Expert Systems With Applications 117 (2019) 103–111

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multiple instance learning for histopathological breast cancer image classification

P.J. Sudharshan[a], Caroline Petitjean[b,*], Fabio Spanhol[c], Luiz Eduardo Oliveira[c], Laurent Heutte[b], Paul Honeine[b]

[a] Indian Institute of Information Technology - D&M, Jabalpur, India
[b] Université de Rouen, France
[c] Department of Informatics (DInf), Federal University of Paraná, Curitiba, PR, Brazil

ARTICLE INFO

Article history:
Received 5 June 2018
Revised 31 August 2018
Accepted 23 September 2018
Available online 24 September 2018

Keywords:
Biomedical image processing
Breast cancer
Histopathology
Image classification
Multiple Instance Learning

ABSTRACT

Histopathological images are the gold standard for breast cancer diagnosis. During examination several dozens of them are acquired for a single patient. Conventional, image-based classification systems make the assumption that all the patient's images have the same label as the patient, which is rarely verified in practice since labeling the data is expensive. We propose a weakly supervised learning framework and investigate the relevance of Multiple Instance Learning (MIL) for computer-aided diagnosis of breast cancer patients, based on the analysis of histopathological images. Multiple instance learning consists in organizing instances (images) into bags (patients), without the need to label all the instances. We compare several state-of-the-art MIL methods including the pioneering ones (APR, Diverse Density, MI-SVM, citation-kNN), and more recent ones such as a non parametric method and a deep learning based approach (MIL-CNN). The experiments are conducted on the public BreaKHis dataset which contains about 8000 microscopic biopsy images of benign and malignant breast tumors, originating from 82 patients. Among the MIL methods the non-parametric approach has the best overall results, and in some cases allows to obtain classification rates never reached by conventional (single instance) classification frameworks. The comparison between MIL and single instance classification reveals the relevance of the MIL paradigm for the task at hand. In particular, the MIL allows to obtain comparable or better results than conventional (single instance) classification without the need to label all the images.

1. Introduction

Supervised learning is a subfield of machine learning where a predictive function is inferred from a set of labeled training examples, in order to map each input instance to its output label. In a conventional setting, the training dataset consists of instances equipped with their corresponding label. While instances are relatively easy to obtain, the expensive data-labeling process with human-based ground-truth descriptions remains the major bottleneck to have large-scale datasets. This issue gives rise to a novel paradigm in machine learning, with the so-called weakly supervised learning, namely when having a partially-labeled training dataset (Zhou, 2017).

Multiple Instance Learning (MIL) provides an elegant framework to deal with weakly supervised learning. In comparison with strong (*i.e.*, fully-labeled) supervised learning where every training instance is assigned a discrete or real-valued label, the rationale of MIL paradigm is that instances are naturally grouped in labeled bags, without the need that all the instances of each bag have individual labels (Fig 1). In the binary classification case, a bag is labeled positive if it has at least one positive instance; on the other hand, a bag is labeled negative if all its instances are negative (Foulds & Frank, 2010). With such training data grouped in labeled bags, MIL algorithms seek to classify either unseen bags (*i.e.*, bag-level classification) or unseen instances (*i.e.*, instance-level classification).

Whereas MIL has many applications in medical imaging, as shown in a recent review (Quellec, Cazuguel, Cochener, & Lamard, 2017), there is a growing interest for the usage of MIL for histopathological image classification (Jia, Huang, Chang, & Xu, 2017a; Mercan et al., 2018; Xu et al., 2014), in particular. Histopathological images are microscopic images of the tissue for disease examination, which prevail as the gold standard for cancer diagnosis (Rubin, Strayer, Rubin, & McDonald, 2008). Estab-

* Corresponding author.
E-mail address: caroline.petitjean@univ-rouen.fr (C. Petitjean).

https://doi.org/10.1016/j.eswa.2018.09.049
0957-4174/

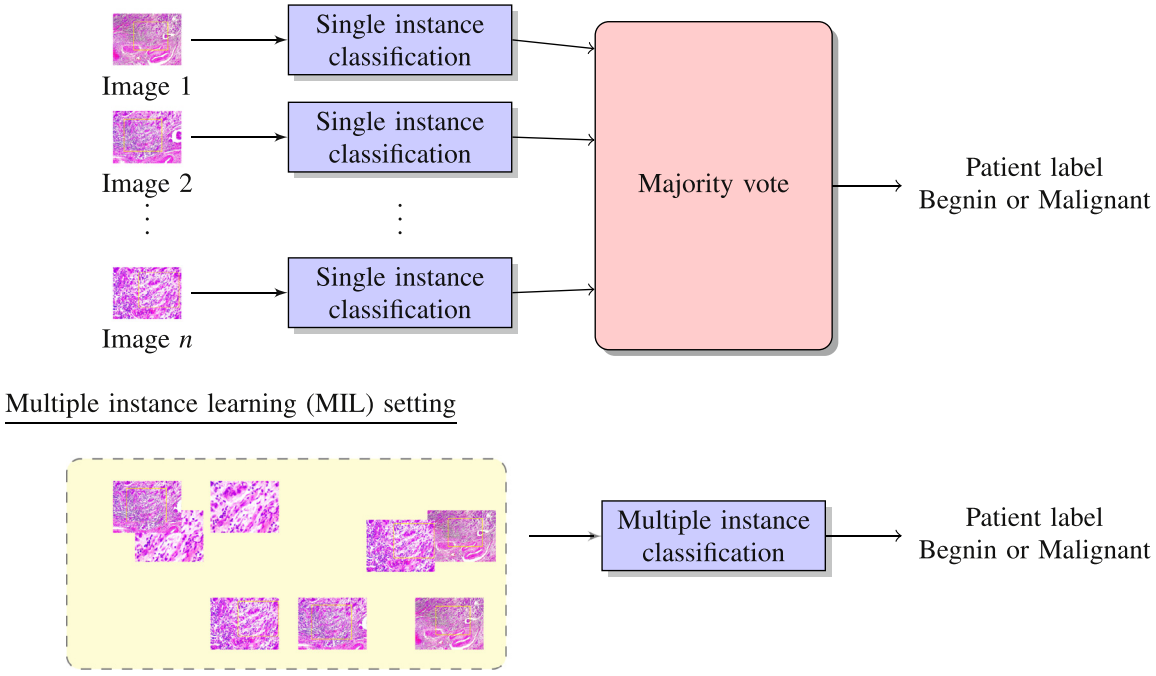

Fig. 1. Multiple instance learning vs single instance classification. In this figure, an instance is an image, and the bag is the patient. One can also consider the case where an instance is a patch, the bag is the image.

lishing a diagnosis with histopathological images remains a non trivial task. The expert has to identify specific patterns (size, geometry, texture). However the analysis is a highly, time consuming specialized task, dependent on the experience of the pathologists and influenced by factors such as fatigue and decrease of attention. As pointed out by (Gurcan et al., 2009) there is a pressing need for computer-assisted diagnosis (CAD) to relieve the workload on pathologists, by for example filtering obviously benign areas, so that the experts can focus on the more difficult-to-diagnose cases.

The diagnosis is established by the pathologist at the patient scale, from reviewing several images. Another decision can also be taken at the scale of the image: the experts can attribute a label to an image by identifying patterns in particular areas in the images. However fine-grain annotation would be too costly to perform; thus datasets are labeled at the patient or image scale, but not at the pixel or region scale. As these data are weakly labeled by essence, the MIL paradigm should be better suited than single instance classification. To verify this hypothesis, we provide in our paper results obtained on a significant, recently established, publicly available dataset of breast cancer images, called BreakHis (Spanhol, Oliveira, Petitjean, & Heutte, 2016b), which contains about 8000 microscopic biopsy images of benign and malignant breast tumors, originating from 82 patients.

We thus propose to investigate the use of MIL on this large dataset of histopathological images, with the aim of image classification and patient, as a first contribution. We have chosen a representative sample of major MIL methods: the seminal Axis-Parallel Rectangle algorithm (APR) (Dietterich, Lathrop, & Lozano-Prez, 1997), and algorithms based on diversity density (DD) (Maron & Lozano-Pérez, 1998; Zhang & Goldman, 2001), k-NN (Citation-kNN) (Wang & Zucker, 2000) and Support Vector Machines (SVM) (Andrews, Tsochantaridis, & Hofmann, 2002), as well as a recently-proposed non-parametric algorithm (Venkatesan, Chandakkar, & Li, 2015) and a deep learning approach revisiting Convolutional Neural Networks (CNN) for MIL (MILCNN) (Sun, Han, Liu, & Khodayari-Rostamabad, 2016). The second contribution is to show that MIL compares favorably to single instance classification, which is the only framework implemented on this data until now. Of course in this case we suppose that instances inherit labels from the bags. We examine whether it is preferable to cast this problem into a single instance one, or if MIL does indeed bring an added value, both at the image and patient levels (Alpaydin, Cheplygina, Loog, & Tax, 2015).

The remainder of this paper is organized as follows. Section 2 motivates the use of the MIL paradigm for histopathological image and patient diagnosis. Section 3 presents the MIL assumption and provides a brief survey of MIL methods. Section 4 describes the BreaKHis dataset and the conducted experiments with the obtained results. Section 5 concludes the paper.

2. The MIL paradigm for histopathological image analysis

While the multiple instance paradigm arose in many domains prior to the 1990's, MIL was first described explicitly and studied by (Dietterich et al., 1997). The original motivation in MIL was drug activity prediction, where experts provide activity labels to bags of molecules, labeling each individual molecule being costly and hard to set up. MIL is central in many applications in various domains, such as in bioinformatics, text processing, computer vision and image processing, to name a few (Herrera et al., 2016). Indeed, in many applications, ground truth labeling is expensive in general and instances can be often grouped in bags, each bag having a set of partially-labeled instances. An example is facial recognition, where several images of the same person taken from different angles can be considered as instances in a bag (the bag being the person) (Herrera et al., 2016). Note that only the MIL paradigm can apprehend this type of situations.

Of particular interest is image-based pathology classification for medical decision making, since it is relatively easy and part of the clinical protocol to take many images of some organs or tissues (physiology) under study; on the other hand, labeling each image is a time-consuming process dominated by human effort. The relevance of MIL for this type of weakly labeled data, as announced in introduction, is twofold.

The first possibility is to divide each image into subimages or patches and to consider the image as a bag, while patches are the instances. As histopathological images have a high resolution, local methods are appealing to process and identify malignant areas. In the field of natural scene images, this approach is related to region-based image categorization, where each instance encodes color, textural or spatial features related to that specific region (Herrera et al., 2016). In our binary setting, the image would be labeled "positive" (pathological) if it has at least one malignant patch; conversely, an image would be labeled benign if it does not have any portion labeled malignant. This multiple instance formalism is natural, since only a subset of the patches is labeled by the experts, making it possible that entire images might be healthy whereas the patient is diagnosed with a tumor. This is not the case in the conventional strategy used so far, in a single instance classification setting with instances inheriting the label of their image.

The second possibility explores the patient scale. Each patient can thus be described with several dozens of images. Conventional, image-based classification systems classify images independently and merge decisions at the patient level; they make the assumption that all the patient's images have the same label as the patient, which is rarely verified in practice. With MIL, the patient is considered as a bag, with the instances being its associated images or subimages. This makes full sense as the diagnosis (*i.e.,* the label) is established only at the patient level. Furthermore, a patient diagnosed with a malignant tumor can still have some of his images described as tumor-free, *i.e.,* healthy, as just said; and a healthy patient has inevitably all of his images healthy. These facts match the MIL assumption.

In our experiments we will investigate these two settings, and compare both of them to their single instance learning equivalent.

3. MIL methods: A brief overview

Under the standard MIL assumption, positive bags contain at least one positive instance, while negative bags contain only negative instances. We denote by L_B the label of a bag B, defined as a set of instances, each one described by its feature vector: $B = \{b_1, b_2, \ldots, b_N\}$. We denote by l_k the label of each instance b_k. We can now define the label of a bag, following the standard MIL assumption:

$$L_B = \begin{cases} +1 & \text{if } \exists l_k : l_k = +1; \\ -1 & \text{if } \forall l_k : l_k = -1. \end{cases} \quad (1)$$

There are other – more relaxed – assumptions, such as a bag is labeled positive when it contains a sufficient number of positive instances; since they are out of the scope of this paper, we refer the reader to (Foulds & Frank, 2010) for further reading.

MIL methods are usually divided into two groups, depending on how they exploit the information in the data (Amores, 2013). The first group consists of methods that consider the discriminative information at the instance level. Learning algorithms do not focus at the larger scale of a bag, but at the local scale of instances. An advantage of these methods is that they are able to classify instances, when needed. However, they require that instances have a precise label, a requirement not all MIL problems meet. The instance level methods include APR, DD, SVM based approaches. The second group consists of the methods that consider the discriminative information to be at the bag level. These methods are usually more accurate, since they can model the distribution of each class and the relations between classes (Carbonneau, Cheplygina, Granger, & Gagnon, 2016). However, they cannot classify single instances, but only bags. An example of such methods is Citation-kNN (Wang & Zucker, 2000). For a review on MIL methods, we refer the reader to (Amores, 2013; Carbonneau et al., 2016; Herrera et al., 2016).

In the following, we briefly describe the well-established MIL methods that have been implemented and applied to the BreaKHis dataset.

3.1. Axis-parallel hyper rectangle (APR)

The MIL paradigm was first introduced in the seminal work of (Dietterich et al., 1997), motivated mainly by an application in biochemistry. The goal was to predict whether a molecule will be binding to a given receptor or not. Each molecule, which can be considered as a bag, can take many different spatial conformations, namely the instances. The methodology to solve the MIL problem is to design an hyper rectangle (called axis-parallel hyper rectangle (APR)) in the feature space aimed at containing at least one positive instance from each positive bag while excluding all the instances from negative bags. A molecule is classified as positive (resp. negative) if one (resp. none) of its instances belongs inside the APR.

3.2. Diverse Density (DD) and its variants

Diverse density (Maron & Lozano-Pérez, 1998) is closely related to the idea of the APR. The DD defines a function over the feature space, such that it is high at points that are both close to instances from positive bags, and far away from instances which are in negative bags. The DD algorithm attempts to find the local maxima of this function (called the positive instance targets or prototypes) by maximizing diverse density (i.e. conditional likelihood) over the instance space, using gradient ascent with multiple starting points. The DD approach has given rise to many variants, the most known is the Expectation-Maximization method (EM–DD) (Zhang & Goldman, 2001). In this variant, the DD measure is maximized iteratively with the EM algorithm.

3.3. Citation-KNN

The Citation-kNN, an adaptation of k-nearest neighbors (k-NN) algorithm, is the first non-parametric approach (Wang & Zucker, 2000). The principle is to first apply the k-NN algorithm to bags, where the distance between bags is measured with the minimum Hausdorff distance. The latter is defined as the shortest distance between any two instances from each bag:

$$\text{Dist}(A, B) = \min_{a_i \in A} \min_{b_j \in B} ||a_i - b_j||$$

for any two bags A and B, where a_i and b_j are instances from each bag. This distance is used by a k-NN to classify a new bag, as the regular k-NN approach. The citation-kNN method adds a final step that makes the process more robust: in addition to the neighbors of the bag to be classified, other bags, called *citers*, are also considered in the classification rule.

3.4. Mi-SVM and MI-SVM

Two alternative generalizations of the maximum margin idea used in SVM classification have been proposed by Andrews et al. (2002). On one hand, the mi-SVM is based on the instance-level paradigm. Since the instance labels are not known, they are treated as hidden variables subject to constraints defined by their bag labels. The mi-SVM method attempts to recover the instance labels and, at the same time, to find the optimal discriminant function. On the other hand, the bag-level paradigm is adopted by the MI-SVM. Its goal is to maximize the bag margin, defined between the positive instances of the positive bags, and the negative instances of the negative bags. In this setting, the bag is not represented by all its instances, but only by the "extreme"

ones, in the same sense as *support vectors* in conventional SVM. Moreover, mi-SVM and MI-SVM inherit also the kernel trick, thus allowing to use linear, polynomial and RBF kernels.

3.5. Non-parametric MIL

This recent technique is designed as a modified version of the *k*-NN classifier (Venkatesan et al., 2015). The non-parametric MIL approach employs a new formulation based on distances to k-nearest neighbors. The idea is to parse the MIL feature space with a Parzen window technique, using different sized regions. Conversely to the majority vote used in *k*-NN, the vote contributions are the kernelized distances in the feature space. Non-parametric MIL has shown enhanced robustness to labeling noise on various datasets.

3.6. MILCNN

Deep learning networks have been overwhelming machine learning, pattern recognition and computer vision fields for a few years. MIL is no exception to this rule (Hoffman, Wang, Yu, & Darrell, 2016; Jia et al., 2017a; Kraus, Ba, & Frey, 2016; Pathak, Shelhamer, Long, & Darrell, 2014; Sun et al., 2016; Wang, Yan, Tang, Bai, & Liu, 2018; Zhou, Zhao, Yang, Yu, & Xu, 2017). In Sun et al. (2016), a Multiple Instance Learning Convolutional Neural Network (MIL-CNN) is proposed. This framework was initially proposed for the data augmentation problem: in object detection, labels are not always preserved when the images are split for data augmentation. The proposed method considers data augmentation generated images as a bag, by combining a convolutional neural network (CNN) with a specific MIL loss function derived with respect to the bag.

4. Experiments and results

4.1. Description of the breakhis dataset

BreaKHis is a publicly available dataset of microscopic biopsy images of benign and malignant breast tumors (Spanhol et al., 2016b). The images were collected through a clinical study in 2014, to which all patients referred to the P&D Laboratory (Brazil) with a clinical indication of breast cancer were invited to participate. The institutional review board approved the study and all patients provided their written informed consent. All the data were anonymized. Samples were generated from the breast tissue biopsy slides, stained with hematoxylin and eosin (HE). The samples were collected by surgical open biopsy (SOB), prepared for histological study and labeled by pathologists of the P&D Lab. The diagnosis of each case was produced by experienced pathologists and confirmed by complementary exams such as immunohistochemistry analysis.

Images were acquired in RGB color space, with a resolution of 752 × 582 using magnifying factors of 40×, 100×, 200× and 400×. Fig. 2 shows these 4 magnifying factors on a single image. This image is acquired from a single slide of breast tissue containing a malignant tumor (breast cancer). The highlighted rectangle (manually added for illustrative purposes only) is the area of interest selected by the pathologist to be detailed in the next higher magnification. To date, the database is composed of 7909 images divided into benign and malignant tumors. Table 1 summarizes the image distribution. For more information about the dataset, we refer to Spanhol et al. (2016b).

4.2. Experimental protocol

Following the standard labeling convention in use in medical studies, the label “positive” (resp. “negative”) refers to malignant (resp. benign) images. The BreaKHis dataset has been randomly divided into a training set (70%) and a testing set (30%), following the protocol described in Spanhol, Oliveira, Petitjean, and Heutte (2016a). Patients used to build the training set are not used for the testing set. We computed the average rate over five trials. Note that the distribution of the samples for each trial is publicly available and allows for a fair comparison of methods. To handle the image high resolution (752 × 582) and to augment data for training, images were divided into patches. The patches will form the instances, whereas bags will be considered at two levels: at the patient level, patches will be collected independently from all the patient’s images; at the image level, patches will be originating from the image of interest.

A size of 64 × 64 pixel was chosen for the patch size, as it has been shown to be particularly relevant for CNN-based classification (Spanhol et al., 2016a). For training, 1000 patches are randomly extracted from each input image. For test, to preserve computational cost, a grid of non-overlapping patches is extracted, yielding around 100 patches per image. Each patch is described with a 162-long feature vector of Parameter-Free Threshold Adjacency Statistics (PFTAS) features (Coelho et al., 2010; Hamilton, S Pantelic, Hanson, & Teasdale, 2007). These features have shown to be particularly relevant for this dataset, when assessed against many others such as local binary patterns (LBP), completed LBP, local phase quantization, gray-level co-occurrence matrices, as well as computer vision features such as ORB (oriented FAST and rotated BRIEF) (Spanhol et al., 2016b).

Twelve MIL methods were evaluated on the BreaKHis dataset, as described in Section 2: APR, DD and EM-DD, citation-kNN, mi-SVM and MI-SVM, both with linear, polynomial and RBF kernels,

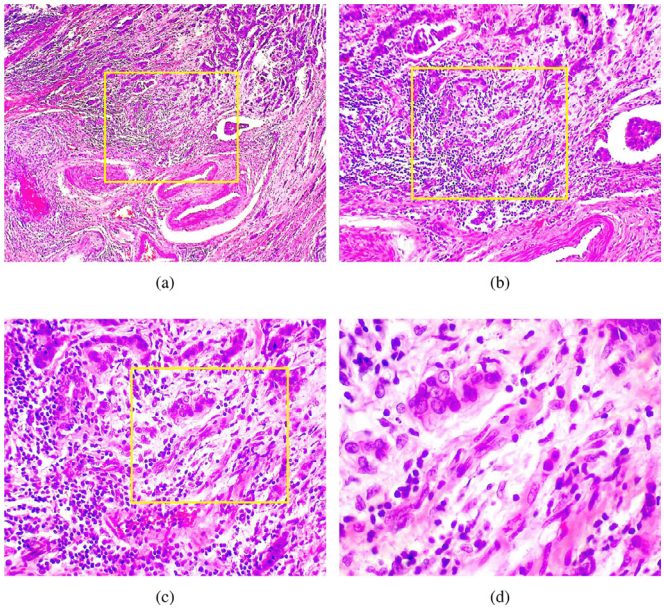

Fig. 2. A slide of breast malignant tumor seen in different magnification factors of the same image: (a) 40×, (b) 100×, (c) 200×, and (d) 400 × .

Table 1
Image distribution by magnification factor and class.

Magnification	Benign	Malignant	Total
40×	625	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Total	2480	5429	7909
# Patients	24	58	82

Table 2
Accuracy rate at respective levels. Best results columnwise are in bold. For statistical significance, please see text.

	Patient as bag				Image as bag			
	40×	100×	200×	400×	40×	100×	200×	400×
Iterated-discrim APR	73.8 ± 3.8	66.5 ± 4.1	84.2 ± 4.9	68.0 ± 5.6	70.4 ± 2.4	65.1 ± 5.0	81.3 ± 5.5	67.3 ± 4.9
DD	70.5 ± 6.1	64.5 ± 4.3	68.3 ± 3.6	71.2 ± 3.3	71.2 ± 5.9	66.1 ± 5.4	66.7 ± 2.9	70.8 ± 3.8
EM-DD	78.3 ± 5.6	80.6 ± 5.2	77.1 ± 6.3	78.7 ± 5.7	73.1 ± 5.4	76.4 ± 4.8	78.2 ± 5.2	76.2 ± 5.6
Citation-kNN	73.7 ± 4.6	72.8 ± 5.4	75.7 ± 3.1	77.2 ± 3.6	73.1 ± 4.3	73.0 ± 5.7	71.3 ± 3.5	78.7 ± 3.1
mi-SVM Linear	79.5 ± 4.3	83.4 ± 4.6	83.6 ± 4.7	81.0 ± 5.2	72.6 ± 4.4	80.6 ± 3.7	80.1 ± 4.9	78.2 ± 5.3
mi-SVM poly	75.2 ± 6.1	79.8 ± 4.8	76.5 ± 3.9	68.5 ± 5.1	75.6 ± 5.7	78.7 ± 4.0	75.2 ± 5.6	69.2 ± 4.8
mi-SVM RBF	77.8 ± 1.6	75.4 ± 1.5	73.8 ± 2.3	72.9 ± 3.4	77.9 ± 2.2	77.3 ± 2.1	74.6 ± 2.9	71.4 ± 3.9
MI-SVM Linear	85.6 ± 5.6	82.1 ± 5.9	84.6 ± 4.8	80.9 ± 4.9	79.5 ± 4.1	78.2 ± 4.4	80.8 ± 4.7	78.9 ± 5.1
MI-SVM poly	84.8 ± 2.7	82.5 ± 4.6	83.9 ± 4.2	81.3 ± 4.2	86.2 ± 2.8	82.8 ± 4.8	**81.7 ± 4.4**	82.7 ± 3.8
MI-SVM RBF	79.0 ± 2.1	71.9 ± 2.9	76.2 ± 1.9	73.0 ± 3.5	78.3 ± 3.2	72.2 ± 3.0	76.8 ± 1.6	71.9 ± 2.4
Non-parametric	**92.1 ± 5.9**	**89.1 ± 5.2**	**87.2 ± 4.3**	82.7 ± 3.0	**87.8 ± 5.6**	**85.6 ± 4.3**	80.8 ± 2.8	**82.9 ± 4.1**
MILCNN	86.9 ± 5.4	85.7 ± 4.8	85.9 ± 3.9	**83.4 ± 5.3**	86.1 ± 4.2	83.8 ± 3.1	80.2 ± 2.6	80.6 ± 4.6

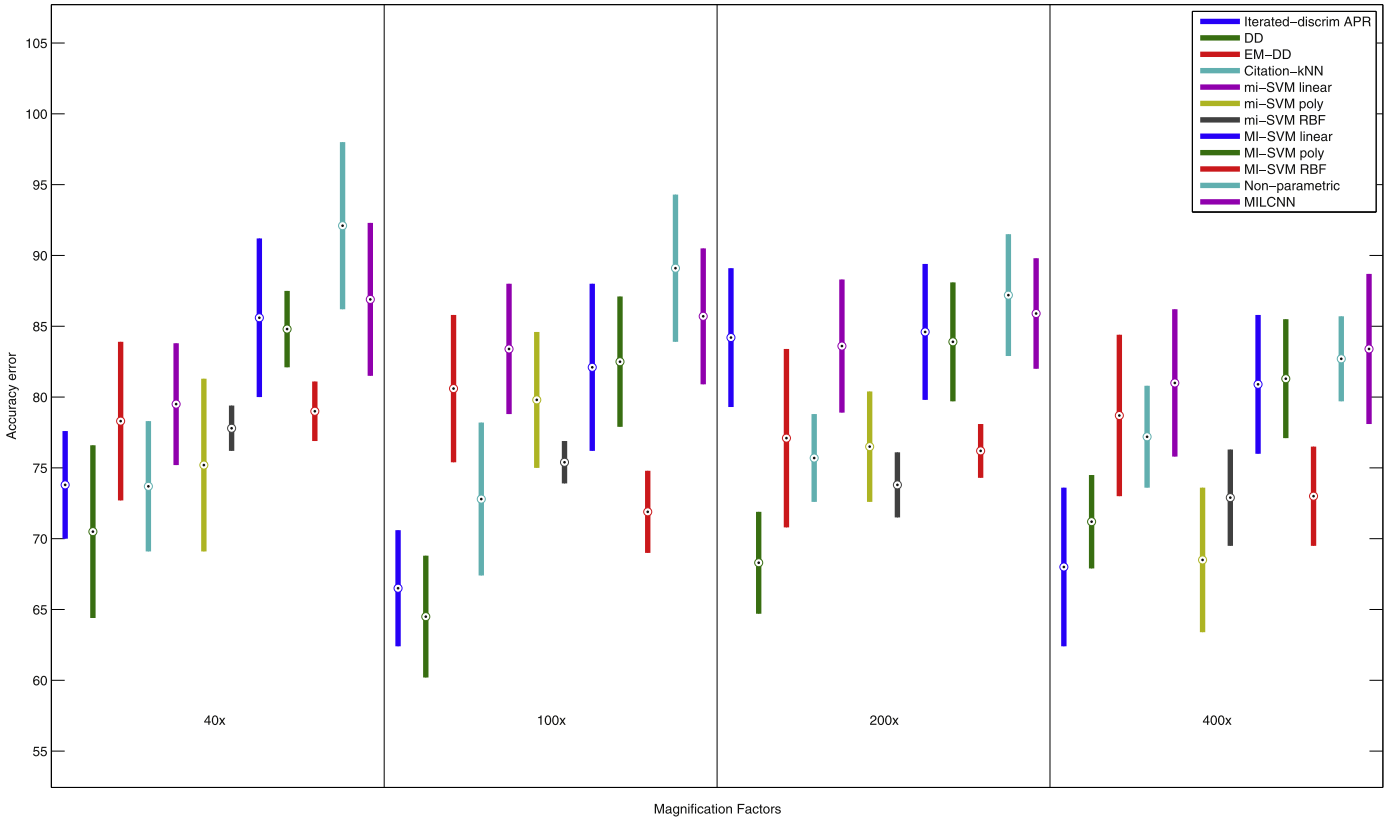

Fig. 3. Accuracy results of MIL benchmark with patient as bag (left part of Table 2). Best viewed in color.

non-parametric MIL, and MILCNN. For all methods except the non-parametric and the MILCNN, we used the implementation of the J. Yang's MIL Library[1] with MATLAB 2017a. The non-parametric MIL algorithm was obtained from the author's website[2]. For the implementation of MILCNN in Python, Keras and Theano were used (Chollet, 2015). The hyper-parameters for each method were optimized using grid search (cf Appendix A). Regarding the non-parametric MIL method, the main (and almost only) hyperparameter is the number of neighbors. It is found with grid search and typically depends on the dataset. The other hyperparameters are related to implementation and train/test setting; they include the maximum number of iterations to maximize the training accuracy and the number of runs on which accuracy is averaged. These hyper parameters should be set as high as possible, or at least as the result of a tradeoff between computation time and accuracy, and do not have a tremendous influence on the results.

In the following, we first show the benchmark of MIL methods, and then assess the best MIL method against single instance classification frameworks.

4.3. Results

4.3.1. MIL Benchmark on breakhis dataset

We provide results for two different settings, as aforementioned. In the first setting, each patient is considered as a bag, which is labeled with its diagnosis, and the instances are the patches extracted from the images. As can be seen in Table 1, in average 25 images are available for each patient, for each magnification factor. Since around 100 patches are extracted per image in test, each bag (or patient) contains around 2500 instances. In the

[1] CMU MIL toolbox: http://www.cs.cmu.edu/~juny/MILL/.
[2] https://github.com/ragavvenkatesan/np-mil.

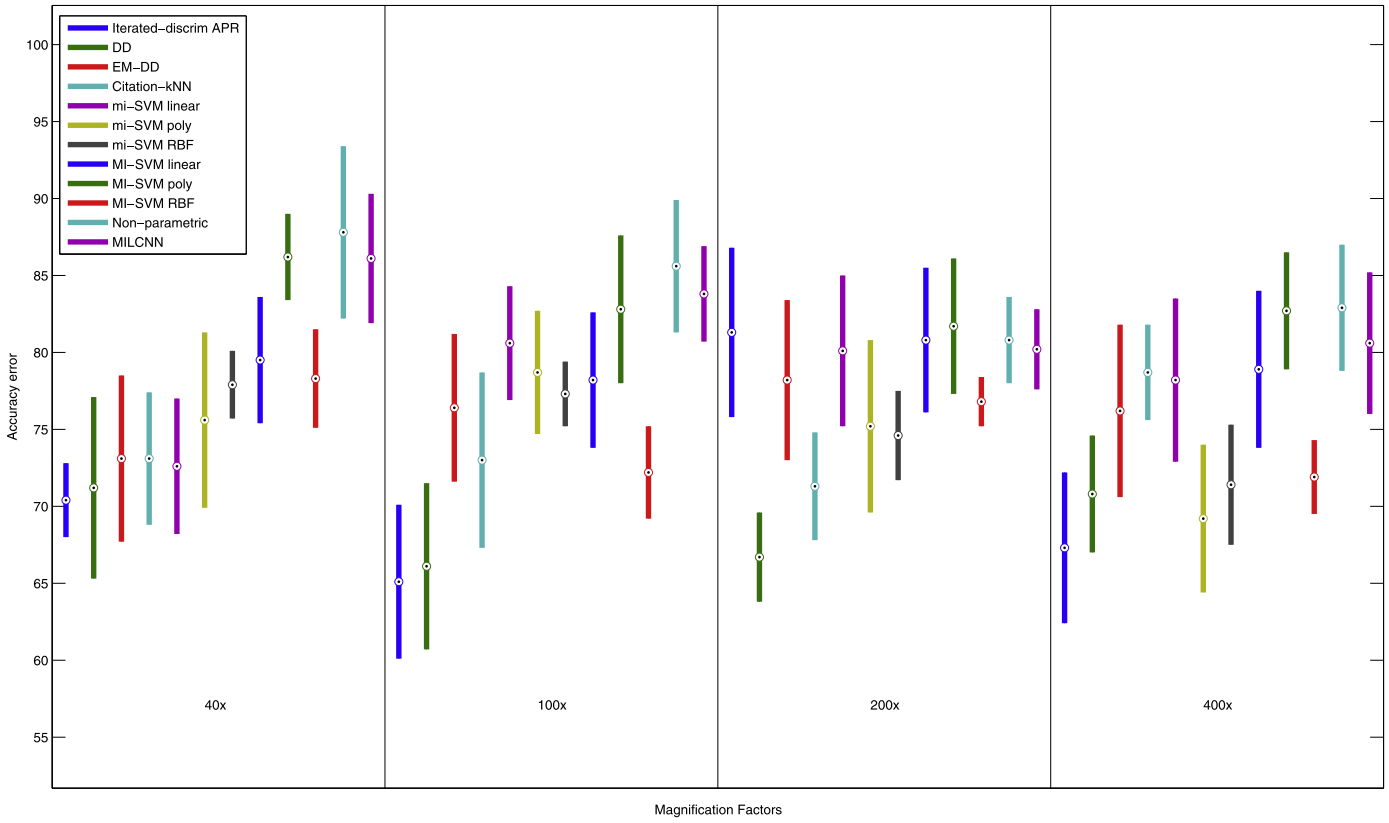

Fig. 4. Accuracy results of MIL benchmark with image as bag (right part of Table 2). Best viewed in color.

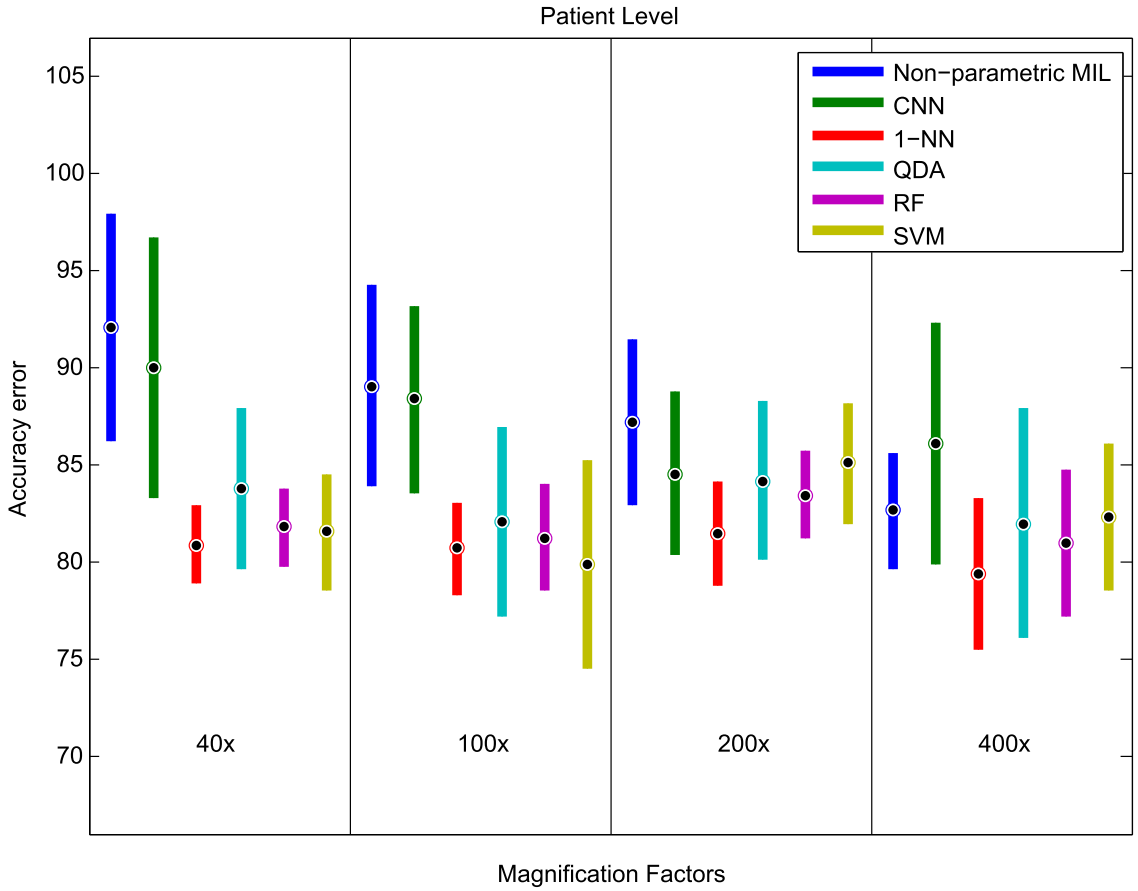

Fig. 5. Accuracy results: MIL vs SIL at patient level (left part from Table 3). Best viewed in color.

Table 3
Comparison of MIL (non-parametric) vs single instance classification (SIL). Best results columnwise are in bold. For statistical significance, please see text.

		Patient as bag (MIL) or level (SIL)				Image as bag (MIL) or level (SIL)			
		40×	100×	200×	400×	40×	100×	200×	400×
MIL	Non-parametric	**92.1 ± 5.9**	**89.1 ± 5.2**	**87.2 ± 4.3**	82.7 ± 3.0	**87.8 ± 5.6**	**85.6 ± 4.3**	80.8 ± 2.8	**82.9 ± 4.1**
SIL	CNN	90.0 ± 6.7	88.4 ± 4.8	84.6 ± 4.2	**86.1 ± 6.2**	85.6 ± 4.8	83.5 ± 3.9	83.1 ± 1.9	80.8 ± 3.0
	1-NN	80.9 ± 2.0	80.7 ± 2.4	81.5 ± 2.7	79.4 ± 3.9	79.1 ± 2.1	77.8 ± 3.0	79.6 ± 1.9	77.6 ± 4.0
	QDA	83.8 ± 4.1	82.1 ± 4.9	84.2 ± 4.1	82.0 ± 5.9	82.8 ± 3.6	80.7 ± 4.9	83.3 ± 3.0	80.5 ± 5.6
	RF	81.8 ± 2.0	81.3 ± 2.8	83.5 ± 2.3	81.0 ± 3.8	80.2 ± 1.9	80.4 ± 3.8	82.4 ± 2.3	80.0 ± 4.5
	SVM	81.6 ± 3.0	79.9 ± 5.4	85.1 ± 3.1	82.3 ± 3.8	79.9 ± 3.7	77.1 ± 5.5	**84.2 ± 1.6**	81.2 ± 3.6

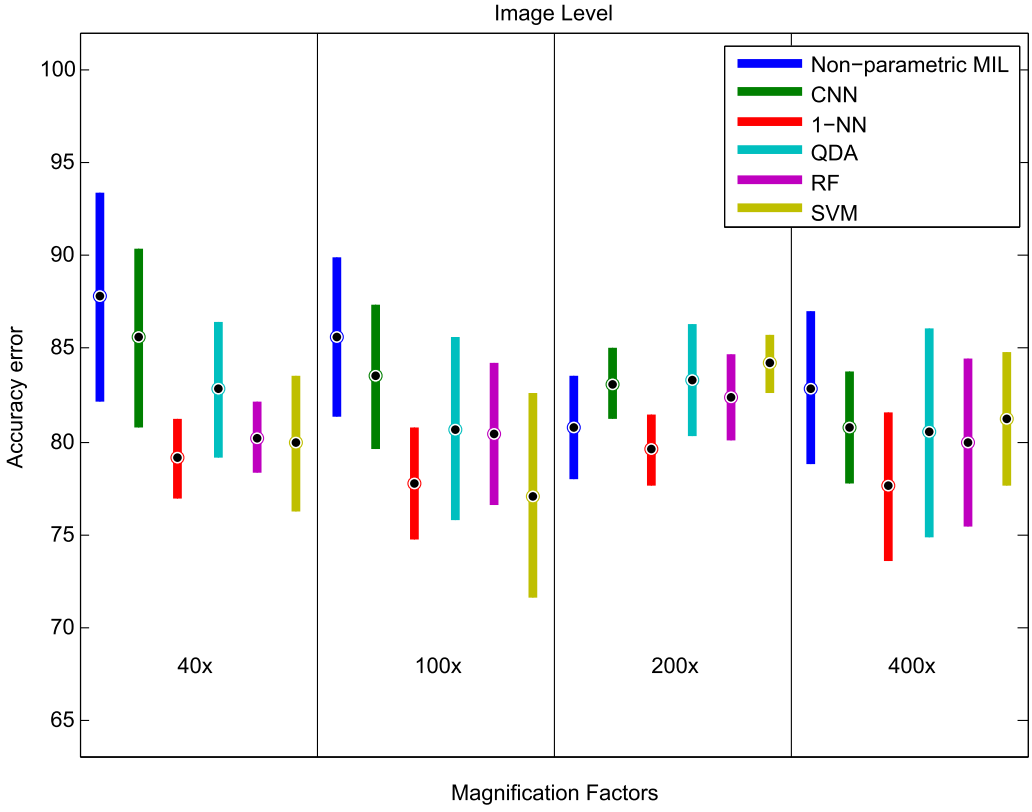

Fig. 6. Accuracy results: MIL vs SIL at image level (right part from Table 3). Best viewed in color.

second setting, we consider each image as a bag; in this case, the instances are the patches, and a bag contains approximately 100 instances.

Results are presented in Table 2, and Figs. 3 and 4. Due to the large standard deviation values, we performed paired *T*-tests on the accuracy rates on all trials, in order to fairly compare the methods and considered the difference to be significant when the *p*-value is inferior to 0.05. As expected, DD-based approaches, APR and citation-kNN yield the poorest results which leads us to think that positive instances are not clustered in a single area of the feature space. SVM-based approaches perform better. In particular, MI-SVM leads to enhanced results, compared to mi-SVM, with a *p*-value of 0.0029, which shows that a bag level paradigm is better suited to the data. At last, best classification rates are obtained with the non-parametric MIL and the MILCNN approaches, which both outperform all the other methods, with *p*-values inferior to 0.001 (when comparing to either MILCNN or non-parametric MIL); however there is no significant difference between MILCNN and non-parametric MIL ($p = 0.77$). To compare MIL against single instance classification methods in the following section, we arbitrarily retain the non-parametric MIL as the best MIL method.

4.3.2. MIL Vs single instance learning

Using the same protocol (same trials, same training and test set distribution), we compare results of non-parametric MIL with state-of-the-art, single instance classifiers, namely 1-NN, quadratic discriminant analysis (QDA), random forest (RF), and SVM, partly obtained from previous experiments (Spanhol et al., 2016b). Hyper-parameters of these classifiers were tuned using grid search and only the best results were retained. These classifiers take as input the PFTAS feature vector describing each image. For the CNN approach, we used AlexNet (Krizhevsky, Sutskever, & Hinton, 2012; Spanhol et al., 2016a). Decisions are taken on each patch and are fused together using the Max Fusion Rule. For further details, we refer the reader to Spanhol et al. (2016a,b).

Results are reported in Table 3, Fig. 5 (patient level) and Fig. 6 (image level). Apart from the single instance CNN, we can observe that the non-parametric MIL, trained using the same PFTAS feature vector, is better than all the other four SIL methods; *p*-values for all methods 1-NN, QDA, RF, SVM against MIL are inferior to 0.0073. This suggests that instances, namely patches, may provide only partial, complementary information for the image or the patient level (Alpaydin et al., 2015), and that a bag-based analysis is valuable for the analysis of histopathological images.

Now, the single instance CNN is not only better than the other single instance learning models trained with hand-crafted textual descriptors (in accordance with previous experiments (Spanhol et al., 2016a) and (Han et al., 2017)), its performance are similar to the non-parametric MIL approach ($p = 0.65$). Even if both methods are not significantly different in terms of accuracy, the non-parametric MIL still has the advantage of not requiring the labeling at the image level, but only at a macro (i.e. patient) level: MIL can leverage weakly labeled data, an advantage in the field of medical imaging where fine-grain labeling is especially costly.

5. Conclusions and future works

Multiple instance learning provides a classification framework that is particularly adapted to computer-aided diagnosis based on histopathological image analysis. In the case of the BreaKHis dataset, several hundreds of images are available per patient. The patient can thus be considered as a bag, which is labeled with its diagnosis.

Our MIL benchmark shows that the recently proposed non-parametric MIL and MILCNN are particularly efficient for the tasks of patient and image classification. Patient classification rates can reach up to 92.1% for the 40 × magnification factor, a level never reached by conventional classification frameworks, which enhances the fact that instances are complementary and can be fruitfully considered in a MIL framework. MIL can thus leverage digital histopathological image classification and analysis to improve computer-aided diagnosis, without the need to label all the images.

As future work, we are currently engaged in experimenting other deep learning frameworks (Spanhol, Cavalin, Oliveira, Petitjean, & Heutte, 2017). With the acceleration of proposals in this area, no doubt that some more efficient networks will be proposed in the near future. Today's usage of CNN and more generally AI or learning based technology is often limited to assist the clinician for the final decision. In addition to improving the accuracy of the decision making process, research should also focus on extracting those features that matter for the cancer image classification. These features will give insight on specific areas to be examined, and the experts will be able to focus on these areas (Xu et al., 2017). By considering the image as a bag and pixels as the instances, MIL offers an adequate framework for histopathological image segmentation, to identify malignant region position (Jia, Huang, Chang, & Xu, 2017b; Kraus et al., 2016; Pathak et al., 2014; Xu et al., 2014).

Acknowledgment

The authors acknowledge the CRIANN (Centre des Ressources Informatiques et Applications Numériques de Normandie, France) for providing computational resources. Fabio Spanhol was supported for this work by a grant of the Programa de Doutorado Sanduíche no Exterior 88881.135972/2016-01 agency (Coordination for the Improvement of Higher Education Personnel, Brazil).

Appendix A. Method hyper-parameterization

For APR (Dietterich et al., 1997):

- Kernel Width: 0.999
- Outside Probability: 0.023
- GridNum: 25000

For DD and EM-DD (Maron & Lozano-Pérez, 1998):

- Scaling: 1
- Aggregate: average
- Threshold: 0.5
- No. of runs: 100 (DD), 500 (EM-DD)
- Iteration Tolerance (for EM-DD): 0.08

For Citation-kNN (Wang & Zucker, 2000):

- Bag Distance Type: minimum
- Instance Distance Type: Euclidean
- Reference nodes considered: 5
- CiterRank: 11

For mi-SVM and MI-SVM (Andrews et al., 2002):

- Kernel: Linear, poly, RBF
- KernelParam - NA/degree/gamma: (NA), 4, 0.32 (mi-SVM), (NA), 5, 0.17 (MI-SVM)
- CostFactor: 1/0.96/1 (mi-SVM), 1/1/1 (MI-SVM)
- NegativeWeight: 1/1/1
- Threshold: 0.5/0.5/0.5

For non-parametric MIL (Venkatesan et al., 2015):

- Averaged accuracy over 100 runs
- Range of k: 50 (using elbow method)
- Maximum No. of iterations to maximize the training accuracy: 3000
- Distance Method: Euclidean

For MILCNN (Sun et al., 2016) the structure is the same as the one for CIFAR10 / CIFAR100 with same values of the parameters.

References

Alpaydin, E., Cheplygina, V., Loog, M., & Tax, D. M. J. (2015). Single- vs. multiple-instance classification. *Pattern Recognition, 48*(9), 2831–2838.

Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence, 201*, 81–105.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2002). Support vector machines for multiple-instance learning. In *Proceedings of the 15th international conference on neural information processing systems*. In *NIPS'02* (pp. 577–584). Cambridge, MA, USA: MIT Press.

Carbonneau, M.-A., Cheplygina, V., Granger, E., & Gagnon, G. (2016). Multiple instance learning: A survey of problem characteristics and applications. *CoRR, abs/1612.03365*.

Chollet, F. (2015). Keras: Theano-based deep learning library. Code: https://github.com/fchollet. Documentation: http://keras.io.

Coelho, L. P., Ahmed, A., Arnold, A., Kangas, J., Sheikh, A.-S., Xing, E. P., et al. (2010). Structured literature image finder: Extracting information from text and images in biomedical literature. In C. Blaschke, & H. Shatkay (Eds.), *Linking literature, information, and knowledge for biology: workshop of the biolink special interest group, ISMB/ECCB 2009, Stockholm, June 28–29, 2009, revised selected papers* (pp. 23–32)). Springer.

Dietterich, T. G., Lathrop, R. H., & Lozano-Prez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial. Intelligence, 89*(1), 31–71.

Foulds, J., & Frank, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review, 25*(1), 1–?25.

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering, 2*, 147–171.

Hamilton, N., S Pantelic, R., Hanson, K., & Teasdale, R. (2007). Fast automated cell phenotype classification. *BMC Bioinformatics, 8*, 110.

Han, Z., Wei1, B., Zheng, Y., Yin, Y., Li, K., & Li, S. (2017). Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific Reports, 7*(4172).

Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Sánchez-Tarragó, D., & Vluymans, S. (2016). Multiple instance learning. In *Multiple instance learning* (pp. 17–33). Springer.

Hoffman, J., Wang, D., Yu, F., & Darrell, T. (2016). Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*.

Jia, Z., Huang, X., Chang, E. I. C., & Xu, Y. (2017a). Constrained deep weak supervision for histopathology image segmentation. *IEEE Transactions on Medical Imaging, 36*(11), 2376–2388.

Jia, Z., Huang, X., Chang, E. I. C., & Xu, Y. (2017b). Constrained deep weak supervision for histopathology image segmentation. *IEEE Transactions on Medical Imaging, 36*(11), 2376–2388.

Kraus, O. Z., Ba, J. L., & Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics, 32*(12), i52–i59.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks* (pp. 1106–1114).

Maron, O., & Lozano-Pérez, T. (1998). A framework for multiple-instance learning. In *Proceedings of the 1997 conference on advances in neural information processing systems 10*. In *NIPS '97* (pp. 570–576).

Mercan, C., Aksoy, S., Mercan, E., Shapiro, L. G., Weaver, D. L., & Elmore, J. G. (2018). Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE Transactions on Medical Imaging, 37*(1), 316–325.

Pathak, D., Shelhamer, E., Long, J., & Darrell, T. (2014). Fully convolutional multi-class multiple instance learning. *CoRR, abs/1412.7144*.

Quellec, G., Cazuguel, G., Cochener, B., & Lamard, M. (2017). Multiple-instance learning for medical image and video analysis. *IEEE Reviews in Biomedical Engineering, 10*, 213–234.

Rubin, R., Strayer, D., Rubin, E., & McDonald, J. (2008). *Rubin's pathology: Clinicopathologic foundations of medicine*. Lippincott Williams & Wilkins.

Spanhol, F., Cavalin, P., Oliveira, L., Petitjean, C., & Heutte, L. (2017). Deep features for breast cancer histopathological image classification. *IEEE international conference on systems, man, and cybernetics, Banff, Canada*.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016a). Breast cancer histopathological image classification using convolutional neural networks. In *International joint conference on neural networks* (pp. 2560–2567).

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016b). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering, 63*(7), 1455–1462.

Sun, M., Han, T. X., Liu, M.-C., & Khodayari-Rostamabad, A. (2016). Multiple instance learning convolutional neural networks for object recognition. In *International conference on pattern recognition (ICPR)* (pp. 3270–3275).

Venkatesan, R., Chandakkar, P. S., & Li, B. (2015). Simpler non-parametric methods provide as good or better results to multiple-instance learning. In *2015 ieee international conference on computer vision (ICCV)* (pp. 2605–2613).

Wang, J., & Zucker, J.-D. (2000). Solving the multiple-instance problem: A lazy learning approach. In *Proceedings of the seventeenth international conference on machine learning*. In *ICML* (pp. 1119–1126).

Wang, X., Yan, Y., Tang, P., Bai, X., & Liu, W. (2018). Revisiting multiple instance neural networks. *Pattern Recognition, 74*, 15–24.

Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., et al. (2017). Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics, 18*(1), 281.

Xu, Y., Zhu, J.-Y., Eric, I., Chang, C., Lai, M., & Tu, Z. (2014). Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis, 18*(3), 591–604.

Zhang, Q., & Goldman, S. A. (2001). EM-DD: An improved multiple-instance learning technique. In *advances in neural information processing systems* (pp. 1073–1080). MIT Press.

Zhou, L., Zhao, Y., Yang, J., Yu, Q., & Xu, X. (2017). Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images. *IET Image Processing*.

Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning. *National Science Review, 00*, 1–10.