IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 9, 2020, accepted November 12, 2020, date of publication November 24, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040106

Detection of Breast Cancer From Whole Slide Histopathological Images Using Deep Multiple Instance CNN

KAUSIK DAS[1], SAILESH CONJETI[2], (Member, IEEE), JYOTIRMOY CHATTERJEE[3], (Member, IEEE), AND DEBDOOT SHEET[1], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Indian Institute Technology Kharagpur, Kharagpur 721302, India
[2]Siemens Healthineers, 91052 Forchheim, Germany
[3]School of Medical Science and Technology, Indian Institute of Technology Kharagpur, Kharagpur 721302, India

Corresponding author: Kausik Das (imkausikdas@iitkgp.ac.in)

ABSTRACT Histopathological Whole Slide Imaging (WSI) has become a standard in the detection of breast cancer. Automated image analysis methods attempt to reduce the workload from the clinicians and Convolutional Neural Networks (CNNs) are a popular choice for this purpose. However, size of a WSI image typically is approximately 40,000 × 40.000 pixels (can reach up to 100,000 × 100.000 pixels). CNNs cannot handle such large images. Moreover, downscaling a WSI image causes degradation of small-scale visual information. Hence, a large number of small patches (containing critical visual information) from a WSI image are extracted by a trained pathologist and are used for training. However, it requires massive amounts of time to precisely search and label appropriate class-representative patches. To address this issue, a Deep Multiple Instance Learning (MIL) based CNN framework has been introduced in this paper. In the proposed framework every slide is represented as a bag of extracted patches. Only the bag label is used for training, thus eliminating the requirement to provide patchwise labels. The patches inherit the label of the bag containing them. A WSI image (i.e. a bag) is labeled benign if all its patches are benign and labeled malignant even if a single patch contains malignant cells. Learning can be carried out at the bag level even with noisy patch labels. Performance of this method was evaluated using the BreakHis, IUPHL and UCSB breast cancer datasets where 93.06%, 96.63%, 95.83% accuracy was achieved respectively.

INDEX TERMS Computer-aided diagnosis, convolutional neural network, multiple instance learning, weakly supervised learning, whole-slide image analysis.

I. INTRODUCTION

Breast cancer (both sexes, all ages) constituted 11.6% (2,088,849 cases) of all types of cancers (18,078,957 cases) in 2018.[1] Moreover, breast cancer based mortality among women is very high, accounting to 42000 deaths in the USA alone, in 2020 [1]. Early detection is the key to reduce the number. Detection of breast cancer is done by palpation, followed by non-invasive mammogram based critical point identification. Invasive microscopic examination of biopsy samples extracted from the critical mass of the breast is carried out for confirmation and full profiling of the disease. Whole Slide Imaging (WSI) has become a standard practice in the field of microscopic investigation of tissue pathologies (including breast cancer). The process starts with collection of biopsy samples excised from a suspicious mass of tissue, followed by fixation, slicing, laying the tissue on a slide and chemically staining it to enhance the visibility of the cell nuclei, cytoplasm and inter-cellular matrix. High resolution microscopic imaging helps in visualizing different tissue structures and cellular features, enabling pathologists to differentiate primarily between benign and malignant tissues or their sub-types. Due to the increasing number of whole slide scans being performed, manual inspection of such huge images is time consuming and strainous. Also chances of missing out small cancerous regions on the slide are very high as malignant cells in some of the cases may be sparsely

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

[1]https://gco.iarc.fr/today/home

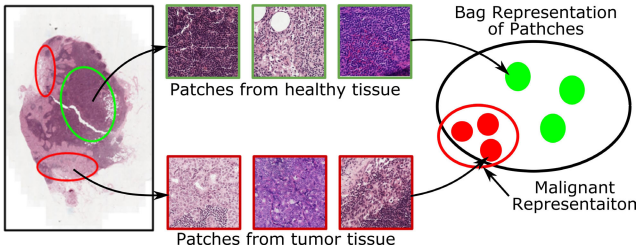


FIGURE 1. **In a WSI, we present the instances of healthy patches (indicated in green) along with cancerous patches (indicated by red), which visually appears very similar, making the task challenging. All the patches are *instances*, which together represent a *bag*, with a single label.**

located. In this context, machine learning based automated inspection techniques allow fast and reliable identification of cancerous cells and regions on the slide.

A typical WSI image is massive in size (varies between 40000 to 100,000 pixels), which a standard CNN [2]–[5] is unable to handle. Alternatively, resizing a WSI image in order to force fit into the CNN, puts down crucial fine-scale descriptive information. Conventionally, a pathologist looks for suspicious regions in a large WSI and inspects those separately. A large number of class representative patches from a WSI must be handpicked and labeled carefully to train the CNN. This process is very costly in terms of man-hours and suffers from inter-observer variability.

To address this problem, we propose a Multiple-Instance Learning (MIL) [6]–[8] based Deep CNN architecture. In MIL, a number of instances which collectively represent a single object or entity, are referred to as a 'bag'. The bag is associated with a label corresponding to that object. In our framework, each patch extracted from a WSI image corresponds to a single instance. All the patches extracted from a particular WSI image, together form a bag. This bag corresponds to a particular patient. It must be noted that the extracted patches can belong to benign, malignant or normal tissue regions of the WSI image. However, only the bag label is considered instead of the patch labels. A WSI image (i.e. a bag) is labeled benign if and only if none of its patches are malignant and labeled malignant even if a single patch contains malignant cells. Fig. 1 illustrates the formation of a bag corresponding to a cancerous WSI image, which comprises instances (patches) from both malignant and normal tissue regions.

In this paper, section II contains the prior art in this area. Section III states the formal definition of the learning problem and section IV contains the detailed discussion of our proposed method. It is followed by experimental validation and discussion of results in sections V and VI respectively. Section VII states the future scopes and concludes this paper.

II. PRIOR ART

Multiple Instance Learning (MIL) was first used for drug activity prediction [9]. Its subsequent variants such as diverse density (DD) [9], expectation-maximization of the DD function (EM-DD) [10], MI-support vector machine (MI-SVM) [11] also gained popularity. MI-SVM [11] typically fuses MI inferencing with the SVM framework. MILBoost [12] is another well-known method, which makes use of an ensemble of multiple weak classifiers, within the MI context. Some of the earlier works to make use of MI neural networks include [13], [14], where instance-level decisions are aggregated for the final classification. However, the techniques mentioned till now suffer from a common drawback. All of them are highly biased towards the presence of a positive instance in the bag, to make the final decision. This discards meaningful information conveyed by the other instances. Some of the recent approaches attempted to overcome this limitation by giving appropriate weightage to all the instances [15], or some of the high responding instances [16], [17] in a bag to aggregate for bag level predictions. Further advancements of MIL such as Multiple-Instance Learning via Embedded instance Selection (MILES) [18], Joint Clustering and Classification for Multiple Instance Learning (JC^{2}MIL) [19] and mi-Graph [20], introduced the concept of bag level comparisons for classification. MIL was first introduced the field of Deep CNN for computer vision [21]–[23], in [24], where bag-level aggregation of instances is performed in the label space. Spatial relation of instances guided by few labeled images was incorporated in [25]. Feature level aggregation within shallow neural network was implemented in [26]. Attention based selective aggregation of features for bag level prediction was done in [27]. Several works [28], [29] have shown hybrid applications of MIL such as classification, localization and pix-wise segmentation of objects.

MIL has been successfully deployed in multiple medical imaging problems, such as mass detection in digitized mammograms [25], [30]–[32], cancer detection in digital pathology images [33]–[37], ultrasound classification [38], diabetic retinopathy screening [33], [39], [40] and malenoma detection [41]. A recent work used MIL with Deep CNNs for WSI image classification [17]. Another recent article on deep MIL [42] proposed a two-stage process, where MI context is used to select the most discriminative patches for boosting CNN classifiers. In [43] sparsity was incorporated in the instance level predictions on a Deep CNN for cancer classification in digitized mammograms.

III. FORMAL DEFINITION

Let us consider $\mathbf{x}_i$ to be the i^{th} patch from the total N patches extracted from a collection of WSI images. In the supervised learning approaches, each of those patches (or instances) are associated with a class label ω_i to form the dataset $\mathcal{D} = \{(\mathbf{x}_1, \omega_1), \cdots, (\mathbf{x}_N, \omega_N)\}$, which is used for training the classifier. The classifier provides decisions in terms of $p(\omega|\mathbf{x}_i)$, estimated as

$$p(\omega|\mathbf{x}_i) = \mathcal{H}(\omega \leftarrow \mathbf{x}_i) \tag{1}$$

where the mathematical function $\mathcal{H}(\cdot)$ is approximated by the trained classifier.

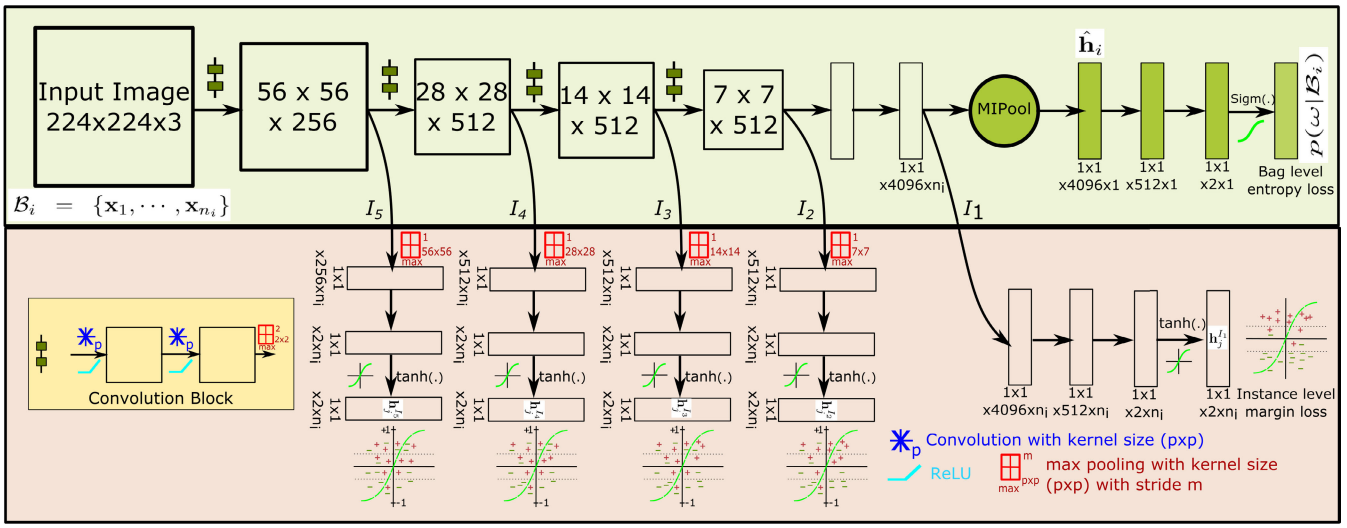


FIGURE 2. An illustration of the deep convolutional neural network with multiple instance level semi-supervised models. Bag level representation is achieved by the strategy of instance features aggregation through feature level max-pooling layer (MIPool) at the feature embedding dimension of 4096. Multiple margin-based instance level loss is introduced at multiple levels of architecture as side level supervisions of the architecture. It aims to perform the instance level classification by inheriting the label from its corresponding bag. In this figure, the upper block comes to play during training and inference both of the time, whereas the lower block comes to play only during training time.

However, in Multiple Instance Learning (MIL), multiple patches extracted from a single WSI image are treated as a bag $\mathcal{B}$, which is then associated with single class label. Thus the i^{th} WSI image is represented by the bag $\mathcal{B}_i = \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_i}\}$, containing n_i number of patches and the dataset used for training is given by $\mathcal{D}_{MIL} = \{(\mathcal{B}_1, \omega_1), \cdots, (\mathcal{B}_{N_B}, \omega_{N_B})\}$, where N_B denotes the total number of WSI images. In contrast to Eq. 1, $\mathcal{D}_{MIL}$ is used for training the classifier which approximates

$$p(\omega|\mathcal{B}_i) = \mathcal{H}_{MIL}(\omega \leftarrow \mathcal{B}_i = \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_i}\}) \tag{2}$$

IV. SOLUTION TO THE PROBLEM

In this article, we propose an architecture that effectively incorporates MIL within a CNN. Our method aims to overcome the drawbacks of the existing methods. These are listed below:

1) Our proposed approach requires only bag level annotations (in contrast to many existing approaches which use individual instance level annotations). Thus in a limited time-span labeling mulltiple WSI images becomes feasible for a pathologist, hence the cost in terms of man-hours is also reduced.
2) We introduce a Multi-Instance Pooling (MIPool) layer, placed at a high-dimensional feature space to aggregate multiple instances in a bag.
3) The MIPool layer in our deep CNN architecture sparsifies the gradients in the training process. To tackle this problem, we introduce single instance losses at multiple locations in the form of side-level supervisory arms. These use margin-level loss functions to prevent incorrect update of network parameters due to the presence of noisy instance labels.

This architecture is an extension of our earlier work [44]. In our previous works, the patches were strongly labeled. In this paper, we introduce a modified architecture that solves the task of classifying large sized WSI images using weakly labeled slides for training.

A. NETWORK ARCHITECTURE

Our MIL enabled deep CNN architecture is based on the pre-trained VGG-19 network [21]. The MIL paradigm is incorporated into the network by a Multiple Instance Pooling (MIPool) layer, which performs a feature level max-pooling operation, along the instances. This forms a unified bag-level descriptor, which is used in the final stages of the decision making process. We have placed the MIPool layer within the fully connected (FC) layers, as given in Fig. 2. In the FC layer, let us assume that the feature response for a given instance $\mathbf{x}_j$, be represented as $\mathbf{h}_j$ which is of dimension 4096×1. The MIPool layer operates on the bag $\mathbf{B}_i$, by taking the row-wise max on the feature response matrix returning the $\hat{\mathbf{h}}_i$ vector. Thus, if the feature responses of the instances are given by $\mathbf{h}_1, \ldots, \mathbf{h}_{n_i}$ for the bag $\mathbf{B}_i$, they are staked columnwise to form a matrix $\mathbf{M}$ of dimension $4096 \times \mathbf{n}_i$. We then obtain $\hat{\mathbf{h}}_i$ vector by stacking the row-wise maximum values from $\mathbf{M}$. The following operation is given as

$$\hat{\mathbf{h}}_{ip} = \max\{m_{pq} : 1 \leq q \leq n_i\} \tag{3}$$

where, $\hat{\mathbf{h}}_{ip}$ is the p^{th} element of $\hat{\mathbf{h}}_i$ and m_{pq} is the element at the p^{th} row and q^{th} column of the aggregation matrix $\mathbf{M}$. Thus $\hat{\mathbf{h}}_i$ characterizes the bag $\mathbf{B}_i$ by aggregating features from all the instances $\forall \mathbf{x}_j \in \mathcal{B}_i$. This is followed by a sequence of operations on $\hat{\mathbf{h}}_i$, consisting of FC layers, ReLU activation functions and dropout layers. These embed the bag descriptor $\hat{\mathbf{h}}_i$ to a further lower dimensional representation. Finally, it is

passed through a softmax layer, to obtain the probabilities of the bag belonging to each of the classes, represented as $\mathbf{h}_i^B := p(\omega|\mathcal{B}_i, \hat{\mathbf{h}}_i)$. Here, the set of CNN parameters $\Theta = \{\mathbf{W}, b\}$ are learned by optimizing the cross-entropy loss function at the bag level, which is achieved by stochastic gradient descent.

Such end-to-end learning of the MIPool based bag-level feature representation helps to (i) attentively boost the meaningful and discriminative instances, (ii) while suppressing the non-informative and weaker instances, (iii) and at the same time working towards the improvement of the bag level classification margin. Since the input to the network is a bag of instances, we removed all *batch-normalization* layers to preserve the descriptive information of all the instance-based feature responses. Batch normalization layers within the network would have blended non-informative features with the informative instance feature responses, hurting the impact of MIPool layer on the proposed architecture.

We also introduced multiple single instances arms $I_1, \ldots, I_K$, at different stages of the architecture, as exhibited in Fig. 2. Each arm consists of spatial max pooling operations of different spatial dimensions followed by multiple FC layers, ReLU activation function, and random weights dropout function. Lastly, the output embedding has been range limited to $[-1, 1]$ by passing it through the hyperbolic tangent activation function ($\mathbf{tanh}(\cdot)$) to generate an instance representation defined as $\mathbf{h}_j^{I_k}$ given the instance $\mathbf{x}_j \in \mathcal{B}_i$ and the k-th instance arm of the network. Here, we used single instance-level margin-based loss functions for training.

The rationale behind using the margin-based losses are: (i) their high generalization capability and (ii) robustness to noisy instance labels (i.e. instances with dissimilar labels from the bag) [40]. Such arms ensure the preservation of instance level information by learning from information-rich feature maps at multiple scales in the network. Apart from that, the dense gradients generated from these arms contribute towards the efficient training of the network. It must be noted that, the MIPool layer highly sparsifies the gradient flow and without the aid of these arms, the training process slows down and might not reach a feasible local minima.

B. BAG LEVEL ENTROPY LOSS

The bag level cross-entropy loss function can be defined as

$$l^B(\Theta) = -\frac{1}{N_B}\sum_{i=1}^{N_B} \omega_l \log\left(\mathbf{h}_i^B\right) \tag{4}$$

where, $\mathbf{h}_i^B = p(\omega_l|\mathcal{B}_i)$ is the estimated probability of the bag $\mathcal{B}_i$ belong to the class l. Here, ω_l is a $c \times 1$ one-hot ground truth vector.

C. INSTANCE LEVEL MARGIN LOSS

Along with cross-entropy loss at the bag level, we introduce margin-loss for the instance level classification problem which is defined for the k-th arm as,

$$l^{I_k}(\Theta) = \frac{1}{N_B}\sum_{i=1}^{N_B}\sum_{j=1}^{n_i} max(0, \lambda - \mathbf{h}_j^{I_k}\alpha_j) \tag{5}$$

where $\lambda \in (0, 1]$ is a tunable margin parameter. Here, $\alpha_j \in [-1, 1]$ is the binary represented ground truth label of instances inherited from its corresponding bag. As shown in Eq. 5, instances will be correctly classified if it lies on or beyond the margin, with $\mathbf{h}_j^{I_k}\alpha_j^{+/-} \geq \lambda$, on the other hand instances get misclassified when $\mathbf{h}_j^{I_k}\alpha_j^{+/-} < \lambda$.

In the margin-based loss, λ is an important hyper-parameter. Setting λ to a larger value will force the instances towards better separability. However, this tends to overfit the network. Similarly, setting λ to a smaller value can force the network to get affected by the noisy and weakly labeled instances in the learning process. Margin-based loss improves the accuracy by updating network parameters only for wrongly classified instances, while correctly classified instances do not contribute to the model updates [40]. On the other hand, the cross-entropy as a loss only captures the errors in the target class, while not getting biased by the prediction probability represented in the negative class, thus focusing only on improving the prediction performance of the true class [40]. Accordingly, entropy as a loss for instance level classification renders the network to be error-prone due to weakly and noisy labeled instances. To perform instance level loss over all the arms, we define $l^{I_K}(\Theta)$ as the sum of the individual arm losses. It is given by

$$l^{I_K}(\Theta) = \sum_{k=1}^{K} l^{I_k}(\Theta) \tag{6}$$

D. LEARNING OF MIL-CNN PARAMETERS

Training of the network is achieved by optimizing the combined loss function defined as

$$\Gamma(\Theta) = \lambda_1 l^B + \lambda_2 l^{I_K} + \lambda_3 \|\mathbf{W}\|_F^2 \tag{7}$$

where λ_1, λ_2 and λ_3 are used to give appropriate weightage to each of the individual losses as in Eq. (4) and Eq. (5) and $\|\mathbf{W}\|_F$ represents the Frobenious norm on W, which is the collection of weights of the proposed architecture. Here we estimate $\Theta = \{\mathbf{W}, b\}$ i.e. the weights of the layers and the bias by minimizing the overall cost function:

$$\Theta^* = \arg\min_{\Theta:\{\mathbf{W},b\}} \Gamma(\Theta) \tag{8}$$

where $\Theta^* = \{\mathbf{W}^*, b^*\}$ is the optimal parameter set that minimizes the overall cost function. The cost function is optimized using the mini-batch stochastic gradient descent (SGD) with back propagation of gradients. The gradients are calculated using the derivative of the cost function w.r.t. Θ i.e. $\frac{\partial\Gamma}{\partial\Theta}$. Now $\frac{\partial\Gamma}{\partial\Theta} = \frac{\partial\Gamma}{\partial\mathbf{h}} \cdot \frac{\partial\mathbf{h}}{\partial\Theta}$, where the second term $\frac{\partial\mathbf{h}}{\partial\Theta}$ is estimated via chain rule leading to the back-propagation of

gradients. The first term estimated from Eq. (7) as:

$$\frac{\partial \Gamma}{\partial \mathbf{h}} = \lambda_1 \frac{\partial l^B(\Theta)}{\partial \mathbf{h}_i^B} + \lambda_2 \frac{\partial l^{I_K}(\Theta)}{\partial \mathbf{h}_j^{I_K}} \tag{9}$$

The derivative term for bag loss is derived from Eq. (4) as

$$\frac{\partial l^B(\Theta)}{\partial \mathbf{h}_i^B} = -\frac{1}{N_B} \sum_{i=1}^{N_B} \frac{\omega_l}{\mathbf{h}_i^B} \tag{10}$$

and the gradients back-propagated through the MIPool layer is derived from Eq. (3)

$$\frac{\partial \hat{\mathbf{h}}_i}{\partial \mathbf{h}_j} = \begin{cases} 1, & \hat{\mathbf{h}}_i = \mathbf{h}_j \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Here, Eq. (11) performs masking operation over the back-propagation of gradients. Next, the derivative term for instance level margin loss in Eq. (5) is derived to be

$$\frac{\partial l^{I_k}(\Theta)}{\partial \mathbf{h}_j^{I_k}} = \begin{cases} \alpha_j, & \mathbf{h}_j^{I_k} \leq \lambda \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

V. EXPERIMENTS AND RESULTS

A. DATA-SETS

The proposed framework has been evaluated on three publicly available data-sets containing WSI images collected from breast tissue biopsies.

D1, Breast Cancer Histopathology Data-set, (*BreakHis* [2]) consists of H&E stained biopsy samples collected from breast tumor tissues, and imaged at four different magnifications, with labeling strategy as described in [45]. In this paper, we make use of only two magnifications 40× and 200× for our experiments. Multiple ROIs represented as instances are acquired from the WSI image (which is considered as a bag). It must be noted that each bag contains a varying number of ROIs or instances.

D2, The Indiana University Health Pathology Lab (IUHPL) dataset [46], has also been used for MIL related studies [47]. It consists of WSI images of histopathology slides collected from 40 patients : 20 of them representing precancerous ductal hyperplasia (UDH) and the other 20 representing ductal carcinoma in situ (DCIS). Due to the high size of these WSI images ($9,000 \times 8,000$ approx.), 653 ROIs each of size 1024×1024 px are collected in total. Multiple patches (as instances) each of size 224×224 px are further extracted from these ROIs (as bags).

D3, the UCSB Breast cancer data-set includes H&E stained breast tissue micro-array (TMA) of biopsy samples : 26 malignant and 32 benign cases. This dataset has also been used for bench-marking different MIL algorithms [34], [48]. Each image is of size 896×768 px. From these images, in our experiment, we have extracted patches of size of 224×224 px, with 50% overlapping between them to generate a set of non exclusive instances that forms a bag.

[2]http://web.inf.ufpr.br/vri/breast-cancer-database

The above three datasets have been tailored according to our MIL approach, keeping in mind : (1) A cancerous bag can contain multiple patches from normal or non-informative tissues along with a very few patches representative of malignancy. (2) On the other hand, it is necessary for a benign or normal bag to have no cancerous instance. Data-augmentation for the under-represented classes was achieved by bag level over-sampling. Rotation, random flipping and color jittering of the bag level instances was performed over every iteration to introduce randomness in the training process. Details of the number of instances per bag and the total number of bags from each of the above datasets, as used in our experiments are given in Tab. 1.

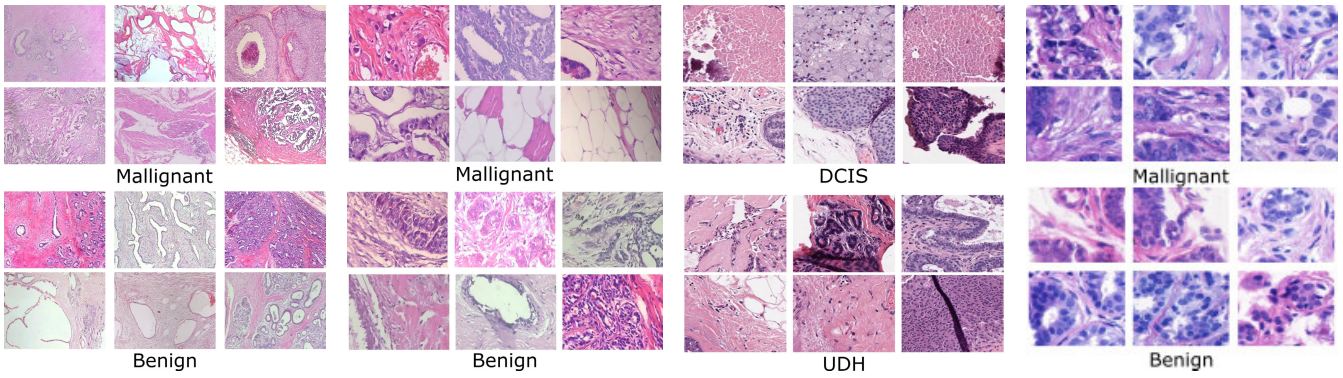
B. EXPERIMENTAL SETTINGS

In all ours experiments, we performed 3-fold cross validation by splitting the datasets into non-overlapping patient-level folds. Further, we resized all instances into the standard size of 224×224 px images to be consistent with the proposed architecture. The hyper-parameters in the loss function of Eq. (7) were set as, $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.001$. This combination was based on experimental observations. The SGD optimizer were used with variable batch size (depending on the number of instances in a bag). The learning rate was initially set to 0.001, decayed by 10% over every 50 epochs. Momentum was set to 0.9. Training was conducted for 300 epochs until early stopping. Early stopping was performed based upon the best validation accuracy. All the experiments were run on a workstation with 2× Intel Xeon ES2620 CPUs, 3× Nvidia Titan X GPUs with 12 GB DDRS RAMs and 128 GB of system RAM. Ubuntu 14.04 LTS was the OS and the experiments were implemented on Torch V7. All the baselines were run till convergence. The evaluation metrics used to quantify the classification performance were accuracy, recall and specificity. The implementation of the proposed method and different baselines are publicly available on the following link https://github.com/KausikDas-10/Deep_MIL_WSI

C. COMPARATIVE METHODS AND BASELINES

1) COMPARATIVE METHODS

The performance of our proposed framework was evaluated against the 10 conventional MIL algorithms and 3 recent deep learning based MIL approaches. The conventional MIL implementations were used from [49]. It must be noted, for implementing the conventional MIL based methods, *SIFT* features at the multiple scales (of 48×48 and 96×96, which helps to extract multi-scale features of the tissue regions from the pathology images) were computed and encoded for each instance using the bag-of-visual-words (BoVW) model with a dictionary size of 500. We also compared our proposed architecture against the state-of-the-art deep CNN methods. Those were: (i) MIL based CNN architecture proposed by *Wu et. al.* [24] (MIL-CNN), (ii) CNN-Vote, where pre-trained VGG network is trained based on the instances, and the final



(a) BreakHis at 40× Mag. (b) BreakHis at 200× Mag. (c) IUPHL datasets (d) UCSB Cancer

FIGURE 3. A set of four dataset images: (a) describes the pathology images from BreakHis datasets on 40×; and (b) on 200×; (c) the pathology images from IUPHL datasets, and (d) represent the pathology images collected from UCSB cancer database.

TABLE 1. Total configuration of the datasets is being tabulated here to fully comprehend the settings of the experiments.

Datasets	Number of Images							
	Positive				Negative			
	Total Patient	Instance	Bag	Instances/Bag	Total Patient	Instance	Bag	Instances/Bag
BreakHis 40×	54	1,370	54	(20-50)	24	625	24	(20-50)
BreakHis 200×	54	1,390	54	(30-60)	24	623	24	(30-60)
IUPHL	20	7,265	392	(10-45)	20	5,361	364	(13-40)
UCSB cancer	26	1,092	26	(40-45)	32	1,344	32	(40-45)

TABLE 2. Configuration of the different incremental baselines are introduced below, to validate the salient attributes of the proposed architecture. Here, FC-Dim defines Fully Connected (FC) Layer at feature dimension (Dim) of the proposed architecture.

Baselines	Loss Function		MIPool Position in the proposed architecture	Number of Instance Arm
	Bag Loss	Instance Loss		
BL-1	Entropy	–	FC-2	–
BL-2	Entropy	–	FC-4096	–
BL-3	Entropy	Margin	FC-4096	I_1
BL-4/Proposed	Entropy	Margin	FC-4096	$I_1 - I_5$
BL-5	Entropy	Entropy	FC-4096	$I_1 - I_5$

bag level labels estimated by average voting, (iii) Single Instance based CNN (SI-CNN), where classifiers are trained on the instance level and labels of the instances are borrowed from their corresponding bags.

2) BASELINES

In addition to the above comparative methods, we also present the six plausible variations of our proposed network, by varying the loss functions as discussed in Sec. IV-B and IV-C, position of the MIPool layer and the number of single instance losses as explored in Fig. 2. The configurations of all the baselines are shown in Tab. 2. The motivations of these baselines are to investigate: (1) the significance of the MIPool layer at different feature dimensions (BL-1-BL-2), (2) the importance of single side level supervision for classification (BL-2-BL-3), (3) the importance of proposed multiple side level instance-based margin losses (BL-3-BL-4), and lastly (4) the comparison between margin-loss and entropy loss at multiple side-level supervisory arms (BL-4-BL-5).

VI. DISCUSSION OF RESULTS

A. ANALYSIS OF PERFORMANCE W.r.t THE MARGIN PARAMETER λ

We have experimentally investigated the system performance with different values of the margin parameter, λ (in Eq. (5)).

TABLE 3. Performance of all the datasets with resepect to different λ parameters value are tabulated here. It can be noticed that the performance with high λ-value decrease over all the dataset which substantiates our observation.

Datasets	λ-Parameter		
	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.8$
BreakHis 40×	87.50	91.07	87.50
BreakHis 200×	93.04	93.04	91.33
IUPHL	96.39	96.63	95.09
UCSB cancer	94.03	95.83	89.55

We report the performance in terms of accuracy w.r.t. λ taking the values 0.2, 0.4 and 0.8 in Tab. 3. We observe that the performance of the proposed architecture performs best with $\lambda = 0.4$. The performance deteriorated with $\lambda = 0.8$, which substantiates our earlier statement that the margin loss with higher λ can make the model sensitive to noisy and weakly labeled instance images. On the other hand, we observe that at a lower value of $\lambda = 0.2$ the model fails to separate between the two classes. We keep the value $\lambda = 0.4$ for all our experiments based on this analysis.

B. COMPARISON WITH OTHER METHODS

We have compared our proposed framework with other MIL methods in Tab. 4, for each of the three data-sets. Across all the data-sets, our framework performed generally performed

TABLE 4. Comparison with different MIL comparative methods. The proposed MIL architecture achieved the best result on all data-sets.

Methods	Datasets					
	BreakHis on 40× Magnification			BreakHis on 200× Magnification		
	Accuracy	Recall	Specificity	Precision	Recall	Specificity
SI-SVM	77.42	**98.15**	11.11	73.25	**98.15**	0
mi-SVM [11]	77.42	98.15	11.11	73.25	98.15	0
MI-SVM [11]	77.42	79.30	72.22	63.41	64.16	61.11
MILBoost [12]	49.64	48.15	55.56	59.36	70.37	27.78
CkNN [50]	78.86	94.44	33.33	77.48	90.63	38.89
EMD-kernel [51]	80.43	88.89	55.56	76.03	88.67	38.89
MInD [52]	80.43	92.59	44.44	76.03	86.71	44.44
EMDD [10]	70.83	83.33	33.33	73.25	84.86	38.89
MILES [18]	71.98	75.60	61.11	66.67	79.08	66.67
migraph [20]	67.81	77.78	38.89	36.48	14.81	100
SI-CNN	80.33	98.15	27.78	83.11	98.15	44.44
CNN-Vote	88.78	88.67	88.89	88.71	90.50	83.33
MIL-CNN [24]	88.81	94.43	77.78	88.83	88.89	88.89
Proposed	**93.06**	94.44	**88.89**	**93.04**	94.44	**88.89**

Methods	Datasets					
	IUPHL			UCSB cancer		
	Accuracy	Recall	Specificity	Accuracy	Recall	Specificity
SI-SVM	57.02	99.44	11.25	56.25	**98.80**	9.06
mi-SVM [11]	57.61	99.15	12.78	56.25	98.80	9.06
MI-SVM [11]	67.11	48.74	86.92	59.25	89.01	17.14
MILBoost [12]	51.90	91.83	0	53.75	94.81	11.60
CkNN [50]	70.88	78.29	72.88	60.33	61.90	59.26
EMD-kernel [51]	81.27	83.36	79.02	66.50	42.86	85.14
MInD [52]	82.14	81.39	82.98	68.75	42.86	88.89
EMDD [10]	70.93	70.40	70.80	56.25	61.90	51.85
MILES [18]	77.14	78.82	78.33	62.50	57.14	66.67
migraph [20]	80.24	67.83	93.63	69.64	76.19	62.96
SI-CNN	85.10	98.87	68.57	87.72	96.80	71.34
CNN-Vote	88.40	88.19	89.66	87.91	86.27	89.71
MIL-CNN [24]	95.35	95.09	**96.63**	93.75	85.71	96.29
Proposed	**96.63**	**96.89**	96.34	**95.83**	90.48	**100**

better than the other methods. Non deep learning approaches like CkNN [50], EMD-Kernel [51], MInD [52] and MILES [18] exhibits similar performance ranges, while our framework out-performs them by a margin of 10-15%, 12-14%, 12-15% and 13-21% respectively, on all the three data-sets. In case of SI-SVM, we observe a significant drop in specificity, as compared to all other MIL methods including ours. This indicates the superiority of MIL approaches over Single Instance Learning (SIL) based approaches. We have also compared with some existing deep learning based approaches. We observe that SI-CNN performed better than all other non deep learning based MIL methods on the basis of bag-level classification performance as reported in Tab. 4. This indicates the fact that the quality of CNN learnt features are much superior than those of the hand-crafted ones. However, SI-CNN doesn't leverage the multi-instance learning strategy, and are also vulnerable to noisy instance labels. A similar behavior is observed for CNN-Vote approaches, where all the instances are given similar importance. Next, we compared with existing MIL-CNN [24], which is philosophically closest to our framework. Interestingly MIL-CNN exhibits the second best performance, substantiating the effectiveness of the MIL based CNN approach. Our approach outperforms MIL-CNN by 1-5% accuracy across all the data-sets. It must be noted that in contrast to pooling multiple instances in a bag after the softmax layer, aggregating them at the feature level provided the extra boost in performance (as evident in Tab. 4). However, a few popular CNN based MIL methods [25], [28], [29] demonstrate state-of-art performance on natural sized image. We could not apply these methods to extremely large sized WSI images and hence comparison with our method couldn't also be done.

C. ANALYSIS OF PERFORMANCE W.R.T. MIPool's POSITION IN THE ARCHITECTURE

Among the baselines, BL-1 uses the pooling at FC-2, whereas BL-2 uses it in FC-4096 of the architecture. Here, FC-2 and FC-4096 defines Fully Connected (FC) Layer at feature dimensions 2 and 4096 (as given in Tab. 2). Comparing the classification accuracies of BL-1 and BL-2 in two of the datasets, we observe a marginal increase of $\sim 2\%$ (as given in Tab. 5). This indicates that aggregating instances at a higher dimensional feature space yields better bag-level representation learning, in contrast to aggregating them at the label space. Thus, aggregating with MIPool at FC-4096 was used throughout rest of experiments. As mentioned in Sec. II, [26] and [27] introduced similar feature level pooling functions and proved its significance on shallow neural networks. However, such pooling layers also introduce the vanishing gradient problem on large CNN networks. In order to solve this

TABLE 5. Performance of different baselines are being tabulated here. Here, Acc., Rec., Spe. and F1-Score represent Accuracy, Recall, Specificity and F1-score repectively.

Baselines	Datasets															
	BreakHis 40×				BreakHis 200×				IUPHL				UCSB cancer			
	Acc.	Rec.	Spe.	F1	Acc.	Rec.	Spe.	F1	Acc.	Rec.	Spe.	F1	Acc.	Rec.	Spe.	F1
BL-1	88.81	94.43	77.78	93.15	88.83	88.89	88.89	92.10	95.34	95.09	96.63	88.01	93.75	85.71	96.29	90.91
BL-2	90.10	92.26	83.33	91.69	88.83	90.74	83.33	92.31	95.90	95.79	**97.26**	88.17	95.83	90.48	100	94.44
BL-3	90.16	90.49	89.80	88.90	90.22	88.87	94.44	93.24	96.05	**97.18**	94.83	**89.70**	95.83	90.48	100	94.44
BL-4/Proposed	**93.06**	**94.44**	**88.89**	**95.28**	**93.04**	**94.44**	**88.89**	**95.24**	**96.37**	96.09	95.84	89.13	**95.83**	**90.48**	**100**	**94.44**
BL-5	91.03	92.37	83.33	91.89	90.28	88.93	94.44	93.27	94.34	94.63	96.05	88.93	89.55	80.95	96.30	86.12

TABLE 6. Results of 3-independent folds on 4-independent data-sets, which ensembles to total 12-independent models are tabulated below.

All Folds	Data-sets									
	BreakHik on 40× magnification					BreakHik on 200× magnification				
	Train (+Ve,-Ve)	Test (+Ve,-Ve)	Accuracy	Recall	Specificity	Train (+Ve,-Ve)	Test (+Ve,-Ve)	Accuracy	Recall	Specificity
Fold 1	(40, 18)	(18, 6)	91.67	94.44	83.33	(40, 18)	(18, 6)	83.33	83.33	83.33
Fold 2	(40, 18)	(18, 6)	91.67	94.44	83.33	(40, 18)	(18, 6)	95.83	100	83.33
Fold 3	(41, 18)	(17, 6)	95.65	94.12	100	(41, 18)	(17, 6)	100	100	100

All Folds	Datasets									
	IUPHL					UCSB cancer				
	Train (+Ve,-Ve)	Test (+Ve,-Ve)	Accuracy	Recall	Specificity	Train (+Ve,-Ve)	Test (+Ve,-Ve)	Accuracy	Recall	Specificity
Fold 1	(276, 254)	(116, 110)	96.01	95.69	96.36	(19, 23)	(7, 9)	100	100	100
Fold 2	(276, 254)	(116, 110)	95.11	96.10	95.91	(19, 23)	(7, 9)	87.50	71.43	100
Fold 3	(276, 254)	(116, 110)	97.67	98.09	95.27	(19, 23)	(7, 9)	100	100	100

issue, we incorporated an instance level arm (I_1) with the MIPool layer. Next section elaborates the significance of I_1 in the proposed architecture.

D. EFFECT OF ADDING THE INSTANCE LEVEL ARM I_1

Here we are compared the two baselines, BL-2 and BL-3 (as given in Tab. 2). As we discussed in Sec. IV-A, the MIPool layer sparsifies the gradients during training. The instance level arm (I_1) tries to make it non-sparse. It does so by associating gradients from itself for those features which do not exhibit a maximum value. These features don't contribute to the MIPool as a result of which remain at zero gradient value during back-propagation. Significant classification performance improvement was noticed (0.94-1.31% classification accuracy) between BL-2 and BL-3 in the BreakHis dataset. A similar observation was found in case of the IUPHL and UCSB cancer datasets where the classification accuracy increased by 0.12% and 0.39% respectively while using BL-3. This substantiates the importance of adding the single instance loss function to our framework.

E. EFFECT OF MULTIPLE INSTANCE LEVEL ARMS

In this section, we choose the final design of the model, through the comparison of two architectures defined as BL-3 and BL-4 (in Tab. 2). BL-3 as discussed in the previous section consists of a single instance level loss arm, whereas BL-4 (proposed final architecture) contains five instance level loss arms at multiple positions as illustrated in Fig. 2. The performance of all of these three baselines are reported in Tab. 5. Increasing the number of side level supervisory instance loss arms improves the trainability of the network but also increases its complexity. Comparing against BL-3, the classification accuracy of BL-4 improved by a range of 0.7-2.8% over all the datasets, which quantitatively substantiates its importance.

F. COMPARISON BETWEEN MARGIN LOSS AND ENTROPY LOSS FOR THE INSTANCE LEVEL CLASSIFICATION TASK

In Sec. IV-C and Sec. VI-A, we emphasized the significance of margin based instance loss over the traditional entropy loss. In order to prove this experimentally, we introduce BL-5 which uses only entropy based loss functions. Comparing against our proposed architecture (BL-4), we observe a fall in performance in the range of 0.29-6.28%. This endorses the importance of the proposed loss function, which works towards improving the classification performance while tackling the weak supervision and noisy label problems of MIL.

G. ANALYSIS OF FOLDED CROSS VALIDATIONS

In order to fully comprehend the expertise of the proposed architecture, we report its 3-fold cross validation performances on three different data-sets. In the BreakHis dataset, we have reported performances on two different magnifications (40× and 200×). Hence, we effectively had a total of four different datasets, which along with the 3 folds, altogether resulted in 12 independent training models. The complete results have been given in Tab. 6. It needs to be discussed a WSI image is typically acquired from a stained tissue slide prepared from a suspicious mass of tissue collected from a subject. It is desirable to have a very high specificity, in order to avoid the risk associated with the adverse effects of unnecessary cancer treatment, and also a very high recall to correctly detect positive cases. We have noticed a significantly high recall and specificity in all the data-sets over all the folds (as given in Tab. 6). In the BreakHis data-sets, due to the imbalance present in data, we have observed variations in performance over folds. However, performance is consistent

in the other two data-sets where the data imbalance issue is absent.

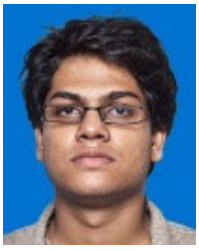
VII. CONCLUSION

In this article, we have proposed an end-to-end MIL guided learning of a CNN architecture for the WSI image classification problem. As compared to standard CNN based methods our proposed MIL based CNN architecture has no dependence on experts to mark the critical regions on a WSI image for the purpose of learning. We showed that positioning the MIPool layer at the higher dimensional feature space helps to create better bag level WSI image representations. We also showed the importance of multiple supervisory instance level classification arms. Our proposed architecture is able to expand the usability and scalability of WSI image based breast cancer detection tasks. Experimental results demonstrate the capability of the proposed algorithm over the other state-of-the-art MIL methods.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA A, cancer J. Clinicians*, vol. 70, no. 4, pp. 7–30, 2020.
[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[3] C. Li, D. Xue, Z. Hu, H. Chen, Y. Yao, Y. Zhang, M. Li, Q. Wang, and N. Xu, "A survey for breast histopathology image analysis using classical and deep neural networks," in *Proc. Int. Conf. Inf. Technol. Biomed.* Cham, Switzerland: Springer, 2019, pp. 222–233.
[4] X. Zhou, C. Li, M. M. Rahaman, Y. Yao, S. Ai, C. Sun, Q. Wang, Y. Zhang, M. Li, X. Li, T. Jiang, D. Xue, S. Qi, and Y. Teng, "A comprehensive review for breast histopathology image analysis using classical and deep neural networks," *IEEE Access*, vol. 8, pp. 90931–90956, 2020.
[5] R. Krithiga and P. Geetha, "Breast cancer detection, segmentation and classification on histopathology images analysis: A systematic review," in *Archives of Computational Methods in Engineering*. Cham, Switzerland: Springer, Aug. 2020, pp. 1–13.
[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.
[7] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.
[8] B. Babenko, "Multiple instance learning: Algorithms and applications," in *View Article PubMed/NCBI Google Scholar*. 2008, pp. 1–19.
[9] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 1998, pp. 570–576.
[10] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1073–1080.
[11] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 577–584.
[12] B. Babenko, P. Dollár, Z. Tu, and S. Belongie, "Simultaneous learning and alignment: Multi-instance and multi-pose learning," in *Proc. Workshop Faces Real-Life Images, Detection, Alignment, Recognit.*, 2008, pp. 1–15.
[13] J. Ramon and L. De Raedt, "Multi instance neural networks," in *Proc. Workshop Attribute-Value Relational Learn. (ICML)*, 2000, pp. 53–60.
[14] Z.-H. Zhou and M.-L. Zhang, "Neural networks for multi-instance learning," in *Proc. Int. Conf. Intell. Inf. Technol.*, Beijing, China, 2002, pp. 455–459.
[15] X. Xu, "Statistical learning in multiple instance problems," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 2003.
[16] O. Z. Kraus, L. Jimmy Ba, and B. Frey, "Classifying and segmenting microscopy images using convolutional multiple instance learning," 2015, *arXiv:1511.05286*. [Online]. Available: http://arxiv.org/abs/1511.05286
[17] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2424–2433.
[18] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.
[19] K. Sikka, R. Giri, and M. Bartlett, "Joint clustering and classification for multiple instance learning," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–71.
[20] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-I.I.D. samples," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 1249–1256.
[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[22] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, vol. 1, no. 2, p. 3.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[24] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3460–3469.
[25] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8290–8299.
[26] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018.
[27] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," 2018, *arXiv:1802.04712*. [Online]. Available: http://arxiv.org/abs/1802.04712
[28] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 642–651.
[29] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 914–922.
[30] B. Krishnapuram, J. Stoeckel, V. Raykar, B. Rao, P. Bamberger, E. Ratner, N. Merlet, I. Stainvas, M. Abramov, and A. Manevitch, "Multiple-instance learning improves cad detection of masses in digital mammography," in *Proc. Int. Workshop Digit. Mammography*. Berlin, Germany: Springer, 2008, pp. 350–357.
[31] P. Lu, W. Liu, W. Xu, L. Li, B. Zheng, J. Zhang, and L. Zhang, "Multi-instance learning for mass retrieval in digitized mammograms," *Med. Imag., Comput.-Aided Diagnosis, Int. Soc. Opt. Photon.*, vol. 8315, Feb. 2012, Art. no. 831523.
[32] C. Li, K. M. Lam, L. Zhang, C. Hui, and S. Zhang, "Mammogram microcalcification cluster detection by locating key instances in a multi-instance learning framework," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2012, pp. 175–179.
[33] M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: A benchmarking study," *Computerized Med. Imag. Graph.*, vol. 42, pp. 44–50, Jun. 2015.
[34] M. Kandemir, C. Zhang, and F. A. Hamprecht, "Empowering multiple instance histopathology cancer diagnosis by cell graphs," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2014, pp. 228–235.
[35] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Med. Image Anal.*, vol. 18, no. 3, pp. 591–604, Apr. 2014.
[36] Y. Xu, J. Zhang, I. Eric, C. Chang, M. Lai, and Z. Tu, "Context-constrained multiple instance learning for histopathology image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2012, pp. 623–630.
[37] Q. Liang, Y. Nan, G. Coppola, K. Zou, W. Sun, D. Zhang, Y. Wang, and G. Yu, "Weakly supervised biomedical image segmentation by reiterative learning," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1205–1214, May 2019.
[38] J. Ding, H. D. Cheng, J. Huang, J. Liu, and Y. Zhang, "Breast ultrasound image classification based on multiple-instance learning," *J. Digit. Imag.*, vol. 25, no. 5, pp. 620–627, Oct. 2012.

[39] G. Quellec, M. Lamard, M. D. Abràmoff, E. Decencière, B. Lay, A. Erginay, B. Cochener, and G. Cazuguel, "A multiple-instance learning framework for diabetic retinopathy screening," *Med. Image Anal.*, vol. 16, no. 6, pp. 1228–1240, Aug. 2012.
[40] S. Manivannan, C. Cobb, S. Burgess, and E. Trucco, "Subcategory classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification," *IEEE Trans. Med. Imag.*, vol. 36, no. 5, pp. 1140–1150, May 2017.
[41] A. Madooei, M. S. Drew, and H. Hajimirsadeghi, "Learning to detect blue–white structures in dermoscopy images with weak supervision," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 779–786, Mar. 2019.
[42] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. N. Metaxas, and X. S. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1332–1343, May 2016.
[43] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 603–611.
[44] K. Das, S. Conjeti, A. G. Roy, J. Chatterjee, and D. Sheet, "Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 578–581.
[45] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016.
[46] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 7, pp. 1977–1984, Jul. 2011.
[47] M. M. Dundar, S. Badve, V. C. Raykar, R. K. Jain, O. Sertel, and M. N. Gurca, "A multiple instance learning approach toward optimal classification of pathology slides," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2732–2735.
[48] A. Ruiz, X. Binefa, and J. Van de Weijer, "Regularized multi-concept MIL for weakly-supervised facial behavior categorization," in *Proc. Brit. Mach. Vis. Conf.*, 2014, p. 8.
[49] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit.*, vol. 77, pp. 329–353, May 2018.
[50] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: A lazy learning approach," in *Proc. 17th Int. Conf. Mach. Learn.* San Francisco, CA, USA: Morgan Kaufmann, 2000, pp. 1119–1125.
[51] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[52] V. Cheplygina, D. M. J. Tax, and M. Loog, "Multiple instance learning with bag dissimilarities," *Pattern Recognit.*, vol. 48, no. 1, pp. 264–275, Jan. 2015.



KAUSIK DAS was born in India, in 1989. He received the B.Tech. degree in electronics and communication engineering from the Kalyani Government Engineering College, India, in 2011, and the M.Tech. degree in signal processing from the Indian Institute of Technology Guwahati, India, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Indian Institute of Technology Kharagpur, India. His research interests include machine learning, deep learning, and medical image processing.

SAILESH CONJETI (Member, IEEE) was born in Chennai, India, in 1990. He received the B.E. degree (Hons.) in electrical and electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 2012, the M.Tech. degree in medical imaging and informatics from the Indian Institute of Technology Kharagpur, India, in 2014, and the Ph.D. degree in computer aided medical procedures from the Technishe Universität Munchen, Germany. He is currently a Product Manager with the AI Products-Imaging Decision Support, Siemens Healthineers, Germany. His research interests include computational medical imaging, machine learning, medical image computing, and biomedical signal processing.

JYOTIRMOY CHATTERJEE (Member, IEEE) received the M.Sc. degree in cytology and molecular genetics and the Ph.D. degree in radiation biology from The University of Burdwan, India, in 1984 and 1990, respectively. He is currently an Assistant Professor with the Indian Institute of Technology Kharagpur, India, where he is also an Investigator of five sponsored research projects. He has authored more than 20 articles and inventor of five patents. His current research interests include medical imaging, image analysis, and system biology for regenerative medicine, cancer diagnosis, and radiation toxicology. He is a member of the Engineering in Medicine and Biology Society. He received the Senior Research Fellowship from the Indian Council of Medical Research for the Postdoctoral Research from 1992 to 1995.

DEBDOOT SHEET (Senior Member, IEEE) was born in Kharagpur, India, in 1986. He received the B.Tech. degree in electronics and communication engineering from the West Bengal University of Technology, Kolkata, India, in 2008, and the M.S. and Ph.D. degrees in medical imaging from the Indian Institute of Technology Kharagpur, India, in 2010 and 2014, respectively. He was a DAAD Visiting Scholar with the Chair for Computer Aided Medical Procedures, Technische Universität Munchen, Germany, from 2011 to 2012. He is currently an Assistant Professor of electrical engineering with the Indian Institute of Technology Kharagpur. His research interests include computational medical imaging, machine learning, image and multidimensional signal processing, and visualization. He has been serving as the Regional Editor for the IEEE Pulse since 2014.