	$\times 5(k=8)$	$\times 2(k=3)$
cv Train	0.77 ± 0.02	0.79 ± 0.03	0.79 ± 0.03	0.80 ± 0.03
cv Validation	0.78 ± 0.10	0.79 ± 0.12	0.78 ± 0.11	0.81 ± 0.10
held-out Test	0.76	0.75	0.75	0.78

4. Results

The reconstruction loss between the compressed and the decompressed representation of the data in embedded space as a function of reduction ratios was generally low in both the training data and a held-out testing set. Reconstruction loss also decreased with smaller reduction ratios. Autoencoder reconstruction loss error was the lowest at 2x times reduction, with 0.06% loss in the training set and 1.7% loss in the held-out testing set. On Table 1, we report the performance gains at each of the steps of the proposed pipeline and for the baseline classifier for comparison.

Recursive feature elimination for the baseline classifier produced an optimal subset of 110 features out of the total pool of features ($n=$ 537). The optimal subset was composed of 9 out of 34 dynamic, 8 out of 19 morphological, 30 out of 44 texture, 8 out of 20 dispersion, 21 out of 80 single-enhancement and 34 out of 340 network connectivity features.

The best diagnostic performance in the completely independent held-out set was achieved at a reduction ratio of 2x times and three cluster centroids. Training AUC was 0.80 ± 0.03 (see Table 1c), while validation AUC was 0.81 ± 0.1 (average AUC ± std across all kth fold validation sets for training and validation sets respectively). Note that the generalization performance on the held-out independent test set is comparable (AUC = 0.78).

Fig. 4a shows the best performance achieved after the dimensionality reduction (step 1) and Fig. 4b after embedded clustering (step 2) with the best hyperparameters tuned during training (two-times reduction (2x) and three clusters or $k=3$). There was an incremental increase in AUC by performing DEC clustering following dimensionality reduction via SAE. Cross-validation AUC increased from 0.70 ± 0.05 to 0.81 ± 0.10, and the difference in performance was statistically significant (p-value= $2.3x10^{-7}$). Difference in AUC demonstrated that performance was statistically higher (validation AUC shown in solid green lines in Fig. 4).

When comparing the validation AUC for the overall proposed pipeline with the performance of the MLP baseline classifier, the validation AUC increased from 0.67 ± 0.08 to 0.81 ± 0.10 (p-value= 4.2×10^{-8}) with the proposed graph-based lesion characterization and deep embedding framework. Significance testing of AUC performance of the proposed framework indicated that generalization AUC was not statistically different from training AUC (p-value= 0.315), indicating that our proposed pipeline classifier in embedded space does not suffer from overfitting the training data. It is important to note that in order to carry out feature selection and to train a classifier in the original feature space, it was necessary to remove the 415 unlabeled cases from the training data, since the fully supervised method can only leverage labeled data (actual training data size, n = 377).

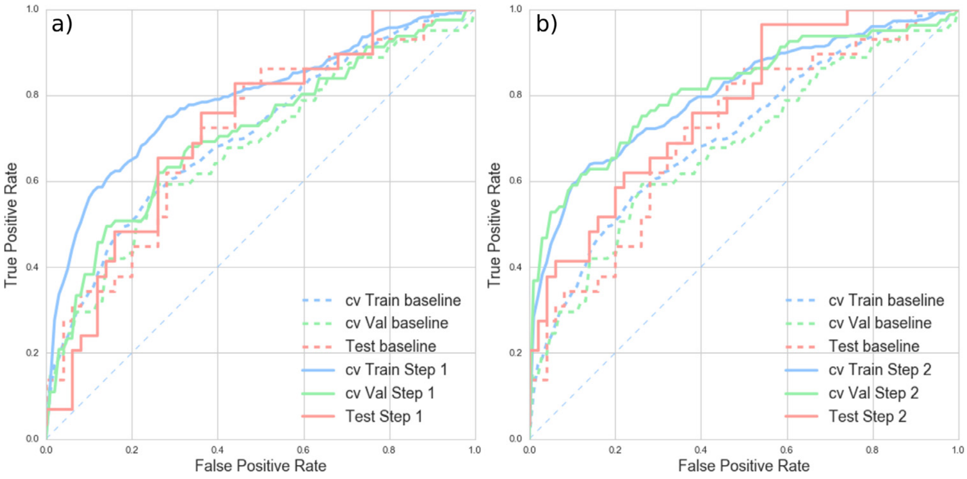

Fig. 4. Comparison of ROC with best diagnostic performance during 5-fold cross-validations for Baseline v.s.: a) after dimensionality reduction component (SAE step1) and b) after semi-supervised clustering component (DEC step2).

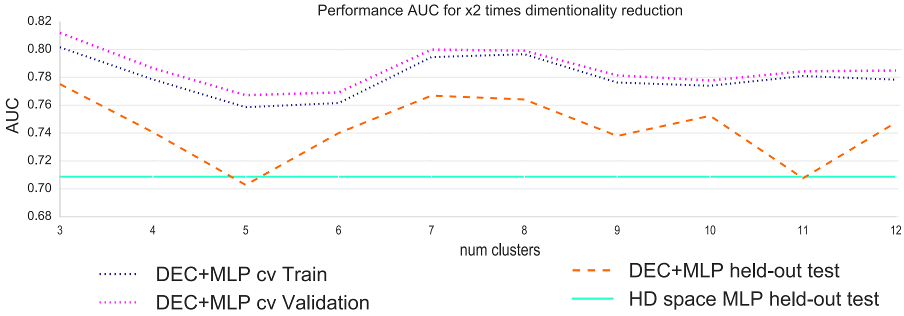

Fig. 5. Performance AUC for the proposed classifier in embedded space for x2 times dimensionality reduction. The DEC+MLP classifier (dotted blue lines) and held-out performance (dashed orange line) outperforms the MLP classifier trained in HD original space (solid cyan line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In Fig. 5 we plot the corresponding performance as a function of the number of centroids or clusters used during DEC-MLP (step 2). The number of centroids had a lesser influence on the diagnostic performance of the proposed classifier. Differences in performance across different number of centroids were not statistically significant. The highest AUC during validation was 0.81 ± 0.10 (average AUC ± std across all folds) and had a generalization AUC in the held-out testing set of 0.78. But, as can be seen in the figure, at x2 times reduction the achieved performance with 7 and 8 cluster centroids was quite similar (average AUC ± std across all folds validation sets of 0.80 ± 0.09 and 0.80 ± 0.10) and a generalization AUC of 0.77 and 0.76 respectively.

A few observations can be made from the results shown in the graphs: First, the proposed method outperforms the fully-supervised MLP classifier in original high-dimensional space (shown by the cyan line plot) for all settings, and shows more robustness to parameter tuning variability. Second, there is a smaller discrepancy between training and validation cv AUC performance for the proposed classifier, which indicates that the classifier trained in embedding space benefits from higher regularization and does not overfit at any setting.

5. Discussion and conclusions

This study presents an improved method for computer-aided diagnosis of nonmass-like enhancements in breast MR imaging. The aims of this work were twofold: First, to extract graph-based features from a network model suitable for nonmass-like lesion representation and second, to reduce efficiently the dimensionality of the resulting feature space while maximizing the diagnostic accuracy of a CAD lesion classifier to discriminate between benign vs. malignant nonmass-like lesions.

We proposed the use of deep embedded clustering (DEC) in combination with a multi-layered perceptron classifier (MLP) in embedded space. The strength of the method lies on the ability to simultaneously learn clusters and embedded feature representations of lower dimensionality. These embedded feature representations can be further optimized by the MLP classifier to better discriminate between lesion classes. DEC produces an embedding based on a centroid-based probability distribution by minimizing the KL divergence to an auxiliary target distribution that improves both cluster assignments and feature representations. In this work, we combined the KL divergence minimization with a negative log likelihood minimization for MLP classification and we were able to improve the discrimination between benign and malignant nonmass-like lesions.

The results of the comparison with more conventional supervised classification techniques previously proposed in CADx, such as a MLP classifier in original feature space with feature selection; showed that unsupervised dimensionality reduction combined with embedded space clustering produced more optimal feature representability for the lesion discrimination task. In another study (Jamieson et al., 2009), t-SNE and Laplacian eigenmaps were used to optimize the feature space for multi-modality breast mass lesion characterization (U.S., DCE-MRI, and full-field digital mammography). The authors also found that lesion classification in unsupervised dimensionality-reduced spaces outperformed traditional supervised feature selection techniques for classification.

The main limitation of this study is the small sample size of labeled data, in particular, the number of samples in the malignant lesion class. This is a well-known challenge in medical image datasets. Nevertheless, DEC handles class imbalances robustly and the proposed methodology exploits the use of unlabeled datasets, which are more abundant and easier to acquire in medical imaging databases. By utilizing information of the emerging clusters in feature space during DEC embedding, the final classifier despite being fully supervised gains class separation due to the inherent structure of feature space and without the need for labels. Another limitation of the method is the necessity to establish the number of clusters apriori. In practice, this quantity is typically unknown. To overcome this challenge, we made the number of clusters a training parameter that was experimentally determined. However, as for DEC a method for determining the optimal number of clusters is needed.

Nonmass-like enhancements in breast MR are a common but diagnostically challenging finding since typical kinetic or morphological descriptors have shown only moderate discriminative power (Baltzer et al., 2010). Improving the discrimination task for nonmass-like lesions in the context of CAD motivated the development of graph-based features presented in this work. To date, there is no consensus in the radiology literature with regards to any particular nonmass-like enhancement pattern with less than a 2% (Giess et al., 2013) chance of malignancy. However, a recent study (Machida et al., 2015) assessed differences in distribution types of 156 nonmass enhancements and found that a subgroup of linearly distributed lesions smaller than 1 cm had 0% PPV for malignancy (i.e none of the lesions in these subgroup were malignant). We believe that CAD may be useful in exploratory research like this to discover features associated with a significant reduction in the malignant likelihood of nonmass-like enhancement. The proposed CAD pipeline could be applied to future work aimed at exploring associations in a larger group of lesions or in specific subsets of patients.

Acknowledgments

This research is funded by the Canadian Breast Cancer Foundation (CBCF) - Ontario Region, and the Ontario Institute for Cancer Research (OICR) through funding provided by the Government of Ontario. The authors thank Sharmila Balasingham, BSc, for help curating patient clinical information from multiple databases and Nim Li, BComm, for technical assistance with the research database.

References

Bahl, M., Barzilay, R., Yedidia, A.B., Locascio, N.J., Yu, L., Lehman, C.D., 2018. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. Radiology 286 (3), 810–818. doi:10.1148/radiol.2017170549.

Baltzer, P.A.T., Benndorf, M., Dietzel, M., Gajda, M., Runnebaum, I.B., Kaiser, W.A., 2010. False-Positive findings at contrast-Enhanced breast MRI: a BI-RADS descriptor study. Am. J. Roentgenol. 194 (6), 1658–1663. doi:10.2214/AJR.09.3486.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z., 2015. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems., 1–6 arXiv:1512.01274. doi:10.1145/2532637.

Gallego-Ortiz, C., Martel, A.L., 2016. Improving the accuracy of computer-aided diagnosis for breast MR imaging by differentiating between mass and nonmass lesions. Radiology 278 (3), 679–688. doi:10.1148/radiol.2015150241.

Giess, C.S., Raza, S., Birdwell, R.L., 2013. Patterns of nonmasslike enhancement at screening breast MR imaging of high-Risk premenopausal women. Radiographics 33 (5), 1343–1360. doi:10.1148/rg.335125185.

Gutierrez, R.L., DeMartini, W.B., Eby, P.R., Kurland, B.F., Peacock, S., Lehman, C.D., 2009. BI-RADS Lesion characteristics predict likelihood of malignancy in breast MRI for masses but not for nonmasslike enhancement. Am. J. Roentgenol. 193 (4), 994–1000. doi:10.2214/AJR.08.1983.

Hylton, N., 2006. Dynamic contrast-Enhanced magnetic resonance imaging as an imaging biomarker. J. Clin. Oncol. 24 (20), 3293–3298. doi:10.1200/JCO.2006.06.8080.

Jamieson, A.R., Giger, M.L., Drukker, K., Li, H., Yuan, Y., Bhooshan, N., 2009. Exploring nonlinear feature space dimension reduction and data representation in breast CADx with laplacian eigenmaps and t-SNE. Med. Phys. 37 (1), 339–351. doi:10.1118/1.3267037.

Jansen, S.A., Fan, X., Karczmar, G.S., Abe, H., Schmidt, R.A., Newstead, G.M., 2008. Differentiation between benign and malignant breast lesions detected by bilateral dynamic contrast-enhanced MRI: a sensitivity and specificity study. Magn. Reson. Med. 59 (4), 747–754. doi:10.1002/mrm.21530.

Khan, F.M., Scott, R., Donovan, M., Fernandez, G., 2017. Predicting and replacing the pathological gleason grade with automated gland ring morphometric features from immunofluorescent prostate cancer images. J. Med. Imaging 4 (2), 021103. doi:10.1117/1.JMI.4.2.021103.

Lourenco, A.P., Khalil, H., Sanford, M., Donegan, L., 2014. High-risk lesions at MRI-guided breast biopsy: frequency and rate of underestimation. Am. J. Roentgenol. 203 (3), 682–686. doi:10.2214/AJR.13.11905.

Maaten, L.V.D., 2009. Learning a parametric embedding by preserving local structure. JMLR Proc. 5, 384–391. (AISTATS).

Machida, Y., Tozaki, M., Shimauchi, A., Yoshida, T., 2015. Two distinct types of linear distribution in nonmass enhancement at breast MR imaging: difference in positive predictive value between linear and branching patterns. Radiology 276 (3), 686–694. doi:10.1148/radiol.2015141775.

Martel, a.L., Froh, M.S., Brock, K.K., Plewes, D.B., Barber, D.C., 2007. Evaluating an optical-flow-based registration algorithm for contrast-enhanced magnetic resonance imaging of the breast.. Phys. Med. Biol. 52 (13), 3803–3816. doi:10.1088/0031-9155/52/13/010.

Morris, E.A., Comstock, C., Lee, C., 2013. ACR BI-RADS® Magnetic resonance imaging. Am. Coll. Radiol..

Nahid, A.-A., Kong, Y., 2017. Involvement of machine learning for breast cancer image classification: a Survey. Comput. Math. Methods Med. 2017 (i), 1–29. doi:10.1155/2017/3781951.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. Proc: an open-source package for r and s+ to analyze and compare ROC curves. BMC Bioinform. 12 (1), 77. doi:10.1186/1471-2105-12-77.

Shao, Z., Wang, H., Li, X., Liu, P., Zhang, S., Cao, S., 2013. Morphological distribution and internal enhancement architecture of contrast-Enhanced magnetic resonance imaging in the diagnosis of non-Mass-Like breast lesions: A Meta-Analysis. Breast J. 19 (3), 259–268. doi:10.1111/tbj.12101.

Sudbø, J., Marcelpoil, R., Reith, A., 2000. New algorithms based on the voronoi diagram applied in a pilot study on normal mucosa and carcinomas. Anal. Cell. Pathol. 21 (2), 71–86.

Tagliafico, A., Bignotti, B., Tagliafico, G., Tosto, S., Signori, A., Calabrese, M., 2015. Quantitative evaluation of background parenchymal enhancement (BPE) on breast MRI. a feasibility study with a semi-automatic and automatic software compared to observer-based scores. Br. J. Radiol. 88 (1056). doi:10.1259/bjr.20150417.

Thomassin-Naggara, I., Salem, C., Darai, E., Bazot, M., Uzan, S., Marsault, C., Chopier, J., 2009. Non-masslike enhancement in breast MRI: the pearls of interpretation? J. Radiol. 90, 269–275. JR-04-2009-90-4-0221-0363-101019-200904045.

Tozaki, M., Fukuda, K., 2006. High-Spatial-Resolution MRI of non-Masslike breast lesions: interpretation model based on BI-RADS MRI descriptors. Am. J. Roentgenol. 187 (2), 330–337. doi:10.2214/AJR.05.0998.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning - ICML '08. ACM Press, New York, New York, USA, pp. 1096–1103. doi:10.1145/1390156.1390294.

Wu, H., 2016. Automatic Computer Aided Diagnosis of Breast Cancer in Dynamic Contrast Enhanced Magnetic Resonance Images. University of Toronto Master of science.

Xie, J., Girshick, R., Farhadi, A., 2016. Unsupervised Deep Embedding for Clustering Analysis. In: The Journal of Machine Learning Research, pp. 478–487. 1511.06335.