

# Guided Soft Attention Network for Classification of Breast Cancer Histopathology Images

Heechan Yang, Ji-Ye Kim, Hyongsuk Kim<sup>✉</sup>, *Senior Member, IEEE*, and Shyam P. Adhikari<sup>✉</sup>

**Abstract**—An attention guided convolutional neural network (CNN) for the classification of breast cancer histopathology images is proposed. Neural networks are generally applied as black box models and often the network's decisions are difficult to interpret. Making the decision process transparent, and hence reliable is important for a computer-assisted diagnosis (CAD) system. Moreover, it is crucial that the network's decision be based on histopathological features that are in agreement with a human expert. To this end, we propose to use additional region-level supervision for the classification of breast cancer histopathology images using CNN, where the regions of interest (RoI) are localized and used to guide the attention of the classification network simultaneously. The proposed supervised attention mechanism specifically activates neurons in diagnostically relevant regions while suppressing activations in irrelevant and noisy areas. The class activation maps generated by the proposed method correlate well with the expectations of an expert pathologist. Moreover, the proposed method surpasses the state-of-the-art on the BACH microscopy test dataset (part A) with a significant margin.

**Index Terms**—Breast cancer, microscopy image, convolutional neural network, guided attention, pattern recognition and classification.

## I. INTRODUCTION

**B**REAST cancer is one of the most commonly occurring cancer in women affecting more than 10% of women worldwide. Its mortality rate is very high when compared to

other types of cancer. Though different imaging techniques like diagnostic mammograms, magnetic resonance imaging, sonography etc. are used to detect and diagnose breast cancer, histopathological analysis of breast tissue by pathologists is the only way to diagnose breast cancer with confidence. Histopathological analysis is a highly time-consuming specialized task which depends on the skill and experience of the pathologist. The diagnosis is at times subjective and can be affected by several human factors such as fatigue and loss of attention. In addition there is often a lack of consensus among the experts regarding the diagnosis. Considering these facts, there is a pressing need for a computer-assisted diagnosis (CAD) system that can automatically detect and categorize the pathology present in histopathological images with high reproducibility and accuracy.

Researchers have devoted a considerable amount of effort and there is a large body of work on automatic classification of histopathological images [1]–[4]. Most of these works use handcrafted features like color and texture descriptors with classifiers like support vector machine (SVM), random forest (RF) and multilayer perceptron (MLP). However, handcrafted features are not generalizable and often fail to capture the extensive structural diversity found in microscopy and whole slide images (WSI). Recent advances in computer vision and neural networks have demonstrated that automatic feature learning using deep neural networks are more successful than hand-engineered features. Specifically, convolutional neural network (CNN) has been extensively used to produce state-of-the-art results in different computer vision and pattern recognition problems [5]–[10]. CNNs have also been used extensively in breast histopathology [11]–[18] and have produced state-of-the-art results in the breast cancer histology images (BACH) [15], [16], [19] challenge. However, the (excessive) irrelevant noisy areas present in the global image [15] and the local patches [16] affect the decision of neural networks. Moreover, CNNs are generally applied as opaque black box models where the results are difficult to interpret.

Activation maps have been investigated by many researchers for interpreting the decision of neural networks [20]–[26]. In [20]–[22] error back propagation based methods are applied to visualize regions that are helpful for predicting a class. Class activation map (CAM) [23] shows that the average pooling layer can help to generate activation maps representing task relevant regions than fully-connected layers. Grad-CAM [24] extended CAM to explain the model decisions. Training CNNs using global image or random image patches like [15], [16] are typically affected by the irrelevant noisy areas. There

Manuscript received August 29, 2019; revised October 9, 2019; accepted October 13, 2019. Date of publication October 17, 2019; date of current version April 30, 2020. This work was supported in part by the Korea Research Fellowship Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT under Grant NRF-2015H1D3A1062316 and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2019R1A6A1A09031717. (H. Yang and J. Y. Kim contributed equally to this work.) (Corresponding author: Shyam P. Adhikari.)

H. Yang is with the Division of Electronics and Information Engineering, Chonbuk National University, Jeonju 567-54896, South Korea (e-mail: yhc3006@jbnu.ac.kr).

J.-Y. Kim is with the Department of Pathology, Ilsan Paik Hospital, Inje University College of Medicine, Goyang 10380, South Korea, and also with the Department of Pathology, Severance Hospital, Yonsei University College of Medicine, Seoul 03722, South Korea (e-mail: alucion@gmail.com).

H. Kim and S. P. Adhikari are with the Division of Electronics Engineering, Intelligent Robots Research Center (IRRC), Chonbuk National University, Jeonju 567-54896, South Korea (e-mail: hskim@jbnu.ac.kr, all.shyam@gmail.com).

This article has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors.

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

Digital Object Identifier 10.1109/TMI.2019.2948026

may be bias towards a noisy distractor with high correlation and generalization performance may suffer when the test data does not exhibit the same correlation as the training data. A method which can exploit the subtle differences in local structures by highlighting the object of interest and suppressing the irrelevant areas is required to improve the performance of the learning system. Networks with attention mechanism can help to suppress the noisy areas from the final decision making process. Attention mechanism makes the decision process of the neural network more transparent and explainable, thus making neural networks more reliable and trustworthy. Attention in CNNs is implemented as a post-processing step in [20]–[23] in order to understand the decision process of neural network for visual object recognition.

Some researchers have applied unsupervised attention mechanisms in analyzing medical images. Guan *et al.* [27], used a three-branch attention guided convolution neural network (AG-CNN) to learn from disease-specific regions to avoid noisy areas on the task of thorax disease classification on chest X-ray images. A multistage training was utilized with unsupervised extraction of salient regions. However, only a single prominent lesion or region of interest (RoI) was extracted and the method has limitations for applications where multiple disjoint RoIs may be present. Schlemper *et al.* [28], proposed an attention gate (AG) model which learns to focus on target structures in an unsupervised manner, and applied for fetal ultrasound scan plane detection and pancreas segmentation. The AG-CNN also learns to guide the network on target structures of varying shapes and sizes by suppressing irrelevant regions in the input image. However, no marked improvement in performance accuracy of AG-CNN over the vanilla network without attention mechanism is reported. Unsupervised attention mechanisms of [27], [28] are derived from the global images, and the attention map in the first place may be affected by the irrelevant noisy areas resulting in incorrect attention.

End-to-end trainable attention mechanisms to focus the gaze of the network on salient regions for image classification and weakly supervised segmentation have also been studied in the literature [25], [26]. Li *et al.* [26] used two streams of network; classification stream and attention mining stream with parameters shared between them to model the attention mechanism explicitly as part of the training. The attention map was generated from the classification branch using the gradient based Grad-CAM [24]. Motivated by computing more complete and accurate attention maps, the resultant attention map was then used to generate a soft mask. A complement of the mask was then applied to the original input image and used to learn the attention mining branch.

Apart from the unsupervised attention mechanisms of [27], [28], which are guided by the target category labels only, researchers have also explored the use of additional supervision for guiding the attention of CNN [17], [18], [29]. Including additional supervision labels for attention can mitigate the problem of incorrect attention map. However, manual annotation of relevant areas required for supervised attention is labor-intensive and highly expensive as it has to be carried out only by trained pathologists. The region based

supervision of [29] provides an alternative for supervising attention with reduced human labor and cost. Li *et al.* [26] also supported the integration of external supervision in addition to the trainable self-attention mechanism. And, it was shown that additional external supervision always leads to increased performance than using the self-attention mechanism alone.

In this work we propose to include additional region level coarse annotation to guide the attention of CNN for the classification of breast cancer microscopy images. The additional supervision consist of region level labels where only the diagnostically relevant regions are annotated without considering the detailed pixel wise contents inside the region. This annotation can be carried out by novice pathologists with minimal labor under the supervision of expert pathologists, thus significantly reducing the cost and time of annotation. Using a multi-task learning framework, we design a CNN that simultaneously localizes RoIs from the global image, and guides the classification network to make decisions based only on the diagnostically relevant regions. Unlike [29] where the region based attention is implicit, we propose to learn the RoI mask and explicitly use the RoI based soft attention mechanism to guide the focus of the classification network. The soft mask is applied directly to the penultimate activation maps of the classification branch, unlike [26] where it was applied to the input. This explicit attention mechanism is shown to be more effective than the implicit assumption, and the proposed network is shown to produce state-of-the-art accuracy on the BACH microscopy image classification task.

In summary the contributions of this work are as follows:

- 1) We propose a guided soft attention network (GuSA) which explicitly guides the focus of the network on diagnostically relevant regions. We evaluate GuSA on BACH microscopy images, where it surpasses the state-of-the-art by a significant margin.
- 2) We visualize the proposed method applied to breast cancer classification task. GuSA generates activation maps that can be directly used for visual explanation without need for any post-processing. On the BACH microscopy dataset, GuSA activates neurons in diagnostically relevant regions while suppressing in noisy and irrelevant regions, thus making the neural network decision transparent and reliable.

Rest of the paper is organized as follows. The dataset used in this study is described in Section II, followed by the proposed region guided soft attention network in Section III. Experiment and results are presented in Section IV.

Detailed class-wise performance of the proposed system is discussed in Section V. The visualization of activation maps highlighting the diagnostically relevant regions produced by the proposed method is presented in Section VI followed by concluding remarks.

## II. DATASET

A subset of the breast cancer histology images (BACH) challenge [19] was used in this study. The BACH challenge is divided into two parts A and B. Part A is related to

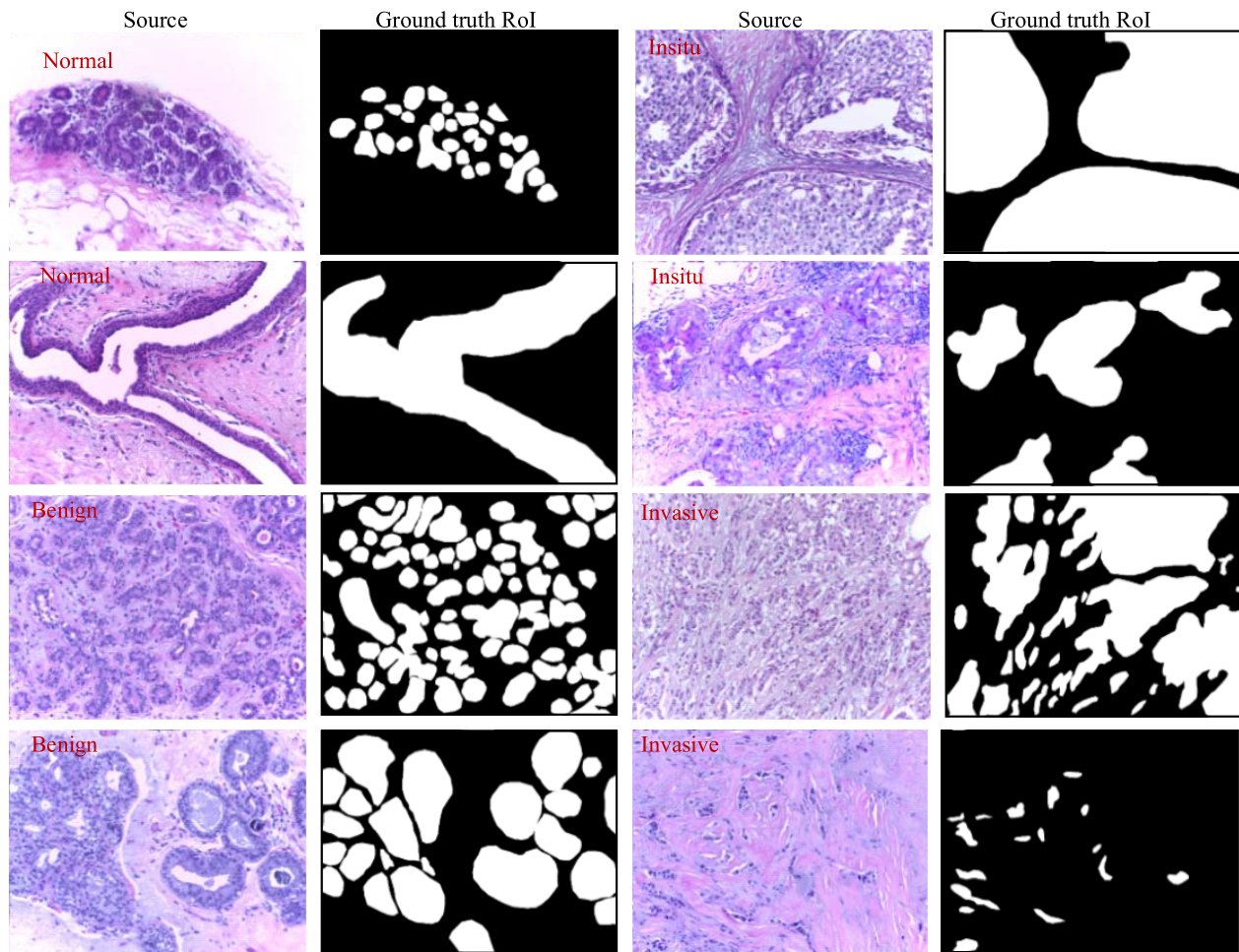


Fig. 1. Train dataset: Breast cancer microscopy images labeled with the four classes; normal, benign, In situ carcinoma, and invasive carcinoma, and the corresponding annotated regions of interest.

the automatic classification of hematoxylin and eosin (H&E) stained breast histology microscopy image in four classes: normal, benign, in-situ carcinoma and invasive carcinoma. Part B of the dataset is related to the pixel-wise segmentation of whole-slide breast histology images into the same four classes. The challenge provides two labeled training datasets; one each for part A and part B. Only the dataset corresponding to part A of the challenge was used in this study. The microscopy dataset consists of 400 labeled training image and 100 test images (labels are hidden) distributed evenly among the four classes. The images were captured using a Leica DM 200) LED microscope and a Leica ICC50 HD camera and all the patients were from Portugal. The microscopy images were annotated image wise by experts from the Institute of Molecular Pathology and Immunology of the University of Porto and from the Institute for Research and Innovation in Health. The microscopy images are RGB images with  $2048 \times 1536$  pixels each and a pixel scale of  $0.42\mu\text{m} \times 0.42\mu\text{m}$ .

In addition to the image wise labels provided by the challenge, the training images were marked for regions of interest by an expert pathologist specializing in breast pathology. RoIs were primarily centered on ductal proliferative regions. This was done because most of breast pathology, including carcinomas, originates from the ductal epithelium. In those

cases with no ductal proliferations, cellular aggregations with atypical nuclear morphology were selected as RoI. Atypical nuclear morphology, in brief, was defined as follows; increased nuclear-to-cytoplasmic ratios, hyperchromatic nucleus and irregular nuclear membrane contour. Some examples of the labeled microscopy images along with their region wise labels are shown in Fig. 1. The annotation of regions is less labor intensive than detailed pixel level detailed annotation.

### III. GUIDED SOFT ATTENTION NETWORK

Since the labeled RoIs denote the areas in the microscopy images which consist of diagnostically relevant regions, we propose a Guided Soft Attention (GuSA) network which aims at localizing these RoIs, and simultaneously use them to guide the classification network. In this way the prediction of the network is based on the regions which a pathologist expects the network to focus on. The architecture of GuSA is presented in Fig. 2.

The network consists of two branches; (a) a RoI prediction branch to localize the diagnostically relevant regions, and (b) a classification branch to identify the type of breast histology image. The backbone of the network is a CNN based feature extraction network. The features at different layers of



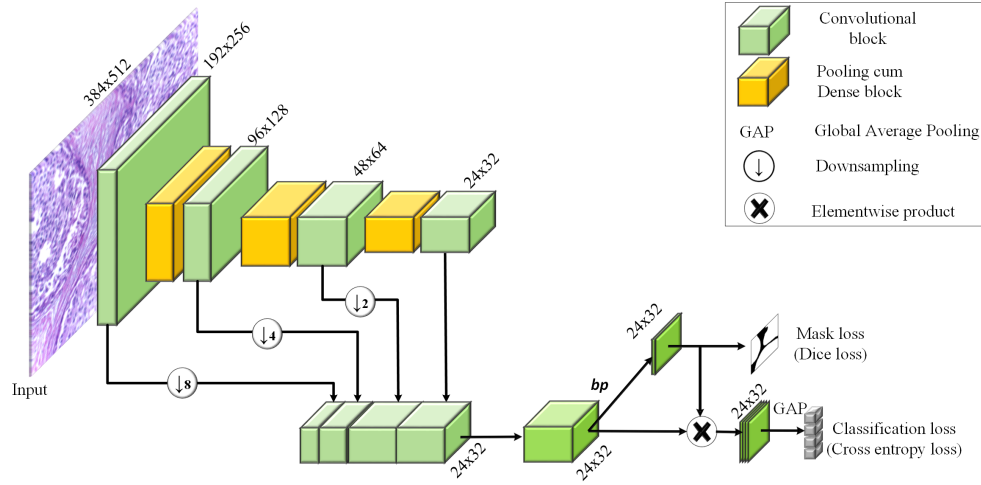


Fig. 2. The proposed explicitly guided attention network for classification of breast cancer microscopy images. Only the spatial dimensions of the feature maps are indicated.

the feature extraction network are aggregated, reduced and then fed to the RoI prediction branch and the classification branch. Given a breast microscopy image, the network is trained to simultaneously (a) predict the region of interest in the image, (b) use the predicted RoI to guide the activation maps for classification, and (c) classify the image into the four target categories.

#### A. Region Supervised RoI Prediction Network

Given a microscopy image as input, the RoI prediction branch outputs a map indicating the diagnostically relevant regions in the image. This branch is trained with the coarse region level labels, described in Section II, where the regions are annotated without considering the detailed pixel wise class contents inside the region. Branching from the backbone network at *bp* (bifurcation point), as shown in Fig 2, this branch consists of an additional  $1 \times 1$  convolutional layer. Finally, a sigmoid layer is added to normalize the output feature map by

$$\tilde{y}_{ij} = p(c|x_{ij}) = \frac{1}{1 + \exp(-x_{ij})} \quad (1)$$

where  $x_{ij}, i \in \{1, 2, \dots, M\}, j \in \{1, 2, \dots, N\}$  are the pixels of the output with size  $M \times N$  and  $p(c|x_{ij})$  represents the probability score of  $x_{ij}$  belonging to the  $c^{\text{th}}$  class,  $c \in \{0, 1\}$ .  $c = 1$  indicates diagnostically relevant pixels whereas other pixels are marked zero. The RoI prediction branch is trained with the region level masks by minimizing the dice loss [30] between the predicted output map  $\tilde{y}$  and the ground truth mask  $y$  given as

$$L_{\text{dice}} = \frac{2 \times \left( \sum_{i=1}^M \sum_{j=1}^N y_{ij} \tilde{y}_{ij} + \varepsilon \right)}{\left( \sum_{i=1}^M \sum_{j=1}^N (\tilde{y}_{ij} + y_{ij} + \varepsilon) \right)} \quad (2)$$

#### B. Guided Soft Attention

The output of the RoI prediction branch is an output map  $\tilde{y}$  where the value of each pixel  $\tilde{y}_{ij} \in [0, 1]$  indicates the

probability of that pixel belonging to a RoI. The values in  $\tilde{y}$  directly indicate the relevance of the spatial location for diagnosis. This output map can be used to locate the diagnostically relevant regions and guide the attention of the network for classification of microscopy images.

Given an input image, let  $A^k(i, j)$  represent the activation at the spatial location  $(i, j)$  in the  $k^{\text{th}}$  channel of the output of *bp* layer, where  $i \in \{1, 2, \dots, M\}, j \in \{1, 2, \dots, N\}, k \in \{1, 2, \dots, K\}$ ,  $M \times N$  is the spatial size of the feature map and  $K$  is the number of channels ( $M = 24, N = 32, K = 512$ ). The attention of the classification branch is then guided by the *soft attention* mechanism [26] whereby the activations  $A^k(i, j)$  are weighed with  $\tilde{y}(i, j)$  along the channels. Hence, the guided activations  $A_G^k(i, j)$  are obtained as,

$$A_G^k(i, j) = A^k(i, j) * \tilde{y}(i, j) \quad (3)$$

The activations belonging to the diagnostically irrelevant regions are suppressed whereas (3) has little effect on the activations belonging to the relevant regions.

Some examples of the guided activations are shown in Fig. 3. A binary mask  $\tilde{y}^B$  with  $\tilde{y}_{ij}^B \in \{0, 1\}$  can also be generated from  $\tilde{y}$ , after thresholding with an appropriate  $\tau$ , to locate the regions of interest and *hard attention* can be used to guide the classification network. However, the value of  $\tau$  controls the shape and size of the resultant RoI and selecting optimal  $\tau$  is not a trivial task. Hence, a *soft attention* mechanism is used in this study.

#### C. Classification Network

For a microscopy input image, the classification branch predicts a category for the input image among the four different target classes; *normal*, *benign*, *in-situ carcinoma* and *invasive carcinoma*. Bifurcating from the backbone network at *bp*, activations of the *bp* layer are guided using RoI as explained in Section III (B). The guided activations are further reduced using  $1 \times 1$  convolutional filters to output four feature maps. Global average pooling (GAP) is then used to further reduce the dimensionality of the feature maps, followed by softmax

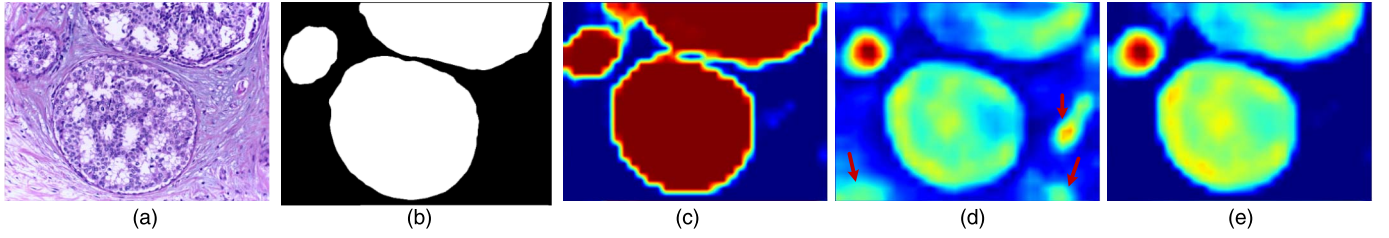


Fig. 3. Guided activations. (a) Source image, (b) ground truth RoI, (c) the predicted RoI map  $\tilde{y}$ , (d) The activations  $A^k$  of  $bp$  layer. For visualization purpose the mean of the activations across all the channels is computed. (e) Guided activations  $A_G^k$  where the irrelevant regions (marked using red arrows in (d)) are suppressed.

to normalize the output GAP values  $Z = \{z_1, \dots, z_C\}$ ,  $C = 4$ , where  $C$  is the number of target classes.

$$\text{softmax}(Z)_j = \frac{e^{z_j}}{\sum_{c=1}^C e^{z_c}}, \quad j \in \{1, 2, \dots, C\} \quad (4)$$

The output of *softmax* is a probability distribution over the  $C$  different possible classes and the final outcome  $y_{pred}$  (predicted class) of classification is the class with the maximum probability,

$$y_{pred} = \arg \max_j (\text{softmax}(Z)_j) \quad (5)$$

The classification branch is trained with training images labeled with the four classes by minimizing the cross entropy loss defined as

$$L_{CE} = - \sum_{c=1}^C 1_y(c) \log(\text{softmax}(Z)_j) \quad (6)$$

where  $1_y(c)$  is a binary indicator function indicating if the class label  $c$  is the correct classification for that image. The overall classification loss is the mean of the cross entropy loss computed over the training batch

$$L_{CE\_total} = \frac{1}{L} \sum_{i=1}^L L_{CE}^i \quad (7)$$

where  $L$  is the batch size.

The proposed network was trained to simultaneously predict the RoIs and classify the microscopy images into four categories. Hence, the total loss for training the network is

$$L_{total} = L_{CE\_total} + \lambda L_{dice} \quad (8)$$

where  $\lambda$  is a weighting parameter to balance the two different objectives. Both the components to the left of (8) are at the same numerical level, hence  $\lambda = 1$  is used in our experiments.

#### IV. EXPERIMENTS AND RESULTS

The performance of the proposed guided soft attention network is evaluated in this section. The experimental dataset preparation, evaluation metrics and the experimental settings are introduced followed by the performance evaluation of classification and RoI localization.

##### A. Dataset Preparation

The proposed approach was evaluated on the BACH dataset. In addition to the image wise labels provided by the challenge for the training images, additional region-wise annotation<sup>1</sup> of diagnostically relevant regions were prepared by our in-house expert as described in Section II. The original RGB microscopy images were down-sampled to  $512 \times 384$  pixels to reduce the memory and computation overhead. The images were color normalized using the normalization technique described in [31]. Extensive data augmentation was performed using vertical and horizontal mirroring, random rotations, addition of random noise, and random change in intensity of the images. The images were finally normalized using the mean and standard deviation computed from ImageNet [32].

A k-fold ( $k = 4$ ) cross validation was used to evaluate the proposed method. The training dataset was divided into four folds with each fold containing 25% of the overall training samples. The number of samples from each class were equally distributed among all the folds. During training three of the folds were used as train set whereas the remaining set was used for validation.

##### B. Network Architecture and Training

DenseNet-169 [7] network pre-trained on ImageNet was used as the backbone network. Only the layers up to Dense Block (DB\_3) of the original DenseNet were retained in the experiments. The details of the inputs, outputs and the size of the convolutional filters used in each layer of the proposed network are given in the supplementary material.<sup>2</sup> The layers retained from DenseNet are shaded in gray.

The network was optimized using Adam [33] with a mini-batch size of 5. All the layers of the network were fine-tuned with a learning rate starting from  $5e^{-5}$ . The learning rate was reduced by 10% after every 10 epochs and all the layers were batch normalized and L2 regularized ( $\lambda_{L2} = 1e^{-5}$ ). The proposed network was implemented with the Tensorflow framework [34] using NVIDIA Titan X GPU.

##### C. Evaluation Protocol

The performance of the RoI detection branch was evaluated using the mean intersection over union metric given as

$$IoU = \frac{y \cap \tilde{y}^B}{y \cup \tilde{y}^B} \quad (9)$$

<sup>1</sup>The mask labels are available from the corresponding author on reasonable request.

<sup>2</sup>Details of the network are given in Table I of the supporting document.

TABLE I  
COMPARISON OF VARIOUS METHODS ON BACH [19] MICROSCOPY IMAGE DATASET

NETWORK	ACCURACY (VALIDATION SET) %	AUC (VALIDATION SET)	TEST ACCURACY (%)		#PARAMETERS
			SINGLE MODEL	ENSEMBLE	
Baseline (B)	87±2.15	0.97721	80	87	12,332,740
B + Feature Aggregation (FA)	87 ± 0.5	0.97895	87	89	7,022,792
B + Guided RoI (GuRoI)	89.25 ± 1.18	0.9865	87	90	6,483,013
B + FA + GuRoI	88.75 ± 2.55	0.98315	84	90	7,024,844
B + Guided Soft attention (GuSA)	89.75 ± 1.65	0.9845	86	91	6,483,013
B + FA + GuSA ( <i>proposed</i> )	<b>90.25 ± 1.84</b>	0.98425	<b>88</b>	<b>93</b>	7,024,844
Attention gate [28]	85.5 ± 1.34	0.97673	85	87	19,368,390
Region Guided Attention [29]	89.25 ± 1.81	0.98413	84	88	7,024,844

where  $y$  is the ground-truth mask and  $\tilde{y}^B$  is the binarized version ( $\tau = 0.5$ ) of the predicted output map  $\tilde{y}$ . As region wise annotation were obtained only for the training set, the IoU metric is reported for the validation data set.

The performance of the classification branch was computed using the overall prediction accuracy, i.e., the ratio between the correctly classified samples and the total number of evaluated samples. The classification accuracy, area under the curve (AUC) and receiver operating characteristic (ROC) curve is reported for the validation set whereas only the classification accuracy is reported for the test set. As four-fold cross validation was used to evaluate the model, the final inference on the test set was obtained from the ensemble of networks trained and validated on the different folds, as shown in Fig. 4.

The final probability distribution over the four classes was obtained as the average of the probability distribution predicted by the individual models of the ensemble.

$$\text{softmax}_{\text{test}}(Z)_j = \frac{1}{L} \sum_{l=1}^L \text{softmax}(Z)_j^l \quad (10)$$

where,  $l \in \{1, 2, \dots, L\}$ ,  $L = 4$ . As the test set labels are not made public (are hidden), the final classification accuracy was obtained by uploading the predictions  $y_{\text{pred}}$  of the ensemble model to the challenge server. Results of single model are also reported on the test set.

## D. Results

The classification accuracy of different networks on the validation and test set are presented in Table I and the details of the networks can be found in the supplementary material.

**1) Baseline Network:** One of the top performing networks on the BACH challenge 2018 is the network of Chennamsetty *et al.* [15], where an ensemble of DenseNet-based networks achieved an accuracy of 87% on the test set. The implementation of the *Baseline* network is similar to [15] except for the image normalization scheme and the ensemble formation method. A single image normalization scheme is used and the ensemble consists of networks with the same architecture but trained on different folds of the training data, as shown in Fig. 4, instead of using different networks. As seen from Table I the accuracy of the baseline implementation on the test dataset is equal to the performance reported in [15].

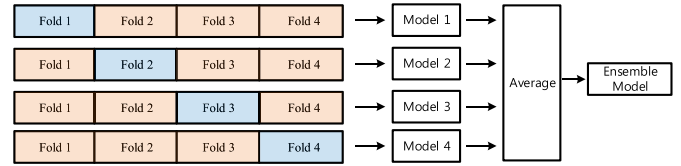


Fig. 4. The final inference was obtained from the ensemble of models trained on different folds.

**2) Multi-Scale Feature Aggregation:** The microscopy image consists of different diagnostic features such as cellular proliferation, architectural complexity of the proliferation, cellular atypia (increased nuclear-to-cytoplasmic ratios, hyperchromatic nucleus, irregular nuclear membrane contour and prominent nucleoli), presence or absence of necrosis etc. at different scales. To capture these features at different scales, the activation maps at different depths of the *Baseline* network are aggregated and reduced for final classification in the *Feature Aggregation (FA)* network. Aggregation of features from different depths of the network improves the accuracy of the baseline model by 2%.

**3) Region Guided Soft Attention (Proposed):** In addition to the classification branch of *Baseline* and *FA*, a region supervised RoI prediction branch is included in the *Guided RoI (GuRoI)* network and trained in a multi-task learning framework. The addition of the RoI prediction branch to localize the diagnostically relevant regions further achieves 3% and 1% improvement over *Baseline* and *FA*, respectively. In the proposed *Guided Soft Attention (GuSA)* network, the predicted RoI is explicitly used as a soft attention mechanism for classification as explained in Section III. *B+GuSA* leads to an improvement of 1% over *B+GuRoI*, and *B+FA+GuSA* achieves 3% improvement in accuracy over *B+FA+GuRoI*. We see that the explicit guidance mechanism in *GuSA* leads to higher accuracy than the simple multi-task learning of *GuRoI*. From Table I, we see that the proposed *B+FA+GuSA* network achieves a 6% improvement over the previous state-of-the-art [15].

In Table I, we also report the results of the best performing single models on the test set to decouple the improvement gain attained due to *GuSA* and the ensemble. As expected, we see that the performance of all the networks is improved by ensembling. The *Baseline* model benefits the most from ensembling

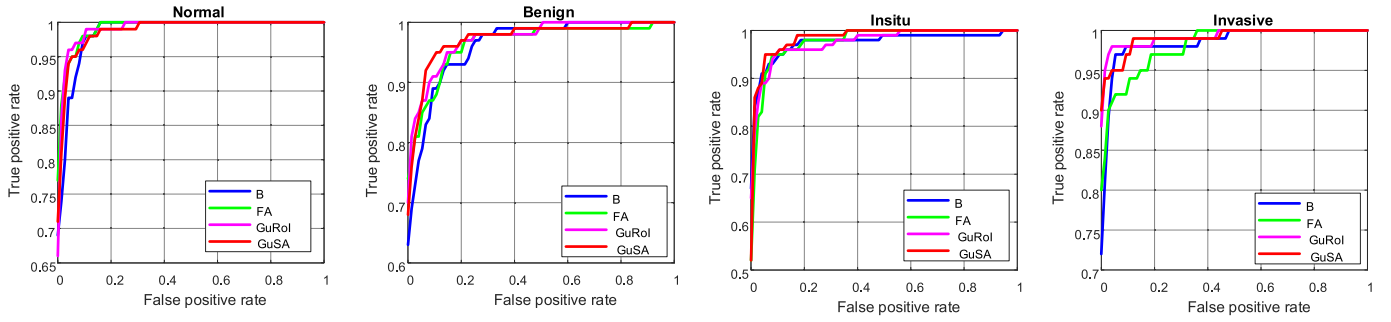


Fig. 5. Class wise ROC curves computed on the validation dataset.

TABLE II

PERFORMANCE OF DIFFERENT NETWORKS ON DETECTING THE ROI

NETWORK	MIOU (VALIDATION SET)
B + FA + GuRoI	0.782 ± 0.011
B + FA + GuSA	0.772 ± 0.014

whereas the benefit for *FA* is marginal. From the single model results, we see that multi-task learning of *GuRoI* and the guided attention of *GuSA*, both lead to significantly higher accuracy than the single model result of *Baseline*. Moreover, the single model performance of *B+GuRoI*, *B+GuSA*, and *B+FA+GuSA* is comparable to that of *Baseline* ensemble.

The performance of the ROI prediction branch of the *GuRoI* and *GuSA* networks are also presented in Table II.

4) *Other Attention Mechanism*: In Table I we also present a comparison of the proposed region supervised attention with the unsupervised attention mechanism of [28] and the implicit region guided attention of [29]. The performance of the unsupervised attention mechanism of [28] when implemented for the BACH dataset produces a performance that is similar to the *Baseline* network. This is due to the fact that this attention mechanism depends on the last convolutional layer of the network which still locates the most discriminative regions only. These regions are not necessarily the diagnostically relevant regions required for improved generalization performance.

The network [29] with additional region guidance mechanism of diagnostically relevant regions shows an improvement of 1% over the *Baseline* network on the test set. This attention mechanism is implicit and does not fully utilize the additional guidance. However, our proposed method explicitly utilizes the external region guidance to suppress the activations of diagnostically irrelevant regions which results in a marked improvement in performance over the implicit region guidance as evident in Table I.

## V. ALL CLASSES ARE NOT EQUAL

As the ground truth labels for test data is not public, we analyze the class-wise performance of the proposed method on the validation dataset. The class-wise receiver operating characteristic (ROC) curves, and area under the curve (AUC) are presented in Fig. 5 and Table III, respectively.

From the ROC curve and AUC table, we see that the performance of the network is better for normal and invasive carcinoma than for benign and in-situ carcinoma. The higher

TABLE III

CLASS-WISE AUC OF DIFFERENT NETWORKS

CLASSES (→) NETWORK (↓)	NORMAL	BENIGN	INSITU	INVASIVE
Baseline (B)	0.9857	0.9627	0.9750	0.9855
B+FA	<b>0.9921</b>	0.9652	0.9776	0.9809
B+FA+GuRoI	0.9912	<b>0.9731</b>	0.9758	<b>0.9925</b>
B+FA+GuSA	0.9882	0.9728	<b>0.9851</b>	0.9909

AUC shows that it is comparatively easier for the network to discriminate between the classes at the extremes of the diagnostic category than the intermediate classes.

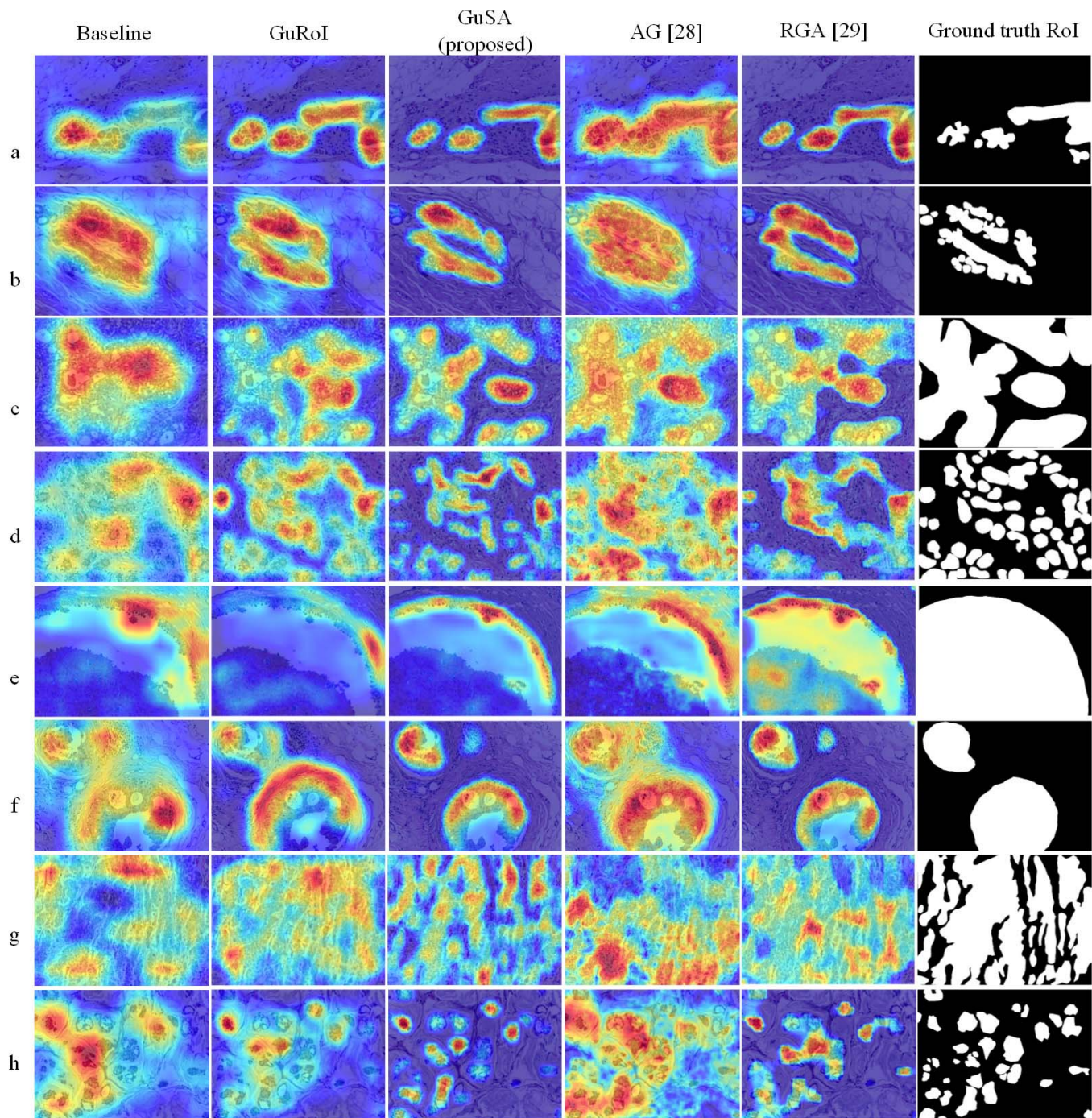
Moreover, from the confusion matrices presented in Table IV, we see that the network has difficulty in discriminating between nearby categories, and the network is less confused while discriminating well separated classes. This is similar to the difficulty experienced by pathologists in actual practice, where the ambiguity of overlapping features of nearby categories often calls for consultations among pathologists and extensive auxiliary studies such as immunohistochemistry and molecular studies. As seen from Table IV, the proposed soft attention network has improved performance in discriminating the nearby categories over the baseline network.

## VI. VISUALIZATION OF ACTIVATION/ATTENTION MAPS

The generated activation (attention) maps for different methods are presented in Fig. 6. The attention maps for the Baseline network are generated using Grad-CAM [24], whereas the attention maps for the *Attention gate*, *RGA*, and the proposed method are generated directly from the network. The attention map of *RGA*, and the proposed method are generated from the *C6* and *RoI\_mask* layers of their respective networks.

From Fig. 6, we can see that the Baseline network has high activations in diagnostically irrelevant regions of non-proliferative mammary stroma. The attention map of the AG network shows high activations in both irrelevant and relevant areas alike. This is because the self-attention mechanism in AG selects the most discriminative regions which may not necessarily be diagnostically relevant. The region guidance of *RGA* [29] shows high activations in relevant regions and comparatively fewer activations in irrelevant areas. However, significantly improved from other models, our proposed method of explicit attention mechanism specifically activates neurons





**Fig. 6.** Visualization of the activation maps for the Baseline, GuRoI, GuSA (proposed), Attention gated network [28], Region guided network [29], and the ground truth RoI. Starting from the top two examples are shown for each of the normal, benign, insitu carcinoma, and invasive carcinoma class. The activation maps are superimposed on the source image taken from validation dataset.

in diagnostically relevant areas guided by the predicted RoI mask.

The RoIs are annotated on diagnostically relevant regions without considering the exact pixel-wise class labels inside the region or the precise boundary of the region. This method is less labor intensive and cost effective than detailed pixel level detailed annotation, and the RoI based attention mechanism improves the accuracy of classification. However, this region level supervision may lead to activations in irrelevant features present inside the RoIs. For example, the entire duct is

annotated as RoI in Fig. 6 (e) & (f), instead of finely annotating only the ductal epithelium as RoI. As seen from Fig. 6 (e), RGA demonstrates nonspecific high activations in the stroma, a region of insignificance in addition to the epithelial cells. However, the proposed method produces high activations only in the epithelial cells. Some of the failure cases of the proposed method are as shown in Fig. 7.

Figure 7 (a) & (b) show that the proposed method incorrectly activates diagnostically irrelevant luminal spaces within the duct. Detailed pixel level annotations may be used to



TABLE IV  
CONFUSION MATRICES OF MICROSCOPY IMAGE CLASSIFICATION

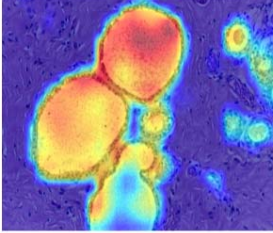
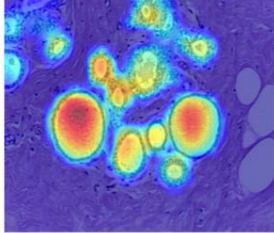
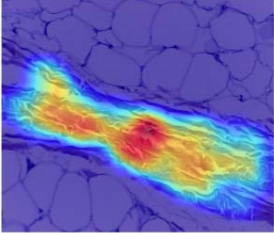
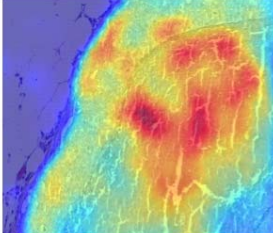
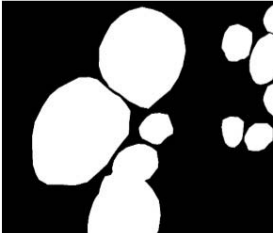







	Normal	Benign	Insitu	Invasive		Normal	Benign	Insitu	Invasive		Normal	Benign	Insitu	Invasive		Normal	Benign	Insitu	Invasive
Normal	0.91	0.06	0.03	0	Normal	0.91	0.07	0.01	0.01	Normal	0.95	0.03	0.01	0.01	Normal	0.93	0.05	0.02	0
Benign	0.1	0.81	0.04	0.05	Benign	0.08	0.84	0.05	0.03	Benign	0.09	0.85	0.05	0.01	Benign	0.1	0.83	0.05	0.02
Insitu	0.03	0.04	0.9	0.03	Insitu	0.05	0.07	0.85	0.03	Insitu	0.04	0.1	0.84	0.02	Insitu	0.01	0.05	0.94	0
Invasive	0.01	0.05	0.08	0.86	Invasive	0.01	0.02	0.09	0.88	Invasive	0.02	0.03	0.04	0.91	Invasive	0.01	0	0.08	0.91
<b>B</b>					<b>FA</b>					<b>GuRoI</b>					<b>GuSA</b>				
																			
																			
																			
(a)					(b)					(c)					(d)				

Fig. 7. Some failure cases of the proposed RoI-based guidance: (top) source image, (middle) ground truth RoI, and (bottom) predicted RoI.

rectify this problem however significant time and cost is incurred in obtaining these annotations. Moreover, CNNs have shown good segmentation performance on larger objects and the performance degrades for small and thin objects [35]. As seen in Fig. 7(c), in the absence of ductal structures, a single nucleus present within the parenchymal stroma was annotated, however due to the scant number of nucleated cells present, the predicted mask is imprecise as higher activations highlight the stroma instead of the target nuclei. This problem can be mitigated by training on larger dataset with thousands of labeled microscopy images. In Fig. 7(d) higher activations is produced from areas of necrosis rather than viable cells. As the dataset consists of limited number of training samples featuring necrosis, hence training on a larger dataset with various scenarios can mitigate this problem as well.

## VII. CONCLUSION

A convolutional neural network with additional region-level supervision which provides explicit guidance on the attention

maps for the classification of breast cancer histopathological images was proposed in this study. The proposed method simultaneously localized the diagnostically relevant regions of interest (RoI) in the microscopy images, and used the predicted RoI to guide the attention of the classification branch. Experiments demonstrating the effectiveness of the proposed method were presented whereby the RoI guided soft attention mechanism specifically activated neurons in the diagnostically relevant areas while suppressing neurons in noisy and irrelevant areas. Moreover, the proposed method surpassed the state-of-the art on the BACH microscopy test set (part A) by a significant margin.

The activations produced by the proposed method can be directly applied to interpret the decision of the neural network without the need of post processing steps like CAM or Grad-CAM. The proposed method makes the neural network decisions transparent and reliable as the produced feature activations are in agreement with the findings that would be expected by a human expert.

The proposed method was trained and evaluated on a limited dataset provided by the BACH challenge. Though the reported test accuracy of different methods were significantly different, the validation accuracy of different models were comparatively similar albeit with large standard error among different validation folds for each model. This is due to the small size and distribution of the training data.

We believe that the reported results are still suboptimal for clinical use. As deep neural networks exhibit improved performance with the increase in the number of training samples, availability of a larger dataset in the future is expected to further enhance the performance of the proposed system and make it suitable for clinical use. Availability of a larger dataset will also help to better ascertain the performance gain obtained by any future enhancements.

Strengthened by the combined benefit of transparent network decision and enhanced diagnostic performance, the proposed method can be used to build a reliable and accurate computer-assisted diagnosis (CAD) system to automatically detect and categorize the pathology present in breast cancer histopathological images.

## REFERENCES

- [1] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak, "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1563–1572, Oct. 2013.
- [2] P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE Trans. Med. Imag.*, vol. 32, no. 12, pp. 2169–2178, Dec. 2013.
- [3] Y. M. George, H. H. Zayed, M. I. Roushdy, and B. M. Elbagoury, "Remote computer-aided breast cancer detection and diagnosis system based on cytological images," *IEEE Syst. J.*, vol. 8, no. 3, pp. 949–964, Sep. 2014.
- [4] A. D. Belsare, M. M. Mushrif, M. A. Pangarkar, and N. Meshram, "Classification of breast cancer histopathology images using texture feature analysis," in *Proc. TENCON IEEE Region Conf.*, Nov. 2015, pp. 1–5.
- [5] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
- [6] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
- [7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.
- [8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
- [9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
- [10] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2097–2106.
- [11] A. Cruz-Roa *et al.*, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," *Proc. SPIE*, vol. 9041, Mar. 2014, Art. no. 904103.
- [12] T. Araujo *et al.*, "Classification of breast cancer histology images using Convolutional Neural Networks," *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0177544.
- [13] B. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Sci. Rep.*, vol. 7, no. 1, Jun. 2017, Art. no. 4172.
- [14] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *J. Pathol. Inform.*, vol. 7, p. 29, Jul. 2016.
- [15] S. Chennamsetty, M. Safwan, and V. Alex, "Classification of breast cancer histology image using ensemble of pre-trained neural networks," in *Image Analysis and Recognition*, A. Campilho, F. Karay, and B. ter Haar Romeny, Eds. Cham, Switzerland: Springer, 2018, pp. 804–811.
- [16] S. Kwok, "Multiclass classification of breast cancer in whole-slide images," in *Image Analysis and Recognition*, A. Campilho, F. Karay, and B. ter Haar Romeny, Eds. Cham, Switzerland: Springer, 2018, pp. 931–940.
- [17] B. Gecer, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks," *Pattern Recognit.* vol. 84, pp. 345–356, Dec. 2018.
- [18] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, "Y-Net: Joint segmentation and classification for diagnosis of breast biopsy images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 893–901.
- [19] G. Aresta *et al.*, "BACH: Grand challenge on breast cancer histology images," Aug. 2018, *arXiv:1808.04277*. [Online]. Available: <https://arxiv.org/abs/1808.04277>
- [20] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. ICLR Workshop*, 2014, pp. 1–8.
- [21] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. ICLR Workshop*, 2015, pp. 1–14.
- [22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.
- [23] B. Zhou, A. Khosla, A. Lapiedra, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, Jun. 2016, pp. 2921–2929.
- [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, Oct. 2017, pp. 618–626.
- [25] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," Apr. 2018, *arXiv:1804.02391*. [Online]. Available: <https://arxiv.org/abs/1804.02391>
- [26] K. Li, Z. Wu, K. C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9215–9223.
- [27] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," Jan. 2018, *arXiv:1801.09927*. [Online]. Available: <https://arxiv.org/abs/1801.09927>
- [28] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.
- [29] J. Son, W. Bae, S. Kim, S. J. Park, and K.-H. Jung, "Classification of findings with localized lesions in fundoscopic images using a regionally guided CNN," in *Computational Pathology and Ophthalmic Medical Image Analysis*. Cham, Switzerland: Springer, 2018, pp. 176–184.
- [30] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 240–248.
- [31] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 5, pp. 34–41, Sep./Oct. 2001.
- [32] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [34] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.
- [35] L.-C. Chen *et al.*, "Searching for efficient multi-scale architectures for dense image prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8699–8710.