Automated Detection of DCIS in Whole-Slide H&E Stained Breast Histopathology Images

Babak Ehteshami Bejnordi*, Maschenka Balkenhol, Geert Litjens, Roland Holland, Peter Bult, Nico Karssemeijer, and Jeroen A. W. M. van der Laak

***Abstract*—This paper presents and evaluates a fully automatic method for detection of ductal carcinoma in situ (DCIS) in digitized hematoxylin and eosin (H&E) stained histopathological slides of breast tissue. The proposed method applies multi-scale superpixel classification to detect epithelial regions in whole-slide images (WSIs). Subsequently, spatial clustering is utilized to delineate regions representing meaningful structures within the tissue such as ducts and lobules. A region-based classifier employing a large set of features including statistical and structural texture features and architectural features is then trained to discriminate between DCIS and benign/normal structures. The system is evaluated on two datasets containing a total of 205 WSIs of breast tissue. Evaluation was conducted both on the slide and the lesion level using FROC analysis. The results show that to detect at least one true positive in every DCIS containing slide, the system finds 2.6 false positives per WSI. The results of the per-lesion evaluation show that it is possible to detect 80% and 83% of the DCIS lesions in an abnormal slide, at an average of 2.0 and 3.0 false positives per WSI, respectively. Collectively, the result of the experiments demonstrate the efficacy and accuracy of the proposed method as well as its potential for application in routine pathological diagnostics. To the best of our knowledge, this is the first DCIS detection algorithm working fully automatically on WSIs.**

***Index Terms*—Computer-aided diagnosis, DCIS Detection, H&E staining, whole-slide imaging.**

I. Introduction

BREAST cancer is the second leading cause of cancer death among women [1]. Approximately 80% of breast cancers arise from epithelial cells lining the ducts (ductal carcinoma). Pathological diagnosis for intraductal proliferative lesions comprise a spectrum with increasing malignant potential, ranging from usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), to invasive ductal carcinoma (IDC) [2]. In this spectrum, DCIS (with cancer cells still being contained within the glandular tissue) and IDC (cancer cells invading the surrounding tissue) are considered malignant, prompting for immediate treatment [1].

DCIS encompasses a heterogeneous group of lesions with highly variable morphology, biomarker expression, genomic profile, and natural progression [3]. Whereas the extremes of the spectrum are easily discernible, the difference between UDH, ADH, and low-grade DCIS is subtle and the classification of such lesions suffers from significant inter-observer variability even among expert pathologists. Introduction of computer aided diagnosis (CAD) systems for breast pathology will be successful if these difficult cases can be handled, with sufficient accuracy. CAD can assist the pathologist in two ways: (1) by detecting all the clinically relevant regions of interest (ROIs) per slide, allowing the pathologist to only focus on the interpretation of these regions, or (2) by providing an accurate assessment of suspicious regions and reducing the variability in pathologists' interpretations. Several recent studies have focused on automated discrimination of DCIS from benign intraductal breast lesions [4]–[6]. Two approaches, based on identification and segmentation of nuclei and the quantification of nuclear features by Dong *et al.* [4] and Dundar *et al.* [5] could discriminate DCIS from UDH with area under the receiver operating characteristic curve (AUC) of 0.86 and 0.93, respectively. Srinivas *et al.* [6] proposed a simultaneous sparsity model to automatically evaluate intraductal breast lesions for cancer diagnosis.

One of the major drawbacks of many published studies on CAD in pathology is the fact that only manually selected ROIs (mostly selected by expert pathologists) were used. A fully automated algorithm that can be used in large-scale histopathological image analysis should automatically identify ROIs in the whole-slide-image (WSI) and discriminate DCIS from different types of benign lesions. This task is particularly challenging for two main reasons: (1) WSIs are large and may contain hundreds of structures which need to be analyzed. Therefore, obtaining a small false positive rate while still retaining a high sensitivity can be hard, and (2) A CAD system that operates on the WSI level should be able to handle a larger set of heterogeneous benign structures (e.g., adenosis, UDH, cysts, etc.) and artifacts (due to staining/cutting) to detect DCIS.

In this paper, we present a fully automated CAD system that can discriminate DCIS from normal/benign conditions in WSI. Our proposed system initially detects epithelial regions in the WSI.A common approach to localize important structures in

Manuscript received February 28, 2016; revised March 25, 2016; accepted March 25, 2016. Date of publication April 05, 2016; date of current version August 30, 2016. This work was supported by the European Union FP7 funded VPH-PRISM project under grant agreement 601040. This work was also supported by the Stichting IT Projecten (STITPRO) and the Radboud Institute for Health Sciences (RIHS), both in Nijmegen, The Netherlands . *Asterisk indicates corresponding author.*

*B. Ehteshami Bejnordi is with the Diagnostic Image Analysis Group, Radboud University Medical Center, 6500HB Nijmegen, The Netherlands (e-mail: babak.ehteshamibejnordi@radboudumc.nl).

M. Balkenhol, P. Bult, and J. A. W. M van der Laak are with the Department of Pathology, Radboud University Medical Center, 6500HB Nijmegen, The Netherlands.

G. Litjens is with the Hamamatsu Tissue Imaging and Analysis Center, University of Heidelberg, D-69120 Heidelberg, Germany.

R. Holland and N. Karssemeijer are with the Diagnostic Image Analysis Group, Radboud University Medical Center, 6500HB Nijmegen, The Netherlands.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2016.2550620

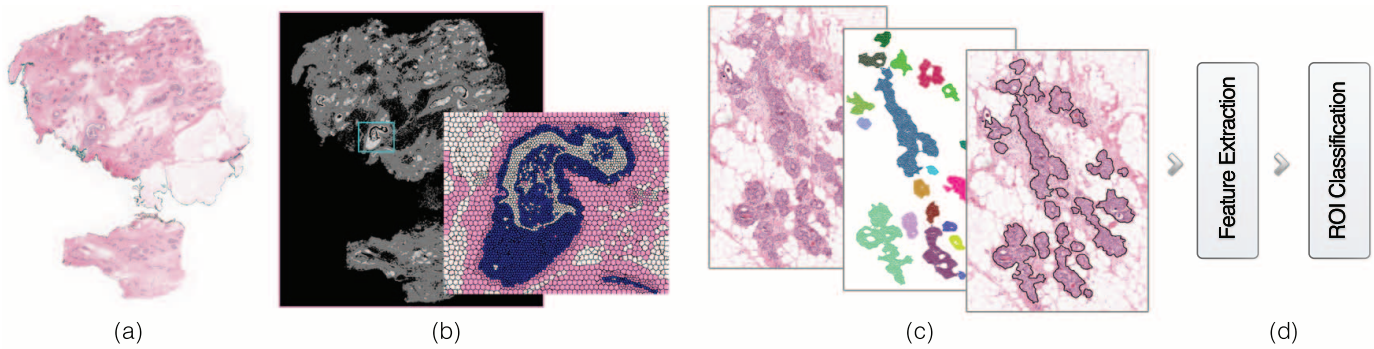

Fig. 1. Overview of the proposed DCIS detection system. (a) Original WSI of a breast tissue slide. (b) Application of multi-scale superpixel classification to classify the image into epithelium, stroma, and background. (c) Graph-based clustering of the epithelium labeled superpixels for delineation of ROIs. (d) Feature extraction and classification of each of the candidate ROIs.

WSIs is to divide the image into rectangular patches and classify them (possibly at multiple resolutions) [7]–[9]. However, these rectangular patches may contain mixtures of class types which will lower the accuracy of the classification. To tackle this problem, our system uses a multi-scale superpixel classification approach [10] to detect epithelial regions in the WSI. Superpixels are classified at multiple resolutions to efficiently detect regions containing epithelium. The superpixels labeled as epithelium are then grouped into histopathologically meaningful regions by application of a graph clustering algorithm. A set of texture and spatial distribution features is then extracted from each candidate region, after which a classifier classifies the region as either DCIS or benign/normal.

Empirical evaluation of the performance of the proposed system is presented in two experiments using two separate datasets. The first dataset comprises 150 WSIs of breast tissue sampled from 150 patients (75 benign/normal and 75 containing DCIS). The second comprises 55 WSIs of breast tissue sampled from 43 patients which are representative of the daily clinical routine samples examined by a pathologist during a specific period of time. The first experiment evaluates the efficacy of the proposed system in detecting and localizing DCIS regions in WSIs using the first dataset. The Dice coefficient is used to evaluate the accuracy of the DCIS segmentation. The second evaluates the performance of the system in classifying a WSI as DCIS at the slide level using the second dataset. This is an important aspect in evaluating the merit of the proposed CAD system because it highlights its potential for application in routine pathological diagnostics.

II. Detailed Description of the Proposed DCIS Detection System

The proposed DCIS detection system takes as input an H&E stained WSI and yields as output the segmentation of the potential DCIS lesions together with a likelihood estimation for each lesion to be DCIS. Fig. 1 presents an overview of the proposed DCIS detection system. The proposed algorithm has 3 basic steps:

a) Multi-scale superpixel classification for finding epithelial areas in the WSI.
b) Graph-based clustering of the superpixels labeled as epithelium and delineation of ROIs.
c) Classification of the segmented regions as benign/normal or DCIS.

Detailed description of the proposed algorithm's steps are discussed below.

A. Multi-Scale Superpixel Classification for Finding Epithelial Areas

Detection of epithelial regions in the WSI is based on the multi-scale superpixel classification algorithm [10]. This algorithm enables subdivision of the WSI into regions which adapt to the underlying image data, such that every superpixel is mostly homogeneous. Accurate classification of the tissue components within the WSI is thereby facilitated. The algorithm initially partitions the image at 1.25X magnification (with pixel size of 3.88 μm$\times$ 3.88 μm) into a set of non-overlapping superpixels using the simple linear iterative clustering (SLIC) algorithm [11]. The generated superpixels each contain approximately 5000 pixels. Image regions containing mainly epithelium or stroma are identified by excluding all the superpixels whose content is more than 90% background. A pixel within a superpixel is classified as background if its overall optical density (i.e., $-\log(I/I_0)$ where I_0 denotes the intensity of the light source) is lower than 0.2 and the density of its r, g, and b channels is lower than 0.25.

Fig. 2 presents the next steps in the multi-scale superpixel classification algorithm. New superpixels are constructed at 5X magnification (pixel size of 0.97 μm$\times$ 0.97 μm) within the areas classified as epithelium or stroma in the previous step (see Fig. 2(b)). These are then again classified into three distinct components: stroma, epithelium, and background (non-tissue containing regions as well as regions containing fat cells and fluid). The size of each superpixel at this magnification was set to be approximately 2000 pixels. Identification of background superpixels is performed similarly to the previous step. For the classification of the remaining superpixels into stroma or epithelium, a set of 54 features were extracted for each superpixel s, including 8 pixel value statistics (minimum, maximum, sum, mean, standard deviation, lower quartile, median, and upper quartile) and 10 uniform local binary pattern features for radius 1 derived from each of the channels of the hue-saturation-density (HSD) color model [12]. In addition, the mean and standard deviation of all of these features for the set of all neighboring superpixels to the superpixel s was included,

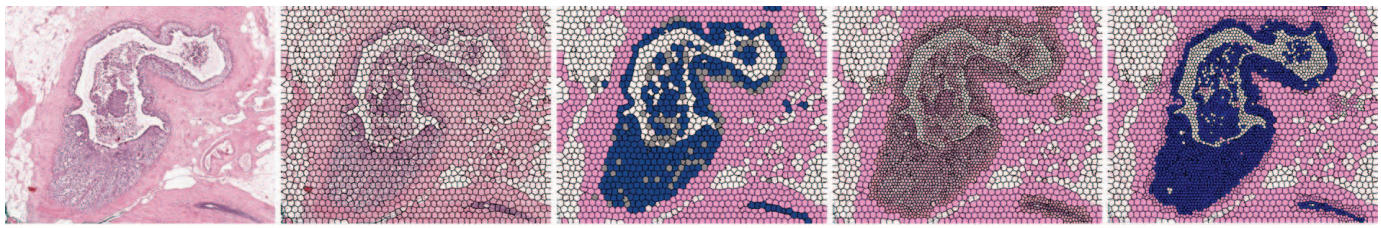

Fig. 2. Illustration of multi-scale superpixel classification. (a) Original Image. (b) Generation of superpixels in the intermediate magnification (5X). (c) Classification of the superpixels into background (white), stroma (pink), or epithelium (blue). Note that gray-colored superpixels are the ones for which the probability score of the classifier for all of the classes was below 0.7. (d) Generating superpixels on the areas requiring more detailed information in higher magnification (20X). Note that the smaller superpixels are built on the highest magnification image while the larger ones are the same superpixels computed in the intermediate magnification. (e) Final classification result.

yielding a total of 162 features. A random forest classifier using 100 decision trees trained on approximately 20,000 manually annotated superpixels (generated on sample patches taken from 30 WSIs in the training set) was employed for classifying the superpixels. Fig. 2(c) shows the results of the classification at the intermediate magnification.

Finally, to achieve a more accurate delineation of histopathological structures, a new set of superpixels was built and classified at the highest magnification (20X) within the areas requiring more detailed information (see Fig. 2(d)). A region required more accurate classification (using the highest magnification image) if it satisfied either of these two conditions: (1) The classifier used to classify the region at the intermediate magnification yielded a low confidence in assigning the output label. The level of uncertainty was assigned according to the output probability for the superpixel classification. Superpixels having a likelihood probability lower than 0.7 for all of the classes were considered uncertain. (2) The region was labeled as epithelium by the classifier in the intermediate magnification. The first condition ensures that a more accurate classification is achieved by using more detailed information present in the higher magnification. The second is to obtain more detailed contouring of the areas that were labeled as epithelium.

The newly generated superpixels in the corresponding areas satisfying the two conditions mentioned above had an approximate size of 1000 pixels. The set of 54 features described previously were extracted for each superpixel in this magnification. Moreover, to incorporate more contextual information for the superpixel s, the set of 162 features previously computed for the parent superpixel s' in the intermediate magnification was appended to the feature list, where s' is the superpixel which has the largest overlap with the corresponding area occupied by the superpixel s. A second stage random forest classifier with 100 decision trees was subsequently utilized to classify these superpixels more accurately. Fig. 2(e) shows the final classification result by the multi-scale superpixel classification approach.

B. Graph-Based Clustering of Superpixels for Delineation of ROIs

The output of the multi-scale superpixel classification algorithm is a set of superpixels with three possible labels (stroma, background, and epithelium). To create regions representing anatomically meaningful structures within the tissue such as ducts or lobules, the superpixels have to be clustered. The aim of this step is not only to merge the superpixels neighboring each other but also splitting distinctive structures lying in the vicinity of each other. To perform the clustering, we propose an algorithm based on local graph structure that models the spatial distribution of the labeled superpixels in the image. Our proposed spatial clustering algorithm prunes the edges of a region adjacency graph built on the centroids of the epithelium-labeled superpixels to cluster them into meaningful tissue regions in the WSI, while still maintaining the overall connectivity of each cluster. The entire algorithm for delineating ROIs can be summarized in three major steps:

1) Step 1: Using a relative neighborhood graph to identify coarse clusters of neighboring superpixels.
2) Step 2: Applying spatial clustering to find spatially homogenous sub-clusters within clusters from the first step.
3) Step 3: Finding the concave-hull of each sub-cluster as the outer boundary of the identified ROI.

Detailed description of each of the steps is discussed below.

1) Step 1: Identifying Coarse Clusters: The algorithm for identifying several isolated groups of coarse clusters can be described as follows:

a) Apply the multi-scale superpixel classification algorithm to the input WSI to obtain labeled superpixels as described in Section II-A.
b) Construct the relative neighborhood graph $RNG(V)$ [13] of the pointset V containing the centroids of the n superpixels with epithelium label in Euclidean space. In the relative neighborhood graph two points v_i and v_j are neighbors if $d(v_i, v_j) \leq \max[d(v_i, v_k), d(v_j, v_k)]$, $\forall k = 1, \ldots, n$ and $k \neq i, j$.
c) Apply a threshold ($T = 200$, equivalent to the diameter of two superpixels) on the maximum edge length of the graph to partition $RNG(V)$ into several local sub-graphs (G^k) and label the entire group of subgraphs using the depth-first search (DFS) algorithm [14].

The threshold on maximum edge length was determined based on the assumption that two superpixels lying further than the diameter of two superpixels away should not be considered neighbors.

2) Step 2: Spatial Clustering of Coarse Clusters Into Anatomically Meaningful Sub-Clusters: The identified coarse clusters in the previous step may contain multiple anatomically meaningful structures which are lying in the vicinity of each other. In this step we cluster each of the identified $G^k = (N, E)$

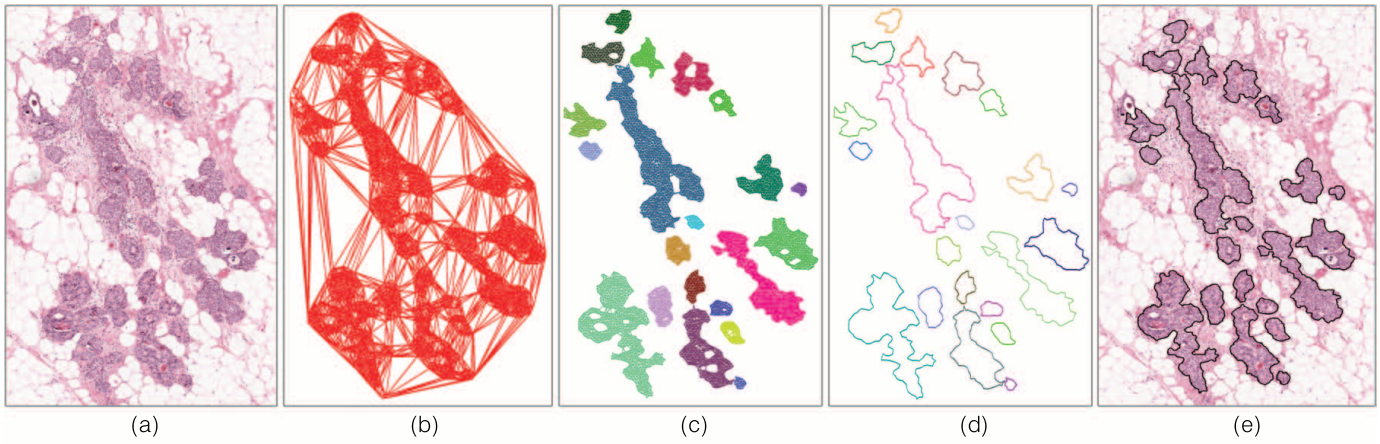

Fig. 3. Graph-based clustering of superpixels for delineation of ROIs. (a) Original Image. (b) Delaunay triangulation built on the set of epithelium labeled superpixels. (c) Application of the graph-clustering algorithm to cluster the graph into several meaningful sub-graphs. (d) Calculating the concave hull for each of the sub-graphs. (e) Final contouring of each ROI.

into several anatomically meaningful tissue regions such as ducts or lobules, where N and E are the set of vertexes and edges of G^k respectively. Fig. 3 shows the processes involved in the proposed spatial clustering algorithm. Our spatial clustering algorithm utilizes Delaunay triangulation (DT), which is a suitable tool for spatial clustering as it implicitly encapsulates vast amount of proximity information (see Fig. 3(b)). The proposed clustering algorithm eliminates the extra simplexes of the triangulation according to a local heterogeneity measure. A DT built on a cloud of points may have both inter- or intra-cluster simplexes. Inter-cluster simplexes are the ones connecting two or multiple anatomically meaningful regions (e.g., ducts or lobules) to each other. Intra-cluster simplexes, however, are the ones connecting multiple vertexes inside a single sub-cluster. Our objective is to extract measures from each simplex in a constructed DT to discriminate between inter- and intra-cluster simplexes and consequently identify separate clusters of points each belonging to separate regions. For this purpose, we define three measures to describe spatial heterogeneity of the simplexes.

The perimeter of the simplex is taken as the first measure describing local topography of DT. According to the density-based definition of clusters, intra-cluster edges are much shorter compared to inter-cluster edges [15], [16]. Consequently, it can be inferred that inter-class simplexes have higher perimeter values than the intra-cluster ones.

The second measure quantifies the elongation of the simplex. Inter-cluster simplexes of a DT tend to have more elongated shapes. To measure the elongation $EL(s)$ of the triangle s, we compute the ratio between the major and minor axes of s's Steiner circumellipse [17]. Steiner's Circumellipse is a unique ellipse whose center coincides with the centroid of the triangle and passes through the vertexes of the triangle. For equilateral triangles, the measure $EL(s) = 1$, and for all other conditions $EL(s) > 1$.

Our final measure quantifies the local shape heterogeneity around the simplex. In this way we can evaluate the tendency of the current simplex to be in the same cluster as its neighboring ones. For this purpose, we compute the standard deviation of the elongation measures over the set $s \cup N(s)$, where $N(s)$ denotes the set of neighboring simplexes of simplex s. Finally, the entire simplex analysis is captured in a criterion function $F(s)$, which is defined as: $F(s) = Perimeter(s) \times EL(s) \times Std(1+EL(s \cup N(s))$. This function takes into account spatial heterogeneity of the simplexes and primarily penalizes large simplex perimeters. The two elongation terms are used as weighting factors that further penalize simplexes that have large elongation and/or neighboring simplexes with heterogeneous elongations. For each simplex s in DT, if $F(s)$ is bigger than a predetermined threshold the simplex is removed from the graph. We found the threshold value of 250 suitable. After eliminating inter-class simplexes and noises, only positive nodes and edges of the graph remain. Through depth-first search we then infer the number of isolated clusters and correspondingly the list of points in each cluster.

As a result of pruning the inter-cluster simplexes we may lose the points lying on the hull of each cluster. To reassign these points to the appropriate cluster we start an iterative graph growing process. Let $S^j \subset G^k$ be a clustered graph within the local sub-graph $G^k = (N, E)$, and let $V(i), \{i \in N\}$ denote the set of points neighboring the vertices at the hull of S^j. At each iteration, a point i in $V(i)$ is assigned to S^j under two conditions; (1) if the Euclidean distance between the node i and it's neighboring node in S^j is less than the maximum edge length in S^j. (2) if i is not neighboring the hull of another clustered graph $S^{j'} \subset G^k$. The first rule reduces the possibility of assigning a noisy node to a cluster, and the second will prevent the merging of two isolated clusters. The assignment of new nodes to the graph is repeated for 3 iterations. Because of the two constraints applied, mostly there are no new nodes added to the sub-graphs after 2 or 3 iterations.

3) Step 3: Finding the Outer Contour of Each Sub-Cluster: The final step is to extract the boundary of the clustered graphs. Let $G^k = (N, E)$ denote a sub-graph created in step 1 of our algorithm and let $S^j = (N(j), E(j))$ denote a clustered graph obtained in step 2 satisfying $S^j \subset G^k$. To find the actual boundary of the S^j graph which corresponds to the concave hull created by the edges $E(j)$ on the point set $N(j)$ we first compute DT of the point set $N(j)$. The boundary of the exterior face of the DT is the convex hull of the point set. Let Γ_{DT} denote the set of edges of the convex hull. By traversing along the edges of

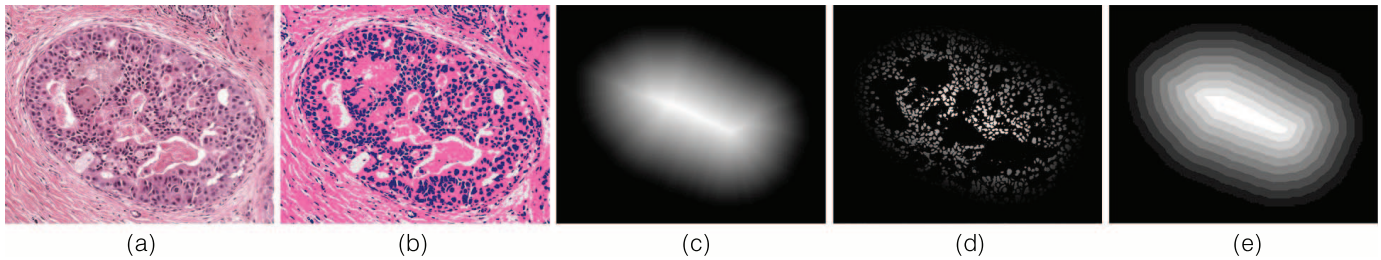

Fig. 4. Illustration of the selected steps required for computation of some of the architectural features. (a) Original image of a DCIS region. (b) The result of pixel classification using the algorithm proposed in [18]. (c) Euclidean distance transform of the inverse DCIS mask. The mask is computed using the output of the ROI delineation algorithm in step II-B3. (d) Portions of *(c)* cut-out by the mask of the hematoxylin stained pixels. These cut-out distances are subsequently used for computation of margination features. (e) Division of the DCIS mask into 10 zones. The ratio between the area of the hematoxylin stained pixels in each zone to the area of that zone are used as measures to characterize the distribution of the nuclei within the candidate ROI.

Γ_{DT} and removing and replacing the edges not present in $E(j)$ with the other two edges of the simplexes in DT to which the removed edges belong to, we can find the concave hull of the S^j graph. Traversing along the edges in Γ_{DT} is continued until the condition $\Gamma_{DT} \subset E(j)$ is met. The edges remained in Γ_{DT} correspond to the outer boundary of the S^j graph. At the end of this step, a more accurate delineation of the ROI is obtained by taking the union of the binary masks of the superpixels lying on the concave hull and within the binary mask of the concave hull itself.

C. Region-Based Feature Extraction and Classification

Cellular and architectural features are the major characteristics considered by a pathologist for diagnosing DCIS. Therefore, the features used in this study to distinguish DCIS from different benign/normal regions are a combination of statistical and structural texture features, and features describing the spatial distribution of the components inside the ROI. A classifier is employed to classify each of the segmented ROIs using the extracted features.

1) Texture Features: To extract texture features, each candidate region identified through our spatial-clustering method is first divided into several superpixels having an approximately equal size of 5000 pixels using the SLIC algorithm (the analyzed image has pixel size of 0.486 μm$\times$ 0.486 μm). For each of the superpixels 5 different types of texture features are extracted from the gray-scale intensities of the image. These features are statistics of the gray level histogram (mean, standard deviation, median, first and third quartiles, interquartile range), 14 statistics calculated from the co-occurrence matrix [19], uniform local binary patterns for radii 1 and 2 [20], and gray level histogram statistics extracted from responses to filter banks in particular Laplacian of Gaussian (LoG) at 5 scales, and Gabor filters at 4 scales and 8 orientations are extracted. These texture features have shown strong discriminatory power in characterizing histopathology images [7], [21], [22]. In total 256 features were extracted for each superpixel. The mean, standard deviation, 5th and 95th percentile of each feature over all superpixels in an ROI yielded a total of 1024 region-based features. Computation of the texture features at the superpixel level rather than pixel level was done to reduce the computation cost of the statistics which are finally derived from them. Moreover, using super-pixels it is possible to extract regions that contain homogeneous tissue structures, therefore the extracted features from these regions tend to be more meaningful and discriminative.

2) Architectural Features: An initial step before computation of the architectural features is classifying the region into different tissue components. For this purpose, we use our recently proposed algorithm [18] for robust stain classification which makes use of spatial information. This algorithm operates at the WSI and automatically extracts training samples for each stain class (the class absorbing mostly hematoxylin and the class absorbing mostly eosin) from the image, obviating the need for manually labeled training data. This algorithm is an intermediate step in the published stain standardization algorithm [18]. Fig. 4(b) shows an example classification result for a detected ROI. The classified image is median filtered (kernel size 5×5) for removing noisy labels from the result.

Following extraction of the masks for different stain classes, we compute the area of the hematoxylin stained, eosin stained and background pixels. These three area measures implicitly include information about the size of the ROI. For this reason, three additional features were included that were normalized to the total area of the ROI.

A subset of features were designed to measure the margination of the nuclei. The margination features characterize the distances of the nuclei to the ROI boundary. The steps required for computation of the margination features are illustrated in Fig. 4. First, the Euclidean distance transform of the inverse mask of the ROI is computed as shown in Fig. 4(c). Next, portions of this image are cut out by the mask of the hematoxylin stained pixels (see Fig. 4(d)). The distribution of the cut-out distances were quantified at five percentiles (10th to 90th in ten percentile steps). Five additional features were included by normalizing the percentile values to the maximum value of the distance transform of the ROI.

A subset of architectural features have been computed to describe the distribution of the nuclei within the ROI. To compute these features, the area inside the ROI is first divided into ten different zones $Z_k = \{i \in ROI \mid D_{\max} * (k-1)/10 < D(i) < D_{\max} * k/10\}$ where i is an arbitrary pixel inside the ROI, $k \in \{1, 2, 3, \ldots, 10\}$, $D(i)$ denotes the distance of the pixel i to the boundary of the ROI and $D_{\max}$ the maximum distance from the ROI boundary. Fig. 4(e) shows example of the division of the DCIS mask into 10 zones using this approach. The ratio between the area of the

hematoxylin stained pixels in region Z_k to the area of Z_k are defined to characterize the distribution of the nuclei within the ROI.

The final subset of architectural features include three measures to quantify the clustering of background and eosin stained pixels. These features are the maximum of the distance transform of the inverse eosin mask, inverse background mask, and the inverse of the union of the two masks.

3) Classification of Anatomically Meaningful Regions: The extracted texture and architectural features yielded a total of 1054 features. The performance of three classifiers were evaluated: logistic regression (LR) with L1 regularization ($\lambda = 1$), support vector machine (SVM) with a radial basis function (RBF) kernel ($\text{gamma} = 10^{-5}$ and $\text{cost} = 10^4$), and gradient boosted classifier with decision trees (GBC) [23] (with 1000 estimators and learning rate of 0.1). The three classifiers were trained and evaluated on separate training and test sets. The parameters of the classifiers were optimized with cross-validation on the training set. All the parameters of the multi-scale superpixel classification algorithm and the graph-clustering algorithm were defined using a subset of images in the training set.

III. Empirical Evaluation

A. Whole-Slide Histopathological Images of Breast Tissue and Ground Truth

Two image datasets were used in this study for empirical evaluation of the proposed DCIS detection system. The first dataset originates from 150 digitized H&E stained breast tissue slides sampled from 150 patients. Each slide was reviewed independently by two expert breast pathologists (RH and PB) and assigned a pathological diagnosis. 75 of the WSIs contained DCIS (grade I (9), grade II (35), grade III (31)) and 56 contained different types of benign lesions (usual ductal hyperplasia (11), adenosis (8), fibrosis (7), duct ectasia (5), fibrocystic (5), hamartoma (4), pseudoangiomatous stromal hyperplasia (3), sclerosing lobular hyperplasia (2), fibroadenoma (1), and mixed benign lesions (14)) and 19 normals. This dataset was taken from the archives of the department of Pathology. To be able to train and test our algorithm on different benign lesions that may occur in routine diagnostics, we enriched the benign class with cases containing all types of benign lesions as listed in the national Dutch breast cancer guidelines. Relative occurrence of these lesions in our dataset is comparable to that encountered in routine diagnostics.

The second dataset used in this study is representative of the daily clinical practice of breast pathology examined by a pathologist during a specific period of time. We took all cases from routine diagnostics of one breast pathologist involved in this study (PB) during the period June 2015 to September 2015, containing either DCIS or normal/benign conditions. This dataset consisted of 55 digitized H&E stained breast tissue slides sampled from 43 patients. This dataset contained 20 slides with DCIS diagnosis (grade I (5), grade II (7), grade III (8)) and 35 benign/normal slides (normal (12), calcification (9), usual ductal hyperplasia (5), fat necrosis (5), cyst (4)). Because the second dataset is taken consecutively from our routine diagnostics, in comparison to the first set which is enriched for benign abnormalities, the second set contains much fewer of the various benign lesion categories present in the first set.

In this study, we excluded the slides containing atypical ductal hyperplasia (ADH). The major problem with ADH is the difficulty in achieving acceptable levels of concordance or consistency in diagnosis [24]. Due to the use of different criteria for defining the characteristics of ADH in the literature [25]–[27] and the difficulty in obtaining reliable ground truth, we chose to exclude this category in our study.

All slides were stained in our laboratory and digitized using the 3DHISTECH Pannoramic 250 Flash II digital slide scanner with a 20X objective lens. Each image has square pixels of size 0.243 μm$\times$ 0.243 μm in the microscope image plane.

A total of 823 regions containing DCIS in abnormal slides from the first dataset were annotated. All the annotations were verified by two pathologists (RH and PB) independently. We included a lesion as ground truth in case both pathologists were in agreement. No annotation was provided for the slides with benign/normal diagnosis, as the training samples from these slides were automatically extracted using our automatic ROI detection algorithm. The ground truth data for the second dataset is only available at the slide level.

B. Experiments

To evaluate the performance of the proposed DCIS detection system two experiments were performed. The first experiment evaluates the efficacy of the proposed system in detecting and localizing DCIS regions in WSIs using the first dataset. The second evaluates the performance of the system in classifying a WSI as DCIS at the slide level using the second dataset.

1) Experiment 1: For this experiment, the first dataset was split into two independent subsets for training and testing. The train set comprises of 50 DCIS slides and 50 benign/normal slides (attempting to balance different DCIS grades and benign lesion categories over train and test sets). The test set comprises 25 DCIS slides and 25 benign/normal slides.

The training samples from the benign/normal category were automatically extracted using the ROI detection and delineation step of our proposed system. The training samples for the DCIS lesions, however, were taken directly from the annotated ROIs.

The performance of the proposed system was evaluated in terms of detecting and localizing the lesion in the slide. A ground truth DCIS lesion was deemed to have been detected if its intersection with the segmentation of the DCIS region performed by the proposed algorithm was non-empty. For the evaluation, free-response receiver operating characteristic (FROC) curve [28] was used. The FROC curve is defined as the plot of sensitivity versus the average number of false positives per image. The FROC curve is computed by varying thresholds on DCIS classification confidence. Considering that not all the DCIS lesions present in abnormal slides may have been annotated, the false positives were only counted in benign/normal slides.

In this experiment, we also evaluate the segmentation performance of the proposed system by computing Dice's overlap measure at the slide level.

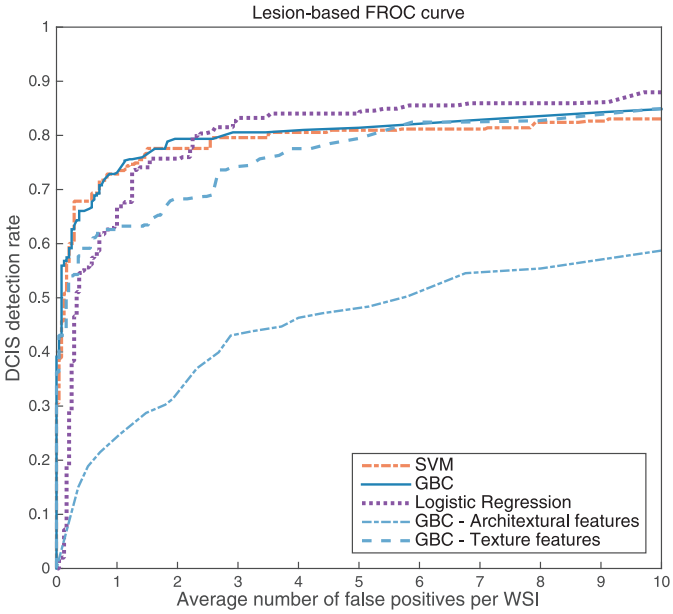

Fig. 5. Lesion-based FROC curve of the proposed DCIS detection system for experiment 1.

2) Experiment 2: The aim of this experiment was to evaluate the performance of the proposed detection system on an independent dataset representing the daily clinical practice of breast pathology examined by a pathologist. Each of the classifiers was trained independently on the entire slides present in the first dataset and evaluated on the second dataset. The parameters of the classifiers were kept the same as the first experiment.

This experiment evaluates the performance of the system in differentiating between the slides containing DCIS and benign/normal slides. To achieve a slide-based score, the highest scored region in a slide is used as the likelihood score that the case contains DCIS.

Slide-based FROC analysis was performed to evaluate the efficacy of the system. The FROC curve in this experiment plots the fraction of slides classified as DCIS divided by the total number of slides with DCIS versus the average number of false positives per WSI.

C. Results

Fig. 5 presents the FROC curve of experiment 1 for the three classifiers. Note that the false positive rate plotted on the horizontal axis is counted on benign/normal slides only. The FROC curve for the GBC is also presented when only texture features or architectural features were used. Table I summarizes the DCIS detection (sensitivity) levels at different average numbers of false positives per WSI, for different classifiers. Overall, the three classifiers achieved comparable performance. SVM and GBC yielded higher sensitivities at smaller numbers of false positives while LR performed better at larger number of false positives. Fig. 6 shows examples of true positives, false positives, as well as false negatives obtained by the CAD system trained by GBC when the performance was fixed at 80% sensitivity.

For the evaluation of the performance of the segmentation algorithm we used Dice's overlap measure. The Dice score at the slide level when considering only the detected DCIS lesions (sensitivity was fixed at 80% in experiment 1) was 0.9243 ± 0.0187 (mean $\pm$ std) over the entire slides in the test set of the first dataset.

TABLE I
RESULTS OF THE EXPERIMENT 1: SENSITIVITY OF DCIS LESION DETECTION IS PROVIDED AT FIVE LEVELS OF AVERAGE NUMBERS OF FALSE POSITIVES (FPS) PER WSI

FPs/WSI	1/2	1	2	3	4
GBC	0.66	0.73	0.80	0.81	0.81
SVM	0.68	0.73	0.78	0.80	0.80
LR	0.50	0.63	0.76	0.83	0.84

In the second experiment GBC yielded the best performance. The FROC curve of Experiment 2 for GBC is shown in Fig. 7. The 95% confidence interval was generated using patient-stratified bootstrapping with 1000 replications. Table II summarizes the slide-based DCIS classification sensitivity at different average numbers of false positives per WSI, for different classifiers. Overall, GBC yielded the best performance, achieving a sensitivity of 95% and 100% at average false positive rates of 2 and 2.6, respectively.

IV. DISCUSSION AND CONCLUSION

In this paper, we presented a CAD system for DCIS detection in digitized H&E stained histopathological breast tissue slides. The proposed algorithm is fully automated, does not require any human interaction, and therefore yields objective and reproducible results. Lesion-based and slide-based evaluation of the performance of the proposed CAD system was presented. Collectively, the results of the experiments demonstrate the efficacy and accuracy of the proposed CAD system as well as its potential for application in routine pathological diagnostics.

To the best of the authors' knowledge, this is the first fully automated DCIS CAD system that operates at the WSI level and has been evaluated on a dataset collected from routine clinical practice. WSI analysis of histopathological slides remains a challenging medical image analysis problem due to technical complexities in dealing with large WSIs and the requirement to have highly specific algorithms to avoid large numbers of false positives. The focus in developing CAD systems for histopathological images, in particular for the task of recognizing DCIS from different types of benign abnormality, has been mainly limited to analyzing small patch images selected by a pathologist [4]–[6]. Existing approaches to localize diagnostically relevant regions in WSIs either analyze the WSI at lower magnification or divide the image into rectangular patches and classify them (possibly at multiple resolutions) [7]–[9]. In this study we proposed a multi-scale superpixel classification scheme for finding epithelial areas in WSIs.

Detection and contouring of diagnostically relevant regions was based on a spatial clustering approach operating on the graphs built on the centroids of epithelium labeled superpixels. Several algorithms have been published for the segmentation of glandular structures with application to prostate and colon tissue [29]–[32]. All of these methods assume an architectural regularity in glandular structure and have detection of lumen as an

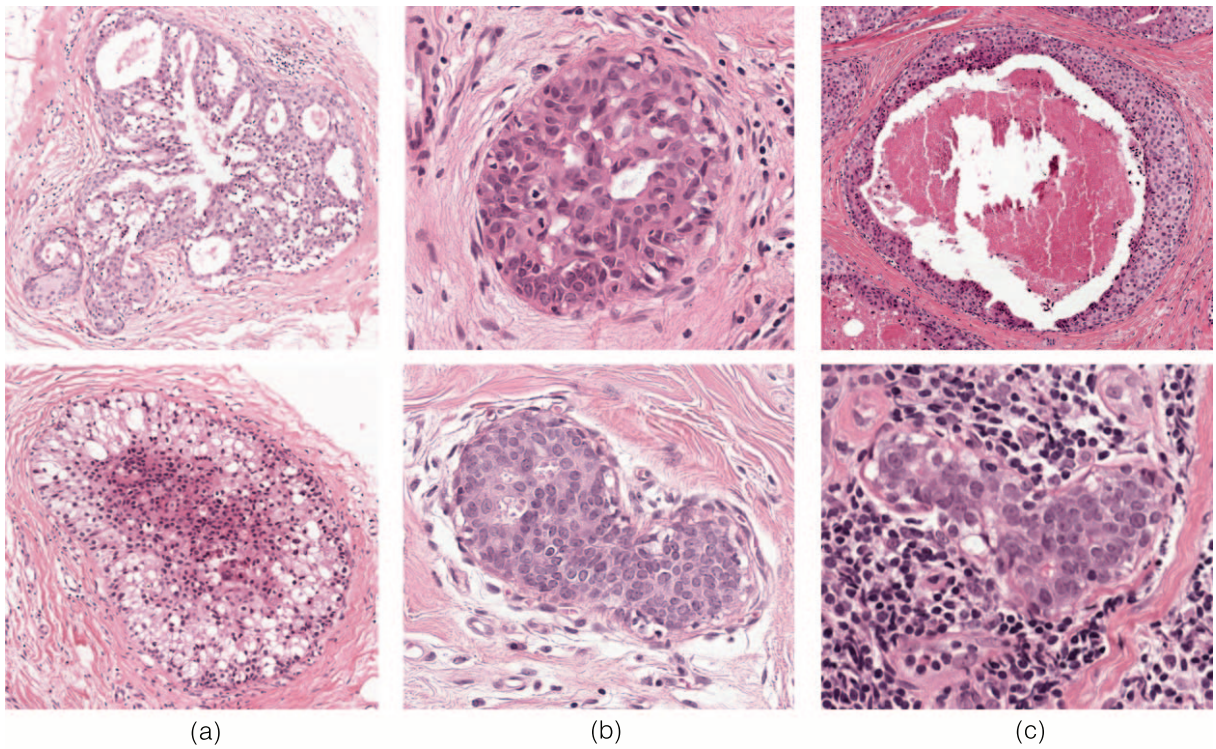

Fig. 6. Examples of true positives, false positives, as well as false negatives. (a) Shows examples of two correctly detected DCIS lesions. (b) Two false positive examples. (c) Examples of two missed DCIS lesions. The image on top shows a DCIS lesion with large amount of necrosis, and the image in bottom shows an example of a DCIS lesion (lobular cancerization) surrounded by lymphocytes.

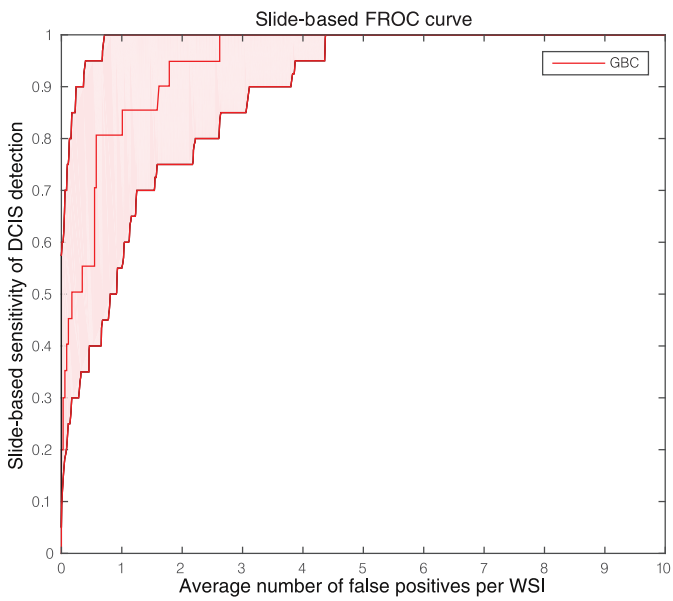

Fig. 7. Slide-based FROC curve and the 95% confidence intervals of the proposed DCIS detection system for experiment 2.

TABLE II
RESULTS OF THE EXPERIMENT 2: SLIDE-BASED SENSITIVITY OF DCIS DETECTION IS PROVIDED AT FIVE LEVELS OF AVERAGE NUMBERS OF FALSE POSITIVES (FPS) PER WSI

FPs/WSI	1/2	1	2	3	4
GBC	0.55	0.80	0.95	1.0	1.0
SVM	0.70	0.70	0.85	0.85	0.85
LR	0.55	0.75	0.90	0.95	0.95

essential step for segmenting the gland. The intraductal proliferation in the breast, however, usually obliterates and distends the ductal lumen [33] which limits the effectiveness of these methods in detecting DCIS lesions of the breast. The approach proposed by Sirinukunwattana *et al.* [34] does not have this limitation as it does not necessitate any strict assumption regarding the arrangement of granular components. However, inferring the number of glands in a clustered population of glands is based on the number of isolated connected components resulted from thresholding the glandular probability map. This means partially connected glands may fall in the same connected component and there is no mechanism in the utilized random polygon model (RPM) to further split these glands. Moreover, although the proposed algorithm yields good results in segmenting glands in colon tissue, due to the stochastic modeling nature of RPM, the proposed model has high computational complexity and may not be suitable for application to WSIs. Our proposed spatial clustering algorithm, in contrast, is robust, efficient and well suited for accurate detection and delineation of breast glandular structures in WSIs. Evaluation of the segmentation performance in experiment 1 demonstrate that our spatial clustering algorithm yields a Dice score of 0.9243 ± 0.0187 for segmenting DCIS regions.

Following the segmentation of the diagnostically relevant regions in the WSI, a set of texture-based and architectural features were extracted from the epithelial structure. Fig. 5 presented the contribution of the proposed architectural features to the performance of the detection system. Our proposed features are efficient to compute and obviate the need to perform nuclear segmentation for describing the distribution of the struc-

tures inside the potential DCIS region. Our evaluations on the first dataset demonstrate the efficacy of the proposed method in detecting and localizing DCIS. Using the proposed system, it is possible to detect 80% of the DCIS lesions in an abnormal slide at an average number of 2 false-positive per WSI. Practically, we expect the time gain in automatically detecting 80% of the DCIS lesions in WSIs outweighs the time lost for looking at the false positives.

In this study, we also presented an evaluation on a dataset collected from routine clinical practice during a four month period. This dataset contains large categories of benign lesions that the pathologist encounters in routine diagnostics. Unlike the previous studies which mainly focus on discrimination of DCIS from UDH, the proposed system was designed to handle a large set of heterogeneous benign categories. Evaluation of the slide-based DCIS detection on this dataset shows that at an average number of 2.6 false-positive per WSI, 100% of the slides that contained DCIS could be detected. This suggests the potential of the proposed method for application in routine pathological diagnostics. Moreover, further reduction of the average number of false positives per WSI can significantly reduce the workload of the pathologist as it would mean that a large number of normal/benign slides can be put aside without the risk of missing slides containing DCIS.

The proposed system has several components, of which only a small number may impact the performance. The multi-resolution superpixel classification algorithm utilizes two classifiers trained on manually labeled superpixels. By using our recently proposed algorithm for standardization of WSIs [18], we can obviate the need for re-training these classifiers when applied to new datasets. However, the images in this study were not standardized as the slides were stained using the same protocol and scanned using the same scanner. We have found that the specific choice of parameters for many components of our system such as the threshold applied to determine the uncertainty of the classification of superpixels, the number of iteration for graph growing operation, and the threshold applied for coarse clustering the structures in the WSI, are relatively unimportant, and serve mainly to reduce computational cost. Moreover, the designed architectural features are based on an initial classification of the image into different stain classes. The utilized method is an intermediate step in our whole slide standardization algorithm and as shown in our paper [18] it is very robust against variations in histopathological images.

The computation time for different steps of the proposed system to analyze a WSI is as follows. The multi-scale superpixel classification algorithm for finding epithelial regions on the WSI takes between 20 to 45 minutes depending on the amount of tissue (in particular epithelial tissue) on the slide. The graph-clustering algorithm is very efficient and takes on average less than two minutes to generate segmented ROIs. The feature extraction and classification stage together take on average 10 minutes. The implementation is done in C++ and the experimental platform was a laptop with an Intel Core i7 CPU (2.4 GHZ) and 16 GB of Ram.

Several limitations of the proposed method must be acknowledged. First, the multi-scale superpixel classification algorithm puts lymphocytic infiltrates within the same category as the epithelial class. This may occasionally cause the system to classify the lesions surrounded by lymphocytes as normal. The major reason for this is that the graph clustering algorithm for delineation of the candidate may result in a region including both lymphocytes and DCIS nuclei, hence polluting the statistics of the DCIS region. Lymphocytes are frequently abundant in benign slides, and less-existent in the annotation of the DCIS regions. A region containing a large number of lymphocytes may therefore be characterized as normal by the system (see Fig. 6(c) for a false negative). Another limitation of the proposed system is in dealing with lesions having large areas of necrosis and very little epithelium. Fig. 6(c) shows an example of a DCIS lesion with such characteristics. Possible reasons explaining the difficulty of the proposed system in dealing with these lesions are the lack of training data for such lesions and the significant deviation of the characteristics presented by these lesions compared to the majority of the DCIS lesions.

The proposed system was primarily designed to aid the pathologist in detecting and localizing the lesions in the WSI and giving a second opinion on the malignancy likelihood of the findings. The proposed system, however, has the potential to be applied to related problems, such as detecting and classifying glands in prostate tissue WSIs. In addition, it has provided another important implication for future research. The proposed system can serve as an important first step for development of systems that aim at finding prognostic and predictive biomarkers within malignant lesions, requiring an accurate delineation of such regions. This will be the major direction for future research.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support from the histology laboratory of the Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA A Cancer J. Clinicians*, vol. 65, no. 1, pp. 5–29, 2015.

[2] I. O. Ellis, "Intraductal proliferative lesions of the breast: Morphology, associated risk and molecular biology," *Modern Pathol.*, vol. 23, pp. S1–S7, 2010.

[3] P. T. Simpson, J. S. Reis-Filho, T. Gale, and S. R. Lakhani, "Molecular evolution of breast cancer," *J. Pathol.*, vol. 205, no. 2, pp. 248–254, 2005.

[4] F. Dong *et al.*, "Computational pathology to discriminate benign from malignant intraductal proliferations of the breast," *PLoS ONE*, vol. 9, no. 12, p. e114885, Dec. 2014.

[5] M. M. Dundar *et al.*, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 7, pp. 1977–1984, Jul. 2011.

[6] U. Srinivas *et al.*, "SHIRC: A simultaneous sparsity model for histopathological image representation and classification," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, 2013, pp. 1118–1121.

[7] M. Peikari, M. Gangeh, J. Zubovits, G. Clarke, and A. Martel, "Triaging diagnostically relevant regions from pathology whole slides of breast cancer: A texture based approach," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 307–315, Jan. 2015.

[8] S. Doyle, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "A boosting cascade for automated detection of prostate cancer from digitized histology," in *Proc. MICCAI*, 2006, pp. 504–511.

[9] O. Sertel *et al.*, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," *Pattern Recognit.*, vol. 42, no. 6, pp. 1093–1103, 2009.

[10] B. E. Bejnordi, G. Litjens, M. Hermsen, N. Karssemeijer, and J. A. van der Laak, "A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images," in *Proc. SPIE Med. Imag.*, 2015, pp. 94 200H–94 200H.

[11] R. Achanta *et al.*, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transact. on Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[12] J. A. van der Laak, M. M. Pahlplatz, A. G. Hanselaar, and P. de Wilde, "Hue-saturation-density (HSD) model for stain recognition in digital images from transmitted light microscopy," *Cytometry*, vol. 39, no. 4, pp. 275–284, 2000.

[13] J. Jaromczyk and G. Toussaint, "Relative neighborhood graphs and their relatives," *Proc. IEEE*, vol. 80, no. 9, pp. 1502–1517, Sep. 1992.

[14] E. F. Moore, "The shortest path through a maze," in *Proc. Int. Symp. Theory Switch.*, 1957, pp. 285–292.

[15] R. Xu *et al.*, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[16] S. E. Schaeffer, "Graph clustering," *Comput. Sci. Rev.*, vol. 1, no. 1, pp. 27–64, 2007.

[17] H. Dörrie*, 100 Great Problems of Elementary Mathematics*. Mineola, NY: Courier, 2013.

[18] B. Ehteshami Bejnordi *et al.*, "Stain specific standardization of whole-slide histopathological images," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 404–415, Feb. 2015.

[19] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man Cybern.*, no. 6, pp. 610–621, 1973.

[20] M. Topi, O. Timo, P. Matti, and S. Maricor, "Robust texture classification by subsets of local binary patterns," in *Proc.. 15th Int. Conf. Pattern Recognit.*, 2000, vol. 3, pp. 935–938.

[21] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *Proc. 5th IEEE Int. Symp. Biomed. Imag. From Nano to Macro*, 2008, pp. 496–499.

[22] S. Naik *et al.*, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *Proc. 5th IEEE Int. Symp. Biomed. Imag. From Nano to Macro*, May 2008, pp. 284–287.

[23] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, pp. 1189–1232, 2001.

[24] S. E. Pinder and I. O. Ellis, "Ductal carcinoma in situ (DCIS) and atypical ductal hyperplasia (ADH)-current definitions and classification," *Breast Cancer Res.*, vol. 5, no. 5, pp. 254–257, 2003.

[25] D. Page, "Atypical hyperplastic lesions of the female breast: A long-term follow-up study," *Plastic Reconstructive Surg.*, vol. 77, no. 4, p. 688, 1986.

[26] D. L. Page and L. W. Rogers, "Combined histologic and cytologic criteria for the diagnosis of mammary atypical ductal hyperplasia," *Human Pathol.*, vol. 23, no. 10, pp. 1095–1097, 1992.

[27] F. Tavassoli, "Intraduct hyperplasias, ordinary and atypical," *Pathol. Breast*, pp. 155–191, 1992.

[28] P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free response approach to the measurement and characterization of radiographic observer performance," in *Application of Optical Instrumentation in Medicine VI.* Bellingham, WA: SPIE, 1977, pp. 124–135.

[29] K. Nguyen, A. Sarkar, and A. K. Jain, "Prostate cancer grading: Use of graph cut and spatial arrangement of nuclei," *IEEE Trans. Med. Imag.*, vol. 33, no. 12, pp. 2254–2270, Dec. 2014.

[30] S. Naik, S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information," in *MIAAB Workshop*, 2007, pp. 1–8.

[31] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer, "Automatic segmentation of colon glands using object-graphs," *Med. Image Anal.*, vol. 14, no. 1, pp. 1–12, 2010.

[32] A. Fakhrzadeh, E. Sporndly-Nees, L. Holm, and C. L. L. Hendriks, "Analyzing tubular tissue in histopathological thin sections," in *Proc. Int. Conf. Digital Image Comput. Tech. Appl.*, 2012, pp. 1–6.

[33] M. Guray and A. A. Sahin, "Benign breast diseases: Classification, diagnosis, and management," *Oncologist*, vol. 11, no. 5, pp. 435–449, 2006.

[34] K. Sirinukunwattana, D. R. Snead, and N. M. Rajpoot, "A stochastic polygons model for glandular structures in colon histology images," *IEEE Trans. Med.l Imag.*, vol. 34, no. 11, pp. 2366–2378, Nov. 2015.