Pattern Recognition 84 (2018) 345–356

Contents lists available at ScienceDirect

Pattern Recognition

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

Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks

Baris Gecer[a,1], Selim Aksoy[a,*], Ezgi Mercan[b], Linda G. Shapiro[b], Donald L. Weaver[c], Joann G. Elmore[d,2]

[a] Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey
[b] Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA 98195, USA
[c] Department of Pathology, University of Vermont, Burlington, VT 05405, USA
[d] Department of Medicine, University of Washington, Seattle, WA 98195, USA

ARTICLE INFO

Article history:
Received 22 August 2017
Revised 13 May 2018
Accepted 16 July 2018
Available online 20 July 2018

Keywords:
Digital pathology
Breast histopathology
Whole slide imaging
Region of interest detection
Saliency detection
Multi-class classification
Deep learning

ABSTRACT

Generalizability of algorithms for binary cancer vs. no cancer classification is unknown for clinically more significant multi-class scenarios where intermediate categories have different risk factors and treatment strategies. We present a system that classifies whole slide images (WSI) of breast biopsies into five diagnostic categories. First, a saliency detector that uses a pipeline of four fully convolutional networks, trained with samples from records of pathologists' screenings, performs multi-scale localization of diagnostically relevant regions of interest in WSI. Then, a convolutional network, trained from consensus-derived reference samples, classifies image patches as non-proliferative or proliferative changes, atypical ductal hyperplasia, ductal carcinoma in situ, and invasive carcinoma. Finally, the saliency and classification maps are fused for pixel-wise labeling and slide-level categorization. Experiments using 240 WSI showed that both saliency detector and classifier networks performed better than competing algorithms, and the five-class slide-level accuracy of 55% was not statistically different from the predictions of 45 pathologists. We also present example visualizations of the learned representations for breast cancer diagnosis.

1. Introduction

Breast cancer is the most widespread form of cancer among women [1]. There can be many types of deviations from a healthy tissue, where some are considered benign and some are indicators for cancer. The detection and categorization of these deviations are not always straightforward even for experienced pathologists. Histopathological image analysis promises to play an important role in helping the pathologists by indicating potential disease locations and by aiding their interpretation.

There is a large body of work on the classification of histopathological images. Most use generic color- or texture-based features and nuclear architectural features with classifiers such as support vector machines (SVM) or random forests (RF) [1,2]. The most common scenario is to use manually cropped regions of interest (ROI) that have no ambiguity regarding their diagnoses. Even though these approaches can provide insights about which features are useful for classification, it is very difficult to design and tune them with respect to the extensive structural diversity found in whole slide images (WSI) that are obtained by digitization of entire glass slides [3]. In particular for breast pathology, the variations in the tissue structure that range from non-proliferative changes to proliferative ones such as usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), and invasive ductal carcinoma (IDC) provide challenges to both experienced and novice pathologists [4]. Furthermore, subtle differences among these categories lead to different clinical actions, and the following treatments with different combinations of surgery, radiation, and hormonal therapy make the diagnostic errors extremely significant in terms of both financial and emotional consequences [4,5].

Unfortunately, the generalizability of the state-of-the-art image features and classifiers that have been designed and evaluated for

* Corresponding author.
E-mail addresses: b.gecer@imperial.ac.uk (B. Gecer), saksoy@cs.bilkent.edu.tr (S. Aksoy), ezgi@cs.washington.edu (E. Mercan), shapiro@cs.washington.edu (L.G. Shapiro), Donald.Weaver@vtmednet.org (D.L. Weaver), jelmore@u.washington.edu (J.G. Elmore).
[1] Present address: Deptartment of Electrical and Electronic Engineering, Imperial College London, UK
[2] Present address: Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, USA

https://doi.org/10.1016/j.patcog.2018.07.022
0031-3203/

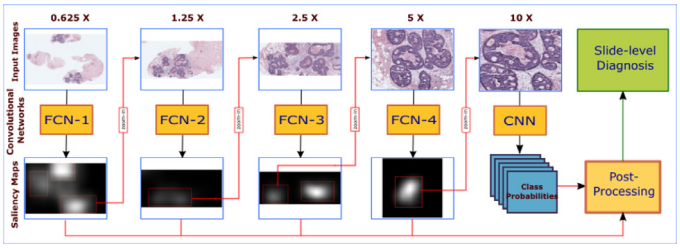


Fig. 1. Overview of the proposed framework. Salient regions are detected on the input WSI by feed-forward processing of FCN-1. Each connected component that has a probability of being diagnostically relevant above a threshold is zoomed in and passed to FCN-2. This process is repeated four times, and the detected salient regions are processed by the classification CNN to obtain the likelihood maps for five diagnostic classes. The detection results and the classification results are fused to obtain the final slide-level diagnosis.

the more restricted, often binary, settings is currently unknown for whole slides that contain multiple areas with different structural deviations that correspond to different levels of diagnostic importance. Even though the final diagnosis is decided based on the most severe one of these areas, existence of different levels of structural anomalies in the same slide often distracts pathologists as shown in eye tracking studies [6]. Thus, automatic detection of diagnostically relevant ROIs can decrease the pathologists' workloads while also assuring that no critical region is overlooked during diagnosis. Such solutions will also benefit computer aided diagnosis by eliminating a significant amount of computation and lead to efficient use of computational resources in more detailed WSI analysis.

In this paper, we study both the detection and the multi-class classification of diagnostically relevant regions in whole slide breast histopathology images using deep networks. Our main contributions are threefold. First, we propose a saliency detection framework for automatic localization of ROIs. Our method uses four separate fully convolutional networks (FCN) trained to imitate the actions of pathologists at different magnifications. Although selecting the right magnification is a common goal in the literature, we go beyond that motivation, and use a data-driven feature learning approach that exploits the recorded viewing behaviors of pathologists where zoom actions are used to construct training samples. These networks progressively eliminate irrelevant areas from lower to higher magnifications, and the combined result provides a saliency map for the WSI. Second, we present another convolutional neural network (CNN) for the identification of five diagnostic categories of ductal proliferations (non-proliferative changes, proliferative changes, ADH, DCIS, and IDC) in whole slides. We consider saliency detection and classification of salient regions as two separate but sequential applications where the proposed modular solutions can also be used in distinct applications. Furthermore, we fuse the outputs of ROI detection and classification steps for slide-level diagnosis. Third, we visualize the resulting networks for better understanding of the learned models in differentiating cancer categories.

An overview of the proposed approach is shown in Fig. 1. The rest of the paper is organized as follows. Section 2 discusses the related work, Section 3 introduces the data set, Section 4 describes the methodology for both ROI detection and classification, Section 5 presents the experimental results, and Section 6 summarizes the conclusions.

2. Related work

The related literature on WSI analysis resorted to restricted classification settings. For example, Dundar et al. [7] used multiple instance learning for discrimination of benign cases from actionable (ADH+DCIS) ones by using whole slides with manually identified ROIs. Dong et al. [8] built a logistic regression (LR) classifier for UDH versus DCIS classification where each WSI was modeled with manually cut ROIs. Some approaches to WSI analysis have focused on efficient applications of existing methods by using multi-resolution [9–11] and multi-field-of-view [12] sliding windows. Even though exhaustive window-based processing of WSIs is an alternative to manually selected ROIs, tiling usually involves arbitrary splits of the image and has the risk of distorting the context. Balazsi et al. [13] tried to overcome the effect of fixed tiling by using color, texture and gradient histogram features extracted from superpixels with RF classifiers for IDC versus normal classification. They concluded that generic features were sufficient for detecting invasive carcinoma but differentiating DCIS from IDC was still a problem. We recently introduced a multi-instance multi-label learning framework to study the uncertainty regarding the correspondence between the pathologists' slide-level annotations and the candidate ROIs extracted from their viewing records for weakly supervised learning using WSI [14].

As one of the rare studies on automatic ROI detection, Bahlmann et al. [15] used color histograms of square patches with linear SVMs for classification as relevant versus irrelevant. Numerical results were given only for a small set of patches. We developed a bag-of-words model using color and texture features of image patches as well as superpixels with SVM and LR classifiers trained using samples extracted from the logs of pathologists' image screenings for ROI detection [16,17]. The results of the proposed method are compared to the results of this model in Section 5. Bejnordi et al. [18] classified superpixels at three scales with a large set of features and RF classifiers for progressive elimination of irrelevant areas, and used graph-based clustering of the resulting superpixels with a heuristic set of rules to obtain the ROIs. However, evaluation was done on manually annotated DCIS cases where ADH instances were excluded due to the difficulty of the problem.

Recent advances in computer vision have demonstrated that feature learning approaches using deep networks can be more successful than hand-crafted features. Such approaches have found applications in histopathology as well. For example, Cruz-Roa et al. [19] showed that a three-layer convolutional neural network (CNN) that operated on 100×100 pixel patches at $2.5\times$ magnification was more successful than color, texture, and graph-based features with an RF classifier in the detection of IDC. Litjens et al. [20] used a deep network with 128×128 pixel patches at $5\times$ magnification for the delineation of prostate cancer. Janowczyk and Madabhushi [3] illustrated the use of deep learning for several tasks including IDC detection using 32×32 pixel patches at $2.5\times$ magnification. CNN-based cell features were also shown to improve the accuracy of graph hashing for histopathology image classification and retrieval in [21]. Other popular applications where deep learning methods achieved the top scores in competitions include mitosis detection [3] and metastatic breast cancer detection in lymph nodes [20,22]. The common characteristics that lead to the success of deep learning in these applications are the suitability of finding an appropriate magnification at which the object of interest and the relevant context can be fit within a fixed size patch and the availability of millions of training examples. The large amount of variation in the sizes of the structures of interest and the lack of large amount of labeled data for the multi-class scenario that considers both pre-invasive and invasive stages of breast cancer presents outstanding challenges to both traditional and deep learning-based approaches.

Besides this work, the only other deep learning study that considered this challenging range of histologic categories reflecting the actual clinical practice is [23] that proposed a novel structural feature for breast pathology. First, a multi-resolution network with

Table 1
Distribution of diagnostic classes among the 180 training and 60 test slides.

Class	Training	Test
Non-proliferative changes only (NP)	8	5
Proliferative changes (P)	50	13
Atypical ductal hyperplasia (ADH)	50	16
Ductal carcinoma in situ (DCIS)	55	21
Invasive ductal carcinoma (IDC)	17	5

two multi-path encoder-decoders and input-aware encoding blocks was used for pixel-based segmentation of ROIs into eight tissue types [24]. Then, superpixels were used as the structural elements that aggregated the pixel labels, and the connected components of the sections marked as epithelium, secretion and necrosis were used to estimate the locations of ducts. Finally, the structural feature was extracted by computing histograms of these tissue types within several layers, defined 1-superpixel thick, towards both the inside and the outside of these ductal components. The structure feature was used to classify each ROI by using a four-class SVM (benign, ADH, DCIS and invasive) and by using a sequence of binary SVMs that eliminate one diagnosis at a time (invasive vs. not-invasive, ADH and DCIS vs. benign, and DCIS vs. ADH). The results of that method are also discussed in Section 5.

3. Data set

We used 240 digital breast histopathology images that were collected as part of NIH-sponsored projects [4,25,26]. The haematoxylin and eosin (H&E) stained slides were selected from registries associated with the Breast Cancer Surveillance Consortium by using a random stratified method to include the full range of diagnostic categories from benign to cancer and to represent a typical pathology lab setting. Each slide that belonged to an independent case from a different patient was scanned at $40\times$ magnification, resulting in an average image size of $100{,}000 \times 64{,}000$ pixels. The slides were divided into training and test sets, with 180 and 60 cases, respectively, by using stratified sampling based on age, breast density, original diagnosis, and experts' difficulty rating of the case so that both sets had the same class frequency distribution with cases from different patients. The distribution of classes is given in Table 1. ADH and DCIS cases were intentionally oversampled to gain statistical precision in the estimation of interpretive concordance for these diagnoses [25].

Three experienced pathologists who are internationally recognized in diagnostic breast pathology evaluated every slide both independently and in consensus meetings. The results of these meetings were accepted as the reference diagnosis for each slide including non-proliferative changes (including fibroadenoma), proliferative changes (including intraductal papilloma without atypia, usual ductal hyperplasia, columnar cell hyperplasia, sclerosing adenosis, complex sclerosing lesion, and flat epithelial atypia), atypical ductal hyperplasia (including intraductal papilloma with atypia), ductal carcinoma in situ, and invasive ductal carcinoma. Each slide in the test set also has independent interpretations from 45 other pathologists. The difficulty of the multi-class problem studied here can be observed from the evaluation in [4,26] where the individual pathologists' concordance rates compared with the reference diagnoses were 82% for the union of NP and P, 43% for ADH, 79% for DCIS, and 93% for IDC.

The data collection also involved tracking the experienced pathologists' actions while they were interpreting the slides using a web-based software tool for multi-resolution browsing of WSI data. In addition, the pathologists also marked an example ROI as a representative for the most severe diagnosis that was observed during their examination of each slide. Both these consensus ROIs and the individual viewing records of the three pathologists are used in the following sections. The diagnoses assigned by the other 45 pathologists are also used for comparison. The study was approved by the institutional review boards at Bilkent University, University of Washington, and University of Vermont.

4. Methodology

4.1. ROI detection

In this section, first, we describe how the training data were constructed from the tracking records of pathologists for building fully convolutional networks (FCN) for detection of ROIs in arbitrarily sized images. Then, we present the pipeline of four FCNs that process large images at different magnifications where areas evaluated as non-salient are incrementally eliminated from lower to higher resolutions. This step uses FCNs because they can take arbitrary sized inputs and can generate similar sized predictions that are suitable for detection and segmentation problems [27]. FCNs provide efficiency during both learning via end-to-end backpropagation and prediction via dense feedforward computation that is more advantageous over sliding window-based processing that involves redundant computation because of overlapping regions.

4.1.1. Data set preparation

The online software designed for the pathologists' interpretation of WSI supported pyramid structures with the original $40\times$ magnification as well as layers successively subsampled by a factor of 2 up to $0.625\times$. The software also provided intermediate resolutions by on-the-fly subsampling from the closest higher magnification. The tracking procedure recorded the coordinates of the windows corresponding to the parts of the WSI visible on the screen and mouse events at a frequency of four entries per second. Each of these log entries is named a viewport, and the sequence of viewports from a particular pathologist's interpretation of a particular slide is denoted as $l_t, t = 1, 2, \ldots, T$ in the analysis below. Motivated by the visual search patterns of the pathologists [28], we designed a selection process that evaluated the possibility of pairs of windows, (l_j, l_i), as being related during the pathologist's visual screening. In this process, the following rules were defined to assess whether a visited window (named the destination, l_j) was considered as salient by the pathologist at one of the earlier windows (named the source, l_i):

$$i < j, \tag{1}$$

$$zoom(l_i) < zoom(l_k), \quad \forall k \in \{i+1, \ldots, j\}, \tag{2}$$

$$zoom(l_j)/3 \leq zoom(l_i) \leq zoom(l_j)/1.5. \tag{3}$$

The first rule stated that the source window l_i must be visited before the destination l_j. The second rule ensured that the destination window was viewed at a higher magnification than the source window, and there was no zoom out action going to a lower magnification than the zoom level of the source window between the two windows. The third rule required that the zoom level of the source window was in a particular range so that there was sufficient context around the destination in which it was considered salient (e.g., when the zoom level of the destination l_j was 30, the zoom level of the source must be in the range [10,20]). Each viewport in our data set was evaluated as a potential destination, and if one or more sources that satisfied (1)–(3) were found, the earliest one was used to form the viewport pair. An example is given in Fig. 2. After evaluating all actions, the pairs that contained

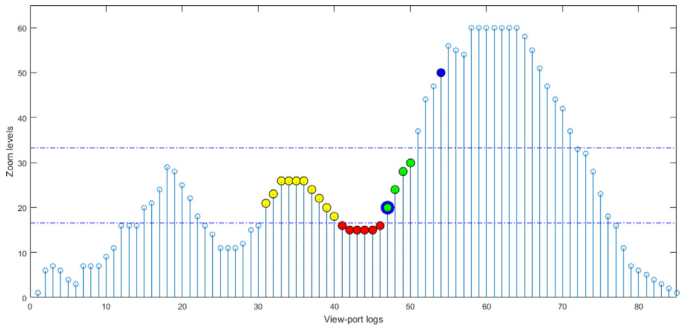


Fig. 2. Training sample generation from the viewport log of a pathologist. The *x*-axis shows the log entry index and the *y*-axis shows the zoom level. The blue dot represents an example destination window ($l_{j=54}$). The horizontal lines indicate the search range of zoom levels for a possible source window as defined in (3). The red dots are eliminated according to this rule. The yellow dots violate (2). The green dots satisfy all three conditions, and the earliest one ($l_{i=47}$), marked with a blue ring, is selected as the source window. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

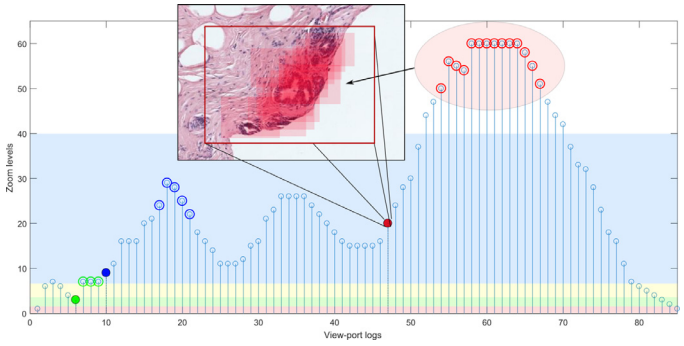


Fig. 3. Training sample generation from the viewport logs (cont.). Grouped viewports are shown with the same color where the filled dots are the source windows and the dots with rings represent the destination windows. A data sample is illustrated for the red group where the source window ($l_{i=47}$) defines the input image data within the WSI, and the union of destination windows ($l_{j=54,\ldots,67}$) are used to construct the saliency label mask. The four ranges of zoom levels that are considered for training four different FCNs are also shown: FCN-1 with $zoom(l_i) = 1$ (red), FCN-2 with $2 \leq zoom(l_i) \leq 3$ (green), FCN-3 with $4 \leq zoom(l_i) \leq 6$ (yellow), and FCN-4 with $7 \leq zoom(l_i) \leq 40$ (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

common source windows were grouped together, and each group was used to create one data sample where the input was the raw image corresponding to the common source window (l_i) and the label was a same sized pixel-level binary mask where the union of destination windows (l_js) in the group were marked as positive (salient). An example is given in Fig. 3.

Training samples were collected from the viewing records of the three experienced pathologists for the 180 training images. The resulting samples were split into four sets according to the zoom levels of the source windows. These sets, shown in Fig. 3, formed the training data for four separate deep networks where each focused on specific contextual cues in a particular range of magnifications. The four training sets consisted of a total of 64,144 images with an average size of 535×416 pixels. The total number of pixels labeled as negative was around five times as many as those that were labeled as positive.

4.1.2. Network architecture for detection

Our network architecture and the related learning parameters were influenced from the deep network presented in [29] because of its success in the ImageNet challenge and simple strategies. Nevertheless, we ensured that the sizes of the receptive fields of the convolutional layers were compatible with the fundamental characteristics of the biopsies such as ductal structures.

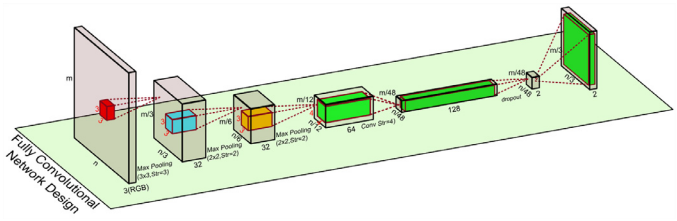


Fig. 4. Illustration of the FCN architecture for ROI detection. The number and size of the filters at each layer are given. All convolutional layers are followed by ReLU nonlinearity. We also show the corresponding image size at each layer for an input of $m \times n$ pixels. Note the deconvolutional layer at the end.

Our fully convolutional network architecture is shown in Fig. 4. The inputs were arbitrary sized RGB images that were collected as in the previous section. Input images were preprocessed by subtracting the overall mean of RGB values of training images from each pixel. The image was then passed through three similar convolutional layers, as in [29], where filters had a very small width and height (3×3) followed by a ReLU nonlinearity unit. Convolutional stride and spatial padding was set to 1 pixel such that the spatial resolution was preserved. ReLU was followed by the max pooling operation with a 3×3 pixel window and a stride of 3 after the first layer, and a 2×2 window and a stride of 2 after the remaining layers. These three convolutional layers were followed by another convolutional layer with a 4×4 window size and a convolutional stride of 4. This layer included a ReLU nonlinearity but no max pooling operation. After that, there was one fully connected layer (which was, in fact, a 1×1 convolutional layer in FCN) followed by a dropout operation with a rate of 0.5. The network continued with a deconvolutional layer with an upsample rate of 16 times and cropping of 32 pixels from all sides. Number of filters in all layers were 32, 32, 64, 128, 2, respectively. The final layer was connected to the 'multinomial logistic loss' (softmax log loss) objective function layer while training, but after training, we removed that layer and added a 'softmax' layer to estimate class (relevant versus irrelevant) probabilities. The hyper-parameters of the network architecture were tuned on one-fifth of the training set as validation data. Given an input image with a size of $m \times n$ pixels, the resulting map of size $m/3 \times n/3$ that was relative to the input was an advantage of the fully convolutional design that improved the precision of detection and localization of salient regions.

4.1.3. Pipeline

We designed a pipeline that gradually eliminated diagnostically irrelevant regions efficiently in four successive stages where the ultimate output was a saliency map of the input image. A given image was processed by four networks that had the same architecture but were trained to handle images at different magnifications as shown in Fig. 1. Let Φ represent the input image at $40\times$ magnification along with the constructed multi-resolution pyramid. For ROI detection, we used the $0.625\times$, $1.25\times$, $2.5\times$, and $5\times$ magnifications, denoted as Φ_1, Φ_2, Φ_3, and Φ_4, respectively, corresponding to the zoom level ranges shown in Fig. 3.

The analysis started with the smallest magnification, Φ_1, being fed to the first network FCN_1 to produce the saliency map Θ_1. Then, the regions with probability of being diagnostically relevant above a particular threshold were fed to the second network. The same procedure was repeated for the remaining stages. The final saliency map Θ was computed as the weighted geometric mean [30] of the thresholded outputs of four networks, Θ_1, Θ_2, Θ_3, and

Θ_4, as

$$\Theta = \prod_{k=1}^{4} \Theta_k^{w_k / \sum_{r=1}^{4} w_r} \tag{4}$$

where $w_k = (\frac{1}{2})^{4-k}$. The weighting scheme assigned larger weights to higher magnifications as their inputs included more details. The output maps were computed in such a way that the pixels below the threshold were set to the minimum value of the pixels above the threshold as

$$\Theta_k(x,y) = \begin{cases} FCN_k\big(\Phi_k(x,y)\big) & \text{if } (x,y) \in \Omega_k, \\ \min\limits_{(x',y') \in \Omega_k} FCN_k\big(\Phi_k(x',y')\big) & \text{otherwise} \end{cases} \tag{5}$$

where Ω_1 was the set of all pixels in the input image ($\Omega_1 = \{(x,y) \in \Phi\}$), and $\Omega_k = \{(x,y) \in |\Theta_{k-1}|_\tau\}$ for $k > 1$ were the sets of pixels above the corresponding thresholds. $|\Theta_k|_\tau$ denotes thresholding Θ_k adaptively such that the lowest τ percentage of the values of Θ_k were removed from the set of pixels to be processed in the subsequent stages. This ensured that the saliency information obtained by earlier FCNs were not lost while preserving the order of pixel values (i.e., the pixels below the threshold could not have higher values than those above it in the geometric mean). Tuning the parameter τ is discussed in Section 5. Note that, all Θ_k maps were scaled to the same resolution, and the geometric mean in (4) was computed pixel-wise.

4.2. ROI classification

In this section, we describe the methodology for both patch-level and slide-level classification of WSIs into five diagnostic categories (NP, P, ADH, DCIS, IDC) using a convolutional neural network (CNN).

4.2.1. Data set preparation

A single WSI often contains multiple areas with different levels of diagnostic importance, and a small area can lead to the pathologist's finalization of the diagnosis. Our data set contained an example ROI that was marked for each WSI as a representative for the most severe diagnosis that was observed during the three experienced pathologists' consensus meetings. We used these consensus ROIs as the training examples for our deep network for classification, and sampled 100×100 pixel patches with 50 pixel strides to form the training data. The $10\times$ magnification was used so that the patches had sufficient context. This combination of patch size and magnification also allowed us to fit a reasonable number of patches in the available GPU memory. The neighboring patches had 50% overlap to achieve translation invariance. The resulting training set consisted of 1,272,455 patches belonging to five categories. Note that, even though the number of samples seems to be large, many of these patches may contain irrelevant content such as empty areas, necrosis, etc., because the consensus ROIs were marked roughly using rectangular boundaries as shown in Fig. 10. We plan to integrate tissue segmentation as a pre-processing stage to perform contextual sampling from epithelial and stromal regions in future work.

4.2.2. Network architecture for classification

Five-class classification was a more challenging task than binary saliency detection; thus, a deeper network and more training data were required. The training set of patches contained approximately ten times as many pixels as the training set used for detection. Furthermore, the design of the network was updated with more layers, filters, and neurons as shown in Fig. 5.

The resulting network accepted $100 \times 100 \times 3$ fixed sized inputs. Input images were normalized by subtracting the overall mean of the three channels. The network consisted of six convolutional layers with 3×3 filters, followed by three fully convolutional layers and a final softmax layer. Except the last layer, all layers were followed by a ReLU nonlinearity. The convolutional layers contained 64, 64, 64, 128, 128, 256 filters in respective order, and the first, second, fourth, and sixth layers were followed by a 2×2 max pooling operation with a stride of 2. The fully connected layers contained 512, 512, 5 neurons, and are followed by a dropout operation with 0.5 probability. The hyper-parameters of the network architecture were tuned on one-fifth of the training set as validation data.

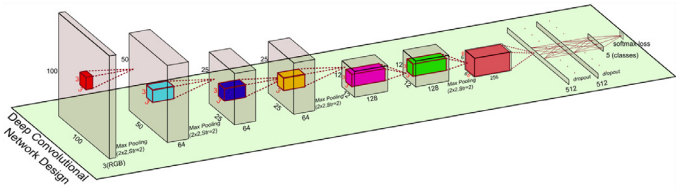


Fig. 5. Illustration of the CNN architecture for ROI classification. The number and size of the filters at each layer are given. All convolutional layers are followed by ReLU nonlinearity.

Our focus was the development of the complete framework, starting from the step that extracts training data from the raw viewing logs of the pathologists to the steps that include the detection of diagnostically relevant ROIs and the ROI-level and slide-level classification of whole slide images. Thus, the network architectures used in this paper were adapted from the network in [29], which has been accepted to be one of the state-of-the-art baselines in many domains. The overall effectiveness can be improved by replacing the networks in Fig. 1 with other suitable architectures from the literature in the future.

4.2.3. Post-processing for slide-level classification

The network provided class probabilities for fixed sized input patches. In order to obtain probability maps for the whole slides, we needed to either classify patches extracted by sliding windows or fully convolutionalize the network. We chose to implement the latter as it enabled more efficient WSI classification that, in fact, implicitly implemented sliding windows with a step size of 16. Therefore, each pixel in the probability maps corresponded to a 16×16 pixel patch in the input space.

The probability maps produced by the above strategy were further downsampled by a factor of seven by bilinear interpolation in order to smooth out the estimates and remove the noise caused by small isolated details. The downsampled maps were then used to determine the final classification such that every pixel voted for the class that had the greatest probability for that pixel. Finally, the class with the majority of the votes was selected as the final diagnosis for the corresponding WSI.

An alternative approach is to learn a slide-level decision fusion model. This has been motivated in the literature [31] for cases in which individual patches may not be discriminative and their predictions can be biased, whereas the learned fusion may model their joint appearance and correct the bias of patch-level decisions. We implemented the method in [31] where a class histogram was generated by summing up all of the class probabilities assigned to all pixels by the patch-level classifier, and a multi-class SVM was trained by using these histograms to produce slide-level predictions.

5. Experiments

In this section, we present the experiments for the detection and classification tasks as well as the visualization of the trained

networks. The training samples for both tasks were further divided into 80% training and 20% validation sets for estimating the hyper-parameters and to avoid overfitting. The implementations were derived from the MatConvNet library [32] with a number of significant modifications, and ran on a system with an NVIDIA GeForce GTX-970 GPU, Intel Xeon ES-2630 2.60 GHz CPU, and 64GB RAM.

5.1. ROI detection

We trained the four FCNs for 50 epochs using the training set. For each FCN, the stochastic gradient descent algorithm was run to optimize a total of 168,290 network parameters on mini batches of 25 images with 0.0001 learning rate, 0.0005 weight decay, and 0.9 momentum. These hyper-parameters were empirically set on a subset of the validation data.

5.1.1. Reference data

Detection of diagnostically relevant ROIs in WSI has not been a well-studied task in the literature, and there is no publicly available data set that is suitable for the evaluation of this task. Therefore, we used the viewport tracking data to generate the annotations for evaluation.

This procedure followed the same approach described in [16,17]. Saliency of the viewports were evaluated by using the following set of rules:

- The pathologist zoomed into a region from the previous window and zoomed out right after. This event was named a zoom peak, and was a local maximum in the zoom level.
- The pathologist slowly slid the viewports while maintaining the same zoom level. This event was named a slow panning, and was represented by the union of the consecutive group of viewports with small displacement.
- The pathologist viewed the same region for more than 2 seconds. This event was named a fixation.

More details can be found in [16,17]. These rules were applied to all viewport logs from the three experienced pathologists, and the union of all windows that satisfied at least one of these rules was computed to create a binary saliency mask for each WSI in the test set. Morphological operations were also used to remove the outer white regions that corresponded to the slide background outside the tissue section because the rectangular viewports often contained such regions. Examples for the saliency masks are shown in Fig. 7. The training and validation labels described in Section 4.1.1 and the test labels described in this section all came from different cases belonging to different patients.

5.1.2. Evaluation criteria

The output of the detection pipeline for each test WSI contains pixel-wise probability estimates in the range [0,1]. These estimates were compared to the reference binary saliency mask for computing pixel-based receiver operating characteristic (ROC) curves by averaging all results from the 60 test cases.

The resulting performance was compared with two alternative approaches. The first one was the classification framework proposed in [17]. The approach in [17] can be considered as a state-of-the-art method that used a bag-of-words model with color and texture features of image patches where a logistic regression classifier trained on the binary saliency masks extracted from the viewport logs of the training slides was used to produce the detection scores.

The second comparison used the U-Net architecture proposed for biomedical image segmentation [33]. The U-Net network consists of 23 convolutional layers where a contracting path is followed by an expansive path. The contracting path uses the typical architecture of a convolutional network, whereas each step in the expansive path performs convolutions on the concatenation of the upsampled version of the previous step in the expansive path and the corresponding feature map from the contracting path. The same training data in four different sets of magnifications were used to train four separate networks that were combined with the same weighting scheme proposed in Section 4.1.3.

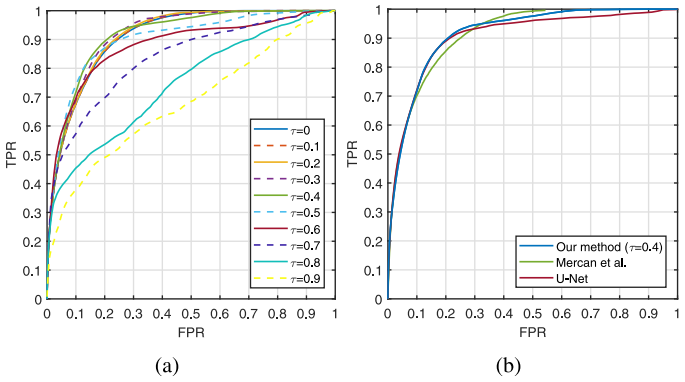


Fig. 6. ROC curves for the proposed saliency detection pipeline with different τ values (a) and comparisons with the method of Mercan et al. in [17] and the U-Net architecture in [33] (b).

5.1.3. Results and discussion

Fig. 6(a) shows the ROC curves for different τ values that were used to eliminate a certain percentage of the pixels for further processing in subsequent stages in the pipeline. The true positive rate (TPR) was considered to represent the effectiveness of the method in identifying all diagnostically relevant ROIs, and the false positive rate (FPR) was considered a suitable metric to evaluate the efficiency of the method to reduce the area to be processed in the following steps as the salient regions usually occupied a relatively small part of a WSI. According to Fig. 6(a), while monotonic improvements on both effectiveness and efficiency were observed until $\tau = 0.4$, further increase in τ corresponded to a decrease in accuracy. Therefore, there is an application dependent trade-off as higher τ values continue to yield more efficiency.

Comparative results are presented in Fig. 6(b). The proposed method attained the best area under the curve (AUC) value for $\tau = 0.4$ as 0.9153, whereas [17] obtained 0.9124 and the network in [33] obtained 0.9043. We also saw that when FPR = 0.2, TPR of [17] and [33] were 0.8552 and 0.8902, respectively, while our method achieved 0.8947. Similarly, in the high TPR region above 0.8, our method obtained smaller FPR values compared to [17] and [33]. Only after a TPR of 0.98, [17] achieved higher TPR at the same FPR. Overall, our method achieved better effectiveness than both [17] and [33], even though [17] used the same set of rules listed in Section 5.1.1 for generating both training and test data whereas our method used a different training set. Furthermore, our method was significantly more efficient than both [17] (with a factor of 74 times for $\tau = 0.4$) and [33] (with a factor of 24 times at the same threshold setting) by operating on lower resolutions and processing only a small portion of the images at utmost $5\times$ magnification in the proposed pipeline, whereas [17] processed entire slides using sliding windows at full $40\times$ magnification and [33] used a much larger network architecture.

The fully convolutional network architecture used in this paper efficiently learned to make dense predictions for per-pixel tasks as the output was aggregated from local computations. Explicit connections from early activations to later layers as in the U-Net architecture have the potential of capturing more detailed location information in the final predictions. However, the resulting networks often need a trade-off for increased complexity in larger

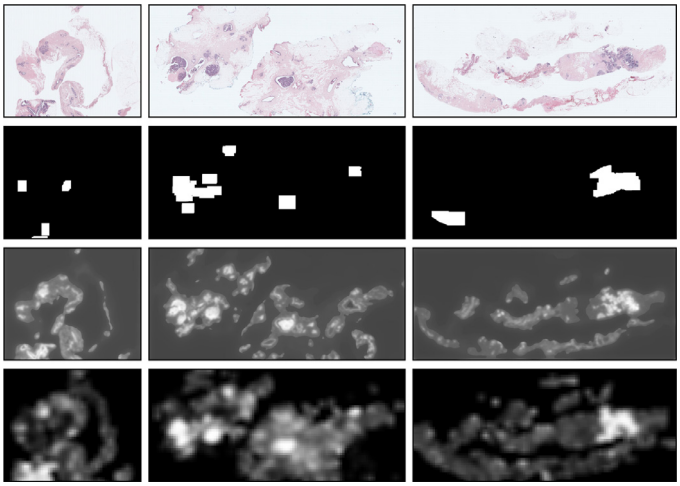


Fig. 7. Example saliency detection results for three WSIs. From top to bottom: RGB WSI, the reference saliency mask, output of the proposed approach for $\tau = 0.4$, output of [17]. The image sizes, from left to right, are $77{,}440 \times 68{,}608$, $128{,}576 \times 65{,}936$, and $132{,}256 \times 55{,}984$, pixels, respectively, at $40\times$ magnification.

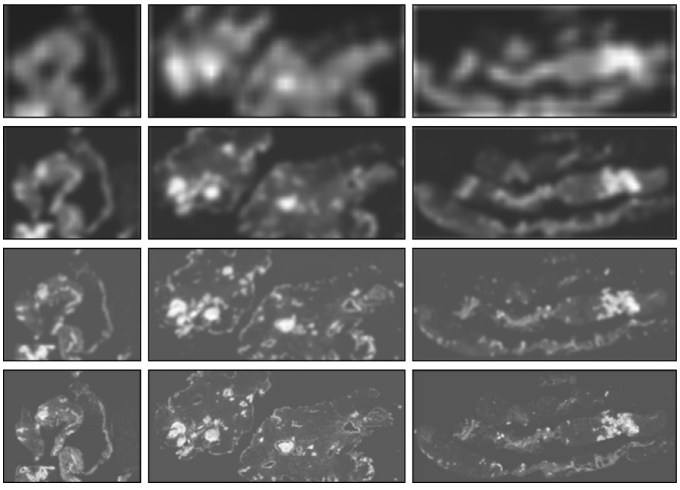


Fig. 8. Details of the individual stages in the saliency detection pipeline for the images shown in Fig. 7. From top to bottom: outputs of the four FCNs, Θ_1, Θ_2, Θ_3, Θ_4, respectively, for $\tau = 0$.

scale problems, such as WSI classification in this paper, via subsampling to keep the filters small and the computational requirements reasonable [27].

Figs. 7–9 present example detection results. Both the full WSI output and the zoomed results showed that the proposed method produced detailed and more precise localization of the relevant regions whereas [17] produced more blurry results because of the windowing effects.

5.2. ROI classification

The CNN used for classification was trained for 50 epochs to optimize a total of 5,576,581 network parameters on mini batches of 256 patches with 0.01 learning rate, 0.0005 weight decay, and 0.9 momentum. These hyper-parameters were empirically set on a subset of the validation data. In order to evaluate the effectiveness of the trained network, we performed experiments for two tasks: classification of 100×100 pixel patches and classification of individual WSIs.

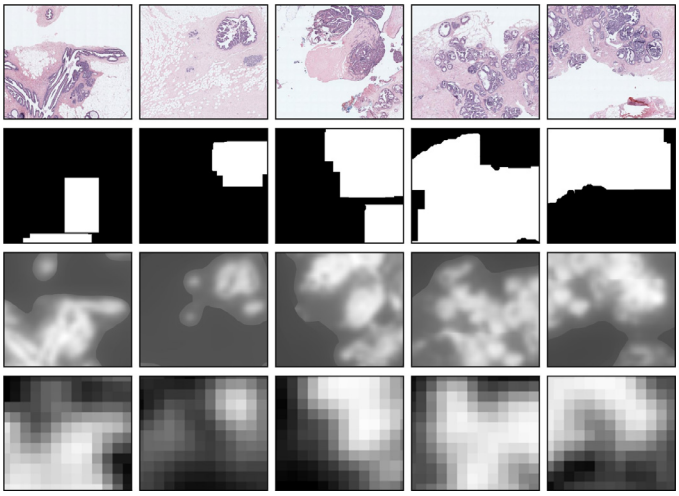
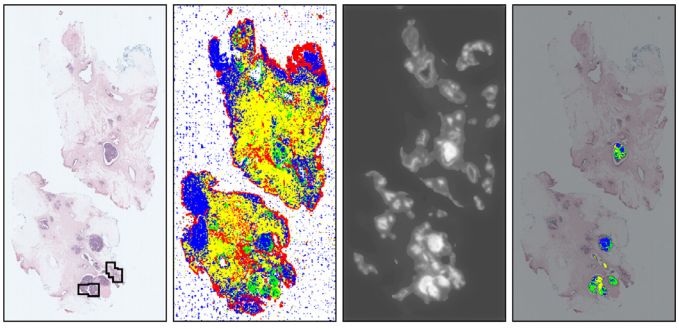
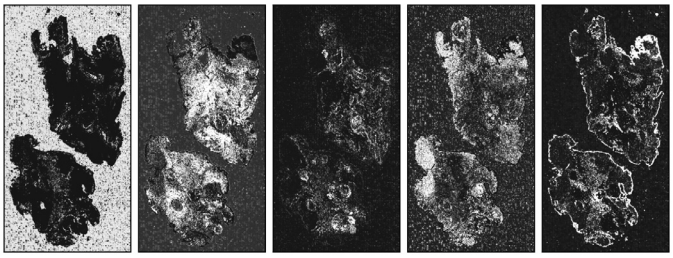


Fig. 9. Zoomed examples from Fig. 7. From top to bottom: RGB image, the reference saliency mask, output of the proposed approach for $\tau = 0.4$, output of [17]. The roughness of the saliency masks used for training and testing can be seen. The proposed method provides more detailed pixel-wise predictions.



(a) Original image (b) Classification (c) Saliency map (d) Fusion



(e) P(NP) (f) P(P) (g) P(ADH) (h) P(DCIS) (i) P(IDC)

Fig. 10. Example classification result for a WSI with a consensus label of ADH. (a) Original image ($65{,}936 \times 128{,}576$ pixels) with consensus ROIs marked with black lines. (b) Patch-level classification for five classes: NP (white), P (yellow), ADH (green), DCIS (blue), IDC (red). (c) Saliency detection (brighter values indicate higher probability). (d) Pixels (in (b)) whose labels were used in the majority voting for slide-level diagnosis after thresholding the saliency map. (e-i) Pixel-wise likelihood maps for five classes. This sample was correctly classified as ADH using the majority voting of the labels shown as overlay in (d). (Best viewed in color with zoom.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2.1. Reference data

The consensus labels assigned by the three experienced pathologists were used as the slide-level reference data. We also used the individual diagnoses provided by the 45 other pathologists on the 60 test cases for comparison. These diagnoses were originally collected for evaluation of the differences between glass slides and digital slides. Therefore, 23 pathologists labeled the same cases by looking at the glass slides, and 22 evaluated the digital slides in WSI format.

Table 2
Confusion matrix for patch-level classification.

		Predicted					TPR &
		NP	**P**	**ADH**	**DCIS**	**IDC**	Recall
True	**NP**	**2477**	945	725	2027	4227	0.2382
	P	503	**7246**	3364	7823	3275	0.3262
	ADH	4092	9249	**5727**	10,572	4511	0.1677
	DCIS	5003	23,074	7068	**47,412**	9550	0.5147
	IDC	661	9145	509	21,491	**18,978**	0.3737
	FPR	0.0515	0.2263	0.0665	0.3566	0.1357	
	Precision	0.1945	0.1459	0.3293	0.5308	0.4681	

Table 3
Confusion matrix for slide-level classification.

		Predicted					TPR &
		NP	**P**	**ADH**	**DCIS**	**IDC**	Recall
True	**NP**	**0**	2	0	1	2	0
	P	0	**9**	2	2	0	0.6923
	ADH	0	4	**4**	8	0	0.2500
	DCIS	0	2	0	**17**	2	0.8095
	IDC	0	0	0	2	**3**	0.6000
	FPR	0.0000	0.1702	0.0455	0.3333	0.0727	
	Precision	–	0.5294	0.6667	0.5667	0.4286	

For the patch classification task, 209,654 patches with 100×100 pixels were sampled from the consensus ROIs of the test cases. Each patch was labeled with the consensus label of the corresponding WSI. However, since the consensus ROIs were roughly drawn as rectangular shapes, some of these patches may contain irrelevant content as in the case of training data generation. The training and validation data described in Section 4.2.1 and the test data described in this section all came from different cases belonging to different patients.

5.2.2. Results and discussion

Patch classification. The accuracy of the CNN for classification of the test patches into five categories was 39.04%. The resulting confusion matrix is shown in Table 2. The errors seemed mostly as underestimations of diagnostic classes as the lower triangle of the confusion matrix added up to 63.21% of the wrong classifications. However, visual inspection of the patches showed that some of them were actually not errors because the whole consensus ROIs were labeled with the same diagnosis without a precise delineation of the ductal regions, and not all patches sampled from these ROIs contained the same level of structural cues that represented the given label. For example, a patch that was sampled from an ROI labeled as ADH could easily contain usual hyperplasia or even stromal regions. Compared to the binary classification tasks of invasive cancer, mitosis, metastasis, etc., detection that have been widely studied in the literature, the labeling of ductal proliferations and hyperplastic changes was a more difficult problem with a higher uncertainty. The fusion of ROI detection and patch classification will recover some of these errors in the next section.

WSI classification. The classification of a WSI by using the fully convolutionalized CNN produced probability maps containing the five-class likelihoods as well as a label map indicating the winning class for each pixel. The class with the highest overall frequency in the whole image (i.e., the majority voting approach described in Section 4.2.3) can be used as the slide-level diagnosis. For robustness to the uncertainty in the output of the patch-based classifier due to the roughness of the consensus ROIs and the corresponding training samples, we also used the saliency map for each WSI, and applied an adaptive threshold so that only the top 15% of the salient pixels remained, where the final slide-level class prediction was obtained for the WSI by using majority voting only among the class labels of the pixels that achieved the highest (top 15%) probability of being salient. The threshold percentage was selected by using the validation data. Fig. 10 shows an example classification. More examples can be found in [34].

We also used the learned decision fusion model by training two separate multi-class SVM classifiers by using the class histograms of the pixels (i.e., the learned fusion approach described in Section 4.2.3) without and with selection by the saliency detection pipeline. The same thresholding protocol was used during selection.

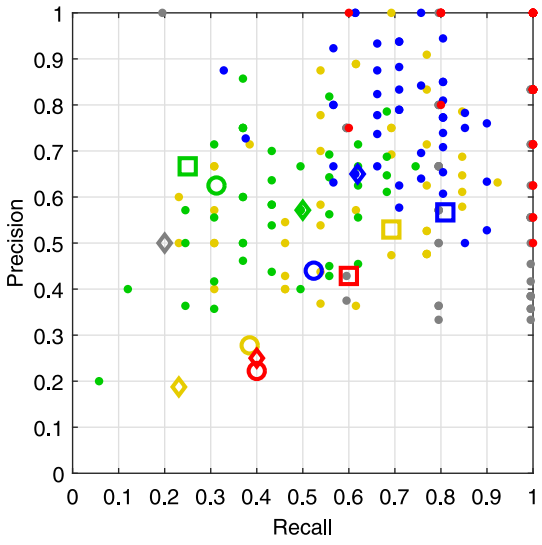


Fig. 11. Class-specific precision versus recall for the proposed method (square), the SVM baseline (diamond), the RF baseline (circle), and the 45 pathologists (dot). Colors represent: NP (gray), P (yellow), ADH (green), DCIS (blue), IDC (red). The variability in the pathologists' predictions, with a very wide range of concordance rates compared with the reference diagnoses particularly for the P, ADH, and DCIS categories, is consistent with the medical literature where inter-rater agreement has always been a known challenge [4,26]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Quantitative evaluation was performed by comparing the final slide-level predictions with the consensus labels and the predictions of the 45 pathologists. We also trained multi-class SVM and RF classifiers with state-of-the-art hand-crafted features including 192-bin Lab histograms (64 bins per channel), 128-bin local binary pattern (LBP) histograms (64 bins for each of the H and E channels estimated via color deconvolution), and 50 nuclear architectural features (as in [12]) with different feature combinations. Both the SVM and the RF classifiers are popular non-deep learning methods for histopathological image classification, and were used as representative baselines in our experiments. The features were computed within $3{,}600 \times 3{,}600$ pixel windows at the highest $40\times$ magnification where the window size was decided based on the observations in [17]. Sliding windows that were inside the consensus ROIs of the training set were used to build the SVM with a linear kernel and the RF classifier where the cost parameter for the SVM and the number of trees and tree depths for the RF were obtained by using cross-validation. The resulting classifiers were then used to label the sliding windows of the test WSIs, and the resulting likelihood maps were combined with the same saliency detection outputs as in our method to obtain the slide-level predictions.

Table 3 shows the confusion matrix for our method. Fig. 11 shows the class-specific precision and recall values for our method, the best performing baselines when all features were combined

Table 4
Classification accuracies for the pathologists, the proposed deep learning-based method, and the state-of-the-art hand-crafted feature representations and classifiers.

	Accuracy (%)
Average and standard deviation of 45 pathologists	65.44 ± 7.07
Proposed method with majority voting without saliency	23.33
Proposed method with majority voting with saliency	55.00
Proposed method with learned fusion without saliency	38.33
Proposed method with learned fusion with saliency	55.00
Lab+LBP+Arch. features with SVM without saliency	28.33
Lab+LBP+Arch. features with SVM with saliency	45.00
Lab+LBP+Arch. features with RF without saliency	31.67
Lab+LBP+Arch. features with RF with saliency	38.33

(370 features), and the 45 pathologists' predictions. Table 4 summarizes all results.

We observed that the learned fusion approach improved the results against majority voting (from 23.33% to 38.33%) when the saliency map was not used. However, considering only the set of pixels with the highest probability of being salient in the slide-level prediction resulted in the same accuracy (55%) when both the majority voting and the learned fusion approaches were used. The 55% classification accuracy achieved by the proposed framework was also 10% higher than the best performing hand-crafted feature and classifier combination as seen in Table 4. This shows that our saliency detection pipeline was very selective and discriminative where majority voting among the most salient pixels was sufficient for the slide-level diagnosis, with the additional benefit of incrementally eliminating most of the image regions in lower magnifications and processing only small portions of the images in higher magnifications. In particular, the average running times for a single whole slide test image (with an average size of $94{,}525 \times 64{,}330$ pixels) could be summarized as follows: saliency detection at $0.625\times$, $1.25\times$, $2.5\times$, and $5\times$ magnifications took 0.50, 1.21, 2.90, and 6.97 s, respectively, for a total of 11.58 s for the whole pipeline when the threshold for eliminating diagnostically irrelevant regions was set to $\tau = 0.4$, and classification of the patches that contained the top 15% of the salient pixels took 55.09 seconds using our Matlab-based implementation on a single core of the CPU.

The overall slide-level classification accuracy of 55% was also comparable to the performances of the 45 pathologists that practice breast pathology in their daily routines. As seen from Fig. 11, there were very mixed performances from the pathologists for the P, ADH, and DCIS classes. In the clinical setting, the pathologists usually agree in their diagnoses for the NP and IDC cases because these are at two extremes of the continuum of histologic features. Given the smaller amount of data used to train the networks, our performance for the NP and IDC classes were lower than the typical pathologist's performance. As there is little clinical difference in how the patients with non-proliferative (NP) and proliferative (P) benign biopsies are managed, we plan to merge the NP and P cases as a single class named benign without atypia in future work. However, when the other more difficult intermediate diagnostic categories with different clinical significance as risk factors for future cancer and with different subsequent surveillance and preventive treatment options were concerned, the proposed method performed better, in terms of recall, than 30 pathologists for P, 5 pathologists for ADH, and 39 pathologists for DCIS. In terms of precision, our method was better than 17 pathologists for P, 34 pathologists for ADH, and 2 pathologists for DCIS.

We also applied McNemar's test [35] to compare the proposed method with the pathologists. Given the predictions of our method and the individual pathologists' for all 60 test cases, 45 tests were carried out at 5% significance level, and in 32 of these tests the null hypothesis could not be rejected, i.e., their performances were not statistically significantly different than ours. Furthermore, we performed a z-test also at 5% significance level, and again we could not reject the null hypothesis, i.e., our scores belonged to the same normal distribution estimated from the performances of the 45 pathologists.

The overall results indicated that the fusion of saliency detection for localization of diagnostically relevant regions and the classification of these regions into five diagnostic classes using deep networks provided a promising solution. The alternative approach of [23] that was tested on the same data set in four-class classification (after merging non-proliferative and proliferative changes as a single category named benign) achieved an accuracy of 56% when the structure feature computed using histograms of eight tissue types within layers of superpixels both inside and around ductal objects was used. Though not directly comparable with our five-class slide-level performance as that accuracy was computed only within the consensus ROIs of the test slides, it provided additional confirmation of the difficulty of the multi-class classification problem involving the full range of histologic categories. Another important finding of that work was that, the structure feature that explicitly incorporated the highly specialized domain knowledge into the classification model was particularly powerful in discriminating ADH cases from DCIS that have not been studied in the published literature. Given the years of training and experience that the pathologists use to diagnose the biopsies, and the importance of objective and repeatable measures for interpreting the tissue samples under the multi-class classification scenario where different classes carry significantly different clinical consequences, our comparable results on this challenging data set showed the promise of deep learning where future work with larger and more precisely labeled data sets and additional computational resources will eventually be practical in a clinical setting.

5.3. Visualization

CNNs are often criticized as black box models. Recent work on the visualization of the inner details of CNNs can also be useful in understanding the representations learned from pathology data. We used the occlusion [36] and deconvolution [37] methods, with implementations from the FeatureVis library [38], to visualize the CNN learned for multi-class classification.

The occlusion method added small-sized random occluders at different locations in a patch and compared the resulting activation after each occlusion with the original one. Fig. 12 shows the visualization results as maps of the importance of different details in example images that affected the classification of particular classes positively or negatively. For example, the first three rows show examples of ductal regions with few layers of epithelial cells around lumens. The fifth and sixth rows show examples of atypical proliferations. The seventh and eighth rows show examples of ducts filled with epithelial cells. The tenth and eleventh rows show examples of intertwined groups of cells with no apparent ductal structure. The ninth and twelfth rows contain examples that were listed as misclassifications that might actually be correct decisions but were counted as errors because of the imprecise delineation of the consensus ROIs and the difficulty of sampling from these large rectangular windows. Finally, the fourth row shows a clear example of the need of the saliency detection step because the almost empty patch confused the CNN and led to activations for multiple classes as similar regions were included in the sample sets for all classes. Note that, it was ignored in the final fused decision because the fully convolutional multi-scale saliency detection pipeline eliminated such areas.

The deconvolution method built reconstructions by projecting the activations back to the input space so that parts of the input

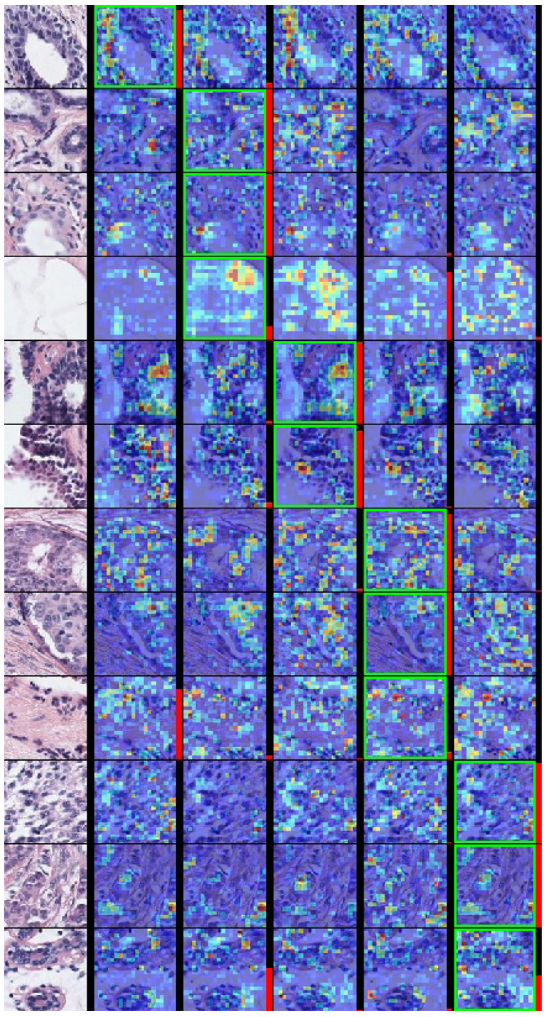


Fig. 12. Visualization of the learned representations using the occlusion method. Each row represents a separate example patch. From left to right: 100×100 pixel input patch, importance of local details overlaid on the input image for individual diagnostic classes NP, P, ADH, DCIS, and IDC. Warmer colors indicate higher impact of that region (either positively or negatively) for the classification of that class. Reference diagnoses are marked by green boxes. The predictions of our method are shown by red bars whose heights indicate the likelihood. (Best viewed in color with zoom.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

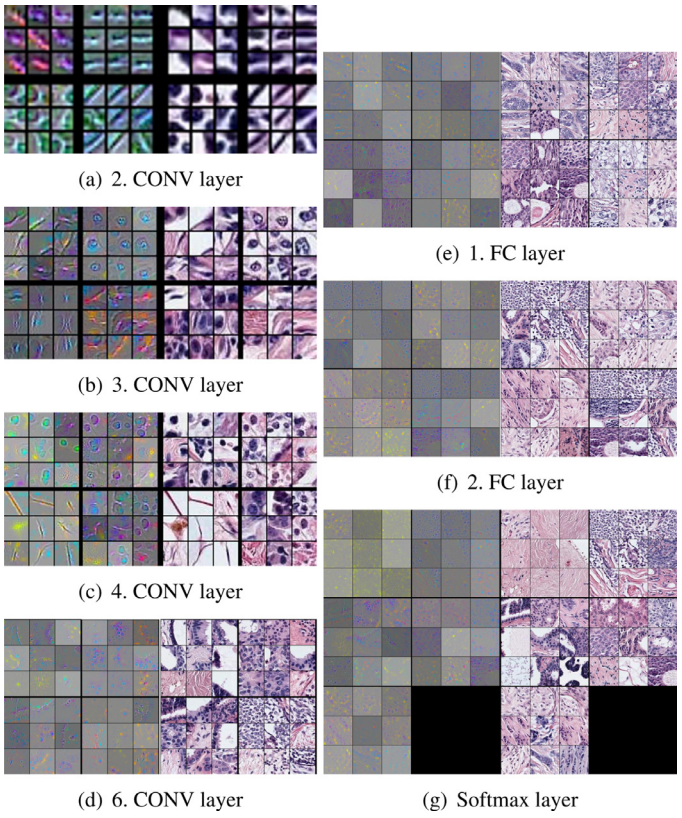


Fig. 13. Visualization of the network layers by using the deconvolution method. For each convolutional (CONV) and fully connected (FC) layer, the top-9 activations (as 3×3 groups) for four example neurons (left) and their corresponding original input patches (right) are shown. The last softmax layer consists of five neurons corresponding to five classes; from left to right and top to bottom: NP, P, ADH, DCIS, IDC. (Best viewed in color with zoom.).

image that most strongly activated particular neurons were found. Fig. 13 illustrates the top-9 responsive patches for example neurons from different layers and the visualization of the contributions of their pixels. The examples showed how the lower layers captured the fundamental features such as edges and blobs, and the higher layers developed more abstract features based on patterns representing particular arrangements of nuclei and other ductal structures. Future work includes more detailed evaluation of these visualizations in a clinical perspective together with the pathologists.

6. Conclusions

We presented a deep learning-based computer aided diagnosis system for breast histopathology. The proposed framework covered the whole workflow from an input whole slide image to its categorization into five diagnostic classes. The first step was saliency detection by using a pipeline of four sequential fully convolutional networks for multi-scale processing of the whole slide at different magnifications for localization of diagnostically relevant ROIs. Both the learning and the inference procedures imitated the way pathologists analyze the biopsies by using the pathologists' recorded actions while they were interpreting the slides. The second step was a patch-based multi-class convolutional network for diagnosis that was learned by using representative ROIs resulting from the consensus meetings of three experienced pathologists. The final step was the fusion of the saliency detector and the fully-convolutionalized classifier network for pixel-wise labeling of the whole slide, and a majority voting process to obtain the final slide-level diagnosis. The deep networks used for detection and classification performed better than competing methods that used hand-crafted features and statistical classifiers. The classification network also obtained comparable results with respect to the diagnoses provided by 45 other pathologists on the same data set. We also presented example visualizations of the learned representations for better understanding of the features that were determined to be discriminative for breast cancer diagnosis. Given the novelty of the five-class classification problem that is important for clinical applicability of computer aided diagnosis, the proposed solutions and the presented results by using a challenging whole slide image data set show the potential of deep learning for whole slide breast histopathology where future work with larger data sets with more detailed training labels have the promise to result in systems that are useful to pathologists in clinical applications.

Conflict of interest

We confirm that there are no known conflicts of interest associated with this work.

Acknowledgments

B. Gecer and S. Aksoy were supported in part by the Scientific and Technological Research Council of Turkey (grant 113E602) and in part by the GEBIP Award from the Turkish Academy of Sciences. E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore were supported in part by the National Cancer Institute of the National Institutes of Health (awards R01-CA172343, R01-140560, and KO5-CA104699). The content is solely the responsibility of the authors and does not necessarily represent the views of the National Cancer Institute or the National Institutes of Health.

References

[1] M. Veta, J.P.W. Pluim, P.J. van Diest, M.A. Viergever, Breast cancer histopathology image analysis: a review, IEEE Trans. Biomed. Eng. 61 (5) (2014) 1400–1411.
[2] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, B. Yener, Histopathological image analysis: a review, IEEE Rev. Biomed. Eng. 2 (2009) 147–171.
[3] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases, J. Pathol. Inform. 7 (29) (2016).
[4] J.G. Elmore, G.M. Longton, P.A. Carney, B.M. Geller, T. Onega, A.N.A. Tosteson, H.D. Nelson, M.S. Pepe, K.H. Allison, S.J. Schnitt, F.P. O'Malley, D.L. Weaver, Diagnostic concordance among pathologists interpreting breast biopsy specimens, J. Am. Med. Assoc. 313 (11) (2015) 1122–1132.
[5] K.H. Allison, M.H. Rendi, S. Peacock, T. Morgan, J.G. Elmore, D.L. Weaver, Histological features associated with diagnostic agreement in atypical ductal hyperplasia of the breast: illustrative cases from the B-Path study, Histopathology 69 (2016) 1028–1046.
[6] T.T. Brunye, P.A. Carney, K.H. Allison, L.G. Shapiro, D.L. Weaver, J.G. Elmore, Eye movements as an index of pathologist visual expertise: a pilot study, PLoS ONE 9 (8) (2014).
[7] M.M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, M.N. Gurcan, Computerized classification of intraductal breast lesions using histopathological images, IEEE Trans. Biomed. Eng. 58 (7) (2011) 1977–1984.
[8] F. Dong, H. Irshad, E.-Y. Oh, M.F. Lerwill, E.F. Brachtel, N.C. Jones, N.W. Knoblauch, L. Montaser-Kouhsari, N.B. Johnson, L.K.F. Rao, B. Faulkner-Jones, D.C. Wilbur, S.J. Schnitt, A.H. Beck, Computational pathology to discriminate benign from malignant intraductal proliferations of the breast, PLOS ONE 9 (12) (2014).
[9] J. Kong, O. Sertel, H. Shimada, K.L. Boyer, J.H. Saltz, M.N. Gurcan, Computer-aided evaluation of neuroblastoma on whole-slide histology images: classifying grade of neuroblastic differentiation, Pattern Recognit. 42 (2009) 1080–1092.
[10] O. Sertel, J. Kong, H. Shimada, U.V. Catalyurek, J.H. Saltz, M.N. Gurcan, Computer-aided prognosis of neuroblastoma on whole-slide images: classification of stromal development, Pattern Recognit. 42 (2009) 1093–1103.
[11] S. Doyle, M. Feldman, J. Tomaszewski, A. Madabhushi, A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies, IEEE Trans. Biomed. Eng. 59 (5) (2012) 1205–1218.
[12] A. Basavanhally, S. Ganesan, M. Feldman, N. Shih, C. Mies, J. Tomaszewski, A. Madabhushi, Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides, IEEE Trans. Biomed. Eng. 60 (8) (2013) 2089–2099.
[13] M. Balazsi, P. Blanco, P. Zoroquiain, M.D. Levine, M.N. BurnierJr., Invasive ductal breast carcinoma detector that is robust to image magnification in whole digital slides, J. Med. Imaging 3 (2) (2016) 1–9.
[14] C. Mercan, S. Aksoy, E. Mercan, L.G. Shapiro, D.L. Weaver, J.G. Elmore, Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images, IEEE Trans. Med. Imaging 37 (1) (2018) 316–325.
[15] C. Bahlmann, A. Patel, J. Johnson, J. Ni, A. Chekkoury, P. Khurd, A. Kamen, L. Grady, E. Krupinski, A. Graham, R. Weinstein, Automated detection of diagnostically relevant regions in H&E stained digital pathology slides, SPIE Medical Imaging Symposium, 2012.
[16] E. Mercan, S. Aksoy, L.G. Shapiro, D.L. Weaver, T. Brunye, J.G. Elmore, Localization of diagnostically relevant regions of interest in whole slide images, in: International Conference on Pattern Recognition, 2014, pp. 1179–1184.
[17] E. Mercan, S. Aksoy, L.G. Shapiro, D.L. Weaver, T.T. Brunye, J.G. Elmore, Localization of diagnostically relevant regions of interest in whole slide images: a comparative study, J. Digit. Imaging 29 (4) (2016) 496–506.
[18] B.E. Bejnordi, M. Balkenhol, G. Litjens, R. Holland, P. Bult, N. Karssemeijer, J.A.W.M. van der Laak, Automated detection of DCIS in whole-slide H&E stained breast histopathology images, IEEE Trans. Med. Imaging 35 (9) (2016) 2141–2150.
[19] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, F. Gonzalez, A. Madabhushi, Accurate and reproducible invasive breast cancer detection in whole slide images: a deep learning approach for quantifying tumor extent, Sci. Rep. 7 (46450) (2017).
[20] G. Litjens, C.I. Sanchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-van de Kaa, P. Bult, B. van Ginneken, J. van der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci. Rep. 6 (26286) (2016).
[21] X. Shi, F. Xing, K. Xu, Y. Xie, H. Su, L. Yang, Supervised graph hashing for histopathology image retrieval and classification, Med. Image Anal. 42 (2017) 117–128.
[22] Y. Liu, K. Gadepalli, M. Norouzi, G.E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P.Q. Nelson, G.S. Corrado, J.D. Hipp, L. Peng, M.C. Stumpe, Detecting cancer metastases on gigapixel pathology images, 2017, arXiv:1703.02442.
[23] E. Mercan, Digital Pathology: Diagnostic Errors, Viewing Behavior and Image Characteristics, University of Washington, Seattle, Washington, 2017 Ph.D. thesis.
[24] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. Elmore, L. Shapiro, Learning to segment breast biopsy whole slide images, in: IEEE Winter Conference on Applications of Computer Vision, 2018.
[25] N.V. Oster, P.A. Carney, K.H. Allison, D.L. Weaver, L.M. Reisch, G. Longton, T. Onega, M. Pepe, B.M. Geller, H.D. Nelson, T.R. Ross, A.N.A. Tosteson, J.G. Elmore, Development of a diagnostic test set to assess agreement in breast pathology: practical application of the guidelines for reporting reliability and agreement studies (GRRAS), BMC Womens Health 13 (3) (2013) 1–8.
[26] J.G. Elmore, G.M. Longton, M.S. Pepe, P.A. Carney, H.D. Nelson, K.H. Allison, B.M. Geller, T. Onega, A.N.A. Tosteson, E. Mercan, L.G. Shapiro, T.T. Brunye, T.R. Morgan, D.L. Weaver, A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis, J. Pathol. Inform. 8 (1) (2017) 1–12.
[27] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.
[28] T.T. Brunye, E. Mercan, D.L. Weaver, J.G. Elmore, Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images, J. Biomed. Inform. 66 (2017) 171–179.
[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015 arXiv:1409.1556.
[30] B. Gecer, G. Azzopardi, N. Petkov, Color-blob-based COSFIRE filters for object recognition, Image Vis. Comput. 57 (2017) 165–174.
[31] L. Hou, D. Samaras, T.M. Kurc, Y. Gao, J.E. Davis, J.H. Saltz, Patch-based convolutional neural network for whole slide tissue image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2424–2433.
[32] A. Vedaldi, K. Lenc, Matconvnet – convolutional neural networks for MATLAB, in: ACM International Conference on Multimedia, 2015. http://www.vlfeat.org/matconvnet/
[33] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.
[34] B. Gecer, Detection and Classification of Breast Cancer in Whole Slide Histopathology Images Using Deep Convolutional Networks, Bilkent University, Ankara, Turkey, 2016 Master's thesis.
[35] T.G. Dietterich, Approximate statistical tests for comparing supervised learning algorithms, Neural Comput. 10 (7) (1998) 1895–1923.
[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene CNNs, in: International Conference on Learning Representations, 2015 arXiv:1412.6856.
[37] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: the all convolutional net, in: International Conference on Learning Representations, 2015 arXiv:1412.6806.
[38] F. Grun, C. Rupprecht, N. Navab, F. Tombari, A taxonomy and library for visualizing learned features in convolutional neural networks, 2016, arXiv:1606.07757.

BARIS GECER received the B.S. degree from Hacettepe University, Ankara, Turkey in 2014, and the M.S. degree from Bilkent University, Ankara in 2016. He is a Ph.D. student of Imperial Computer Vision & Learning Laboratory at the Imperial College, London. His research interests include computer vision, deep learning, face recognition and medical image analysis.

SELIM AKSOY received the Ph.D. degree from the University of Washington, Seattle in 2001. He is currently an Associate Professor at the Department of Computer Engineering, Bilkent University, Ankara, Turkey. His research interests include statistical and structural pattern recognition and computer vision with applications to medical imaging and remote sensing.

EZGI MERCAN, Ph.D., is a researcher at the Craniofacial Center at Seattle Children's Hospital. She earned her PhD (2017) degree in Computer Science and Engineering from the University of Washington in Seattle. Her research is in medical image analysis with a special interest in breast histopathology and craniofacial imaging.

LINDA SHAPIRO received the Ph.D. degree from the University of Iowa in 1974. She is currently Professor of Computer Science and Engineering and of Electrical Engineering at the University of Washington. Her research interests include computer vision, image database systems, artificial intelligence, pattern recognition, and robotics. Dr. Shapiro is a Fellow of the IEEE and a Fellow of the IAPR. She is a past Chair of the IEEE Computer Society Technical Committee on Pattern Analysis and Machine Intelligence and is an Advisory Editor of Pattern Recognition.

DONALD WEAVER, MD, is a Professor of Department of Pathology & Laboratory Medicine, Director of Surgical Pathology Fellowship Program, and Medical Director of the UVM Cancer Center Biobank at the University of Vermont. Dr. Weaver's expertise is breast and cervical pathology.

JOANN ELMORE, MD, MPH, is a Professor of Medicine and Adjunct Professor of Epidemiology at the University of Washington, an Affiliate Investigator with the Fred Hutchinson Cancer Research Center and the Group Health Research Institute. Dr. Elmore enjoys seeing patients as a primary care internist and teaching clinical medicine and epidemiology.