Computer-aided diagnosis from weak supervision: A benchmarking study

Melih Kandemir[1], Fred A. Hamprecht[1]

Abstract

Supervised machine learning is a powerful tool frequently used in computer-aided diagnosis (CAD) applications. The bottleneck of this technique is its demand for fine grained expert annotations, which are tedious for medical image analysis applications. Furthermore, information is typically localized in diagnostic images, which makes representation of an entire image by a single feature set problematic. The multiple instance learning framework serves as a remedy to these two problems by allowing labels to be provided for *groups* of observations, called *bags*, and assuming the group label to be the maximum of the instance labels within the bag. This setup can effectively be applied to CAD by splitting a given diagnostic image into a Cartesian grid, treating each grid element (patch) as an instance by representing it with a feature set, and grouping instances belonging to the same image into a bag. We quantify the power of existing multiple instance learning methods by evaluating their performance on two distinct CAD applications: i) Barrett's cancer diagnosis, and ii) diabetic retinopathy screening. In the experiments, mi-Graph appears as the best-performing method in bag-level prediction (i.e. diagnosis) for both of these applications that have drastically different visual characteristics. For instance-level prediction (i.e. disease localization), mi-SVM ranks as the most accurate method.

Keywords: Multiple instance learning, cancer diagnosis, diabetic retinopathy screening

1. Introduction

Advances in image analysis and machine learning gradually make available more robust algorithms for extracting information from data. An appealing application at the intersection of these two disciplines is computer-aided diagnosis which aims to automate disease diagnosis from images [27]. CAD tools have been shown to be useful to aid the pathologist by pointing out important regions in large biopsy tissue images [11], providing decision support by calculating informative metrics such as cell counting [18], and quantifying the disease risk [21].

A major drawback of many CAD algorithms is their demand for fine-grained expert annotations during training. For tumor diagnosis, pathologists need to indicate the tumor regions, and for diabetic retinopathy, small structures such as microaneurysms have to be annonated by opthalmologists. The use of weakly supervised machine learning techniques can drastically reduce the annotation effort, while keeping prediction performance at an acceptable level.

A common characteristic of diagnostic imaging is the locality of discriminative information. For instance, in cancer histology, a small region within a large slide often determines the final grading, and all the remaining slide is redundant. Similarly, in diabetic retinopathy screening, small structures, such as microaneurysms are much richer in diagnostic information than the texture of the entire image. Hence, application of the standard supervised learning setup to these cases would be problematic. Given a diagnostic image, representing it by a single feature vector would require tedious feature engineering, since when standard feature sets are applied, the uninformative areas in the image would overrule the informative ones. On the other hand, dividing the image into small patches, and representing each patch by a feature vector would result in severe class imbalance.

Multiple instance learning (MIL) [19] provides a learning framework that both allows weak supervision and inherently handles the locality of information problem. In MIL, ground-truth labels are available only for groups of observations, called *bags*. A bag with a positive label indicates that there exists at least one observation within that bag, whose label is positive. For a negatively labeled bag, on the other hand, all observations are known to have a negative label. This framework can directly be applied to CAD by defining each diagnostic image (tissue slide or fundus image) as a bag, and each of its regions (e.g. patches in a Cartesian grid) as an instance. Diseased cases with local lesions are then represented by a positive bag, and healthy cases by a negative bag.

Even though some previous work reports MIL solutions tailored to specific CAD problems [21, 28, 29], the utility of a large set of existing MIL approaches in these applications has not yet been evaluated. Furthermore and more importantly, the generalizability of their success on various CAD problems has not yet been quantified. In this paper, we address these two issues by providing a benchmarking study using a large list of MIL methods [2] on two CAD applications that have clearly distinct visual characteristics: i) diagnosis of Barrett's cancer from H & E stained histology images, and ii) diabetic retinopathy

Email addresses: melih.kandemir@iwr.uni-heidelberg.de (Melih Kandemir), fred.hamprecht@uni-heidelberg.de (Fred A. Hamprecht)

[1]Heidelberg University, HCI, Speyerer Str. 6, D-69115, Heidelberg, Germany

[2]The source code of the MIL methods in our comparison list is available under: http://hci.iwr.uni-heidelberg.de/Staff/mkandemi/MILBundle.tar.gz

screening from eye fundus images. Among the methods under comparison, mi-Graph [33] outperforms the others in both applications in cancer diagnosis (i.e. prediction of bag labels). On the other hand, in the harder problem of cancer localization (i.e. prediction of instances), mi-SVM [2] gives the highest generalization performance.

2. Prior Art

2.1. Cancer diagnosis from histology images

There has been a large volume of studies on application of machine learning methods to histology cancer diagnosis (see [11] for a comprehensive review). Demir et al.[10] propose the classification of brain tumors by constructing graphs from cell topology, and representing the tumor image by a set of graph features. Doyle et al.[8] classify prostate cancer grades from graph-based (e.g. minimum spanning tree of cells), morphological (e.g. nuclear density), and textural features (e.g. Gabor filter responses) using the standard multiclass support vector machine (SVM). Alternatively, Huang et al.[12] show that differential box counting leads to effective prostate cancer grading. Wang [25] demonstrate the successful application of Markov random fields to segmentation of lung tumors. Hang et al.[5] propose a method that combines sparse coding and multiscale histogram intersection kernels for diagnosis of kidney renal carcinoma and glioblastoma. Kandemir et al.[13] perform diagnosis of Barrett's cancer using mi-Graph.

2.2. Automated diabetic retinopathy screening

Agurto et al.[1] introduce an automated diabetic retinopathy screening method that characterizes the texture of regions of interest by their amplitude and frequency properties. Giancardo et al.[9] detect microaneurysms from morphological heuristics, and then apply a standard SVM to predict the disease status. Quellec et al.[22] introduce a content-based image retrieval (CBIR) system for diabetes detection by formulating a probabilistic interpretation of a set of wavelets. In a follow-up study, Quellec et al.[21] improve the state-of-the-art in diabetes detection by extending their CBIR method with multiscale features.

2.3. MIL for computer-aided diagnosis

MIL has comparatively recently started to be used for computer aided diagnosis. Some examplary studies are as follows. Zhao et al.[32] apply the MILES [6] method to patches of slides of 10 different tissue types. Zhang et al.[30] use GPMIL of Kim et al.[15] for classification of skin biopsies. Xu et al.[28, 29] use a multiclass extension of MILBoost [24] for grading of prostate tumors. Quellec et al.[21] build their aforementioned multiscale CBIR method for the MIL setup.

3. The diagnosis pipeline

We use the same automated diagnosis pipeline for both applications. We split a given diagnostic image into a regular grid of patches. We then construct an instance from each patch by extracting a set of features. A group of instances belonging to the same diagnostic image is treated as a bag. The label of the bag is assumed to be $+1$ if it includes the target disease, and -1 otherwise. Consequently, we predict the disease status of a given image (bag) using one of the MIL methods in comparison. Figure 1 illustrates the pipeline.

For both applications, we represent an image patch with a set of intensity histogram and texture features as listed in Table 1. For Barrett's cancer diagnosis, we additionally use a set of cell features. We segment cells using supervised pixel classification and watershed transform as described in [13]. We then extract a set of intensity and morphology features from each cell (see Table 2 for the complete list). Finally, we augment the feature vector of each patch by a set of summary statistics of features of cells lying within that patch, as listed in Table 3.

Table 1: Features extracted from patches of Barrett's cancer histology and fundus images.

	Color features
1	Intensity histogram of RGB channels for 26 bins
	Texture features
2	Mean of local binary pattern histograms of 20x20-pixel grids
3	Mean of SIFT descriptors
4	Box count for grid sizes 2,3,...,8

Table 2: Features extracted from each segmented cell.

1	Central power sums for exponents 1,2,3 and 4,
2	Area, radius, perimeter, and roundness of the segment,
3	Maximum, mean, and minimum intensity, and intensity covariance, variance, skewness, and kurtosis within the region and within its 30-pixel-wide belt for each color channel,
4	Region axes, principal axes, kurtosis, minimum, maximum, and power sums for exponents 1,2,3,4

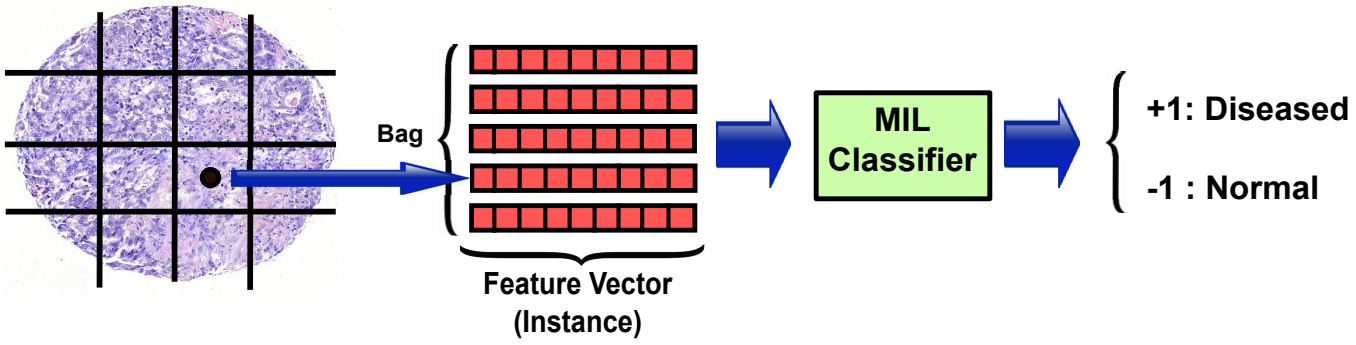
Table 3: Features extracted from cells located within each Barrett's cancer image patch.

Minimum, maximum, mean, standard deviation, skewness, and kurtosis of features (given in Table 2) of all healthy and cancer cells in a patch

4. Multiple instance learning methods

Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ be a data set consisting of N instances, each of which is a D-dimensional feature vector: $\mathbf{x}_i = [x_i^{(1)}, \cdots, x_i^{(D)}]$. The data set is assumed to be partitioned into B bags: $\mathbf{X} = \bigcup_{b=1}^{B} \mathbf{X}_b$, such that $\mathbf{X}_b \bigcap \mathbf{X}_c = \emptyset, \quad \forall b \neq c$, where each bag b consists of N_b instances: $\mathbf{X}_b = [\mathbf{x}_{b1}, \cdots, \mathbf{x}_{bN_b}]$. Let $\mathbf{Y} = [Y_1, \cdots, Y_B]$ be the vector of the corresponding binary bag labels $Y_b \in \{-1, +1\}$. Labels of instances are collected into the vector $\mathbf{y} = [y_1, \cdots, y_N]$, which follows the same partitioning

Figure 1: The block diagram of the used computer-aided diagnosis pipeline. The diagnostic image is first split into rectangular patches. A set of features are extracted from each patch, and an instance is formed by the resultant feature vector. Patches belonging to the same diagnostic image are grouped into a bag. The bag is then classified into *diseased* or *healthy* by a multiple instance learning method.



as instances $\mathbf{y} = \bigcup_{b=1}^{B} \mathbf{y}_b$, such that $\mathbf{y}_b \bigcap \mathbf{y}_c = \emptyset, \quad \forall b \neq c$, where $\mathbf{y}_b = [y_{b1}, \cdots, y_{bN_b}]$. Let $\mathcal{B}^+ = \{b|Y_b = +1\}$ and $\mathcal{B}^- = \{b|Y_b = -1\}$ denote sets of positive and negative bags, respectively. Finally, $\mathbb{I}(\cdot)$ denotes the indicator function that gives 1 if its argument is true, and 0 otherwise.

The central assumption of the MIL setup is that the label of a bag is the maximum of the labels of the instances in that bag: $Y_b = \max(\mathbf{y}_b)$, which we call as the *multiple instance constraint*. A negative bag label implies that all instances within the bag have a negative label. On the other hand, for a positively labeled bag, only existence of *some* positive instances within that bag is known, but the actual instance labels are latent.

The MIL methods included in our benchmarking list are briefly explained below.

4.1. mi-Graph [33]

This simple but effective method represents each bag by a similarity graph. First, the cross-similarities of bag instances are calculated by an instance-level kernel function $k_{inst}(\mathbf{x}_i, \mathbf{x}_j)$. A graph is then constructed by placing a node per each instance within a bag and each node pair is connected by an edge if the two corresponding instances are more similar to each other than a threshold δ. Let $\mathbf{W}_b$ be the affinity matrix of bag b, whose entry is $w^b_{nm} = 1$ if there is an edge between the nodes of instances n and m, and $w^b_{nm} = 0$ otherwise. Consequently, similarity between bags b and c are calculated by the following kernel function

$$k_{bag}(\mathbf{X}_b, \mathbf{X}_c) = \frac{\sum_{n=1}^{N_b} \sum_{m=1}^{N_c} v_{bn} v_{cm} k_{inst}(\mathbf{x}_{bn}, \mathbf{x}_{cm})}{\sum_{n=1}^{N_b} v_{bn} \sum_{m=1}^{N_c} v_{cm}}$$

where $v_{bn} = 1/\sum_{u=1}^{N_b} w^b_{nu}$, $v_{cm} = 1/\sum_{u=1}^{N_c} w^c_{mu}$ are the sum of the weights of the edges incident to nodes (instances) n and m of bags b and c, respectively. An arbitrary kernel learner is then trained on the resultant bag-level Gram matrix.

The intuition behind this kernel is that for instances that are similar to a large number of other instances within the bag, W_{ia} has a smaller value, and for instances different from the rest of the bag, W_{ia} is large. Hence, the influence of odd instances within bags are enhanced, and others are downweighted.

4.2. Gaussian process multiple instance learning (GPMIL) [15]

This method extends the standard Gaussian process classifier to MIL by modifying the sigmoid likelihood to a form that obeys the multiple-instance constraint. Each data point $\mathbf{x}_i$ is assigned a latent decision output variable f_i whose sign indicates the label of the instance and magnitude the decision margin. These decision outputs follow a Gaussian process prior

$$\mathbf{f}|\mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}\Big(\mathbf{f}\Big|\mathbf{0}, \mathbf{K}(\mathbf{X}, \boldsymbol{\theta})\Big)$$

where $\mathbf{f} = [f_1, \cdots, f_N]$ is the vector of decision outputs that shares the same bag partitioning as the data $\mathbf{f} = \bigcup_{b=1}^{B} \mathbf{f}_b$ with $\mathbf{f}_b = [f_{b1}, \cdots, f_{bN_b}]$, and $\mathbf{K}(\mathbf{X}, \boldsymbol{\theta})$ is the Gram matrix calculated by applying a kernel function $k(\cdot, \cdot|\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$ to every pair of instances in $\mathbf{X}$. The sigmoid term used for instance class likelihood in the standard Gaussian process is replaced by the following bag class likelihood that satisfies the multiple instance constraint

$$p(Y_b|\mathbf{f}_b) = \frac{1}{1 + \exp(-Y_b \max(\mathbf{f}_b))}. \tag{1}$$

To make inference tractable, the $\max(\mathbf{f}_b)$ term is replaced by soft-max $\log \sum_{n=1}^{N_b} \exp(f^b_{bn})$, which leads to

$$p(Y_b|\mathbf{f}_b) = \frac{1}{1 + \left(\sum_{b=1}^{N_b} e^{f_{bn}}\right)^{-Y_b}}, \qquad \forall b$$

Following the Bayesian paradigm, the latent $\mathbf{f}$ vector is inferred by posterior estimation. Since the non-conjugate likelihood in Equation 1 does not allow the posterior distribution to be found in closed form, inference is performed via the Laplace approximation

$$p(\mathbf{f}|\mathbf{X}, \mathbf{Y}) \simeq \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \mathbf{H}^{-1}),$$

where $\hat{\mathbf{f}}$ is the estimated mode of the posterior and $\mathbf{H}$ is the negative Hessian of the logarithm of the posterior at its mode. The posterior mode can be estimated using gradient search.

4.3. MILBoost [24]

Boosting is a generic ensemble learning framework, where the idea is to learn a (usually) linear combination of multiple *weak* classifiers (i.e. logistic regression). Let $\mathbf{f} = \{f_1(\mathbf{x}, y), \cdots, f_T(\mathbf{x}, y)\}$ be a set of the decision functions of T weak classifiers for a given instance $\mathbf{x}$ and ground-truth label y, each of which is parameterized by θ_t. AnyBoost [20] learns a vector of classifier weights $\mathbf{w} = [w_1, \cdots, w_T]$ that (typically) combines the decision outputs of weak classifiers linearly

$$p(y_i|\mathbf{w}, \mathbf{x}_i, z_i) = z_i \sum_{t=1}^{T} \Big(w_t p(y_i|f_t, \mathbf{x}_i) \Big)$$

where

$$p(y_i|f_t, \mathbf{x}_i) = \frac{1}{1 + \exp\left(-f_t(\mathbf{x}_i, y_i)\right)} \tag{2}$$

is the probability of instance $\mathbf{x}_i$ to belong to class y_i, and z_i is the weight of instance i indicating its importance in classification. Within this framework, the data likelihood reads

$$\mathcal{L}(\mathbf{X}, \mathbf{y}, \theta, \mathbf{w}, \mathbf{z}) = \sum_{i=1}^{N} \Bigg(\mathbb{I}(y_i = +1) \log p(y_i|\mathbf{w}, \mathbf{x}_i, z_i) + \mathbb{I}(y_i = -1)(1 - \log p(y_i|\mathbf{w}, \mathbf{x}_i, z_i)) \Bigg).$$

In each iteration, the algorithm updates classifier weights $\mathbf{w}$ and instance weights $\mathbf{z}$ using gradient search,

$$\mathbf{z} \leftarrow \mathbf{z} - \alpha \frac{\partial \mathcal{L}(\mathbf{X}, \mathbf{y}, \theta, \mathbf{w}, \mathbf{z})}{\partial \mathbf{z}},$$
$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial \mathcal{L}(\mathbf{X}, \mathbf{y}, \theta, \mathbf{w}, \mathbf{z})}{\partial \mathbf{w}},$$

where α is the step size. The classifier parameters θ are then updated by retraining them with the new values of $\mathbf{w}$ and $\mathbf{z}$.

MILBoost [24] extends the AnyBoost method by replacing the instance-level class-conditional probability (Eq. 2) with Noisy-OR bag-level probability

$$p(Y_b = +1|f_t, \mathbf{X}_b) = 1 - \prod_{n=1}^{N_b} \Big(1 - p(y_{bn} = +1|f_t, \mathbf{x}_{bn})\Big),$$

where $p(y_{bn}|f_t, \mathbf{x}_{bn})$ is modeled by passing the classifier response f_t through a sigmoid function as in Eq. 2. Extension of MILBoost to multi-instance multi-class classification is shown to be successful in prostate cancer grading [28, 29].

4.4. MI-SVM [2]

MI-SVM modifies the standard SVM optimization problem from instance level to bag level: It has one slack variable and one constraint per bag. The large margin constraint in the standard SVM is also replaced by the multiple instance constraint in MI-SVM. The resultant optimization problem is

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}^2\| + C \sum_{i=1}^{N} \xi_b,$$
$$s.t. \ \ \forall b : Y_b \max_{\mathbf{x}_{bn} \in \mathbf{X}_b} (\mathbf{w}^T \phi(\mathbf{x}_{bn})) \geq 1 - \xi_b, \qquad \xi_b \geq 0,$$

where $\mathbf{w}$ is the vector of model parameters defining the planar decision boundary, C is the regularization constant, ξ_b are slack variables, and $\phi(\cdot)$ is a function that maps an instance from the original feature space to a Reproducing Kernel Hilbert Space (RKHS)[23]. This problem can be solved approximately by a two-step iterative algorithm. In each iteration, first, a standard SVM is trained on a set of representative instances. The learned classifier is then used to choose the representative instance set for the next iteration. The most representative instance of a bag in this context is the one having the highest probability of being positive.

4.5. mi-SVM [2]

This method approaches MIL as a semi-supervised learning problem, treating the labels of positive bag instances as latent variables. These latent variables are added to the optimization problem and inferred from data

$$\min_{\mathbf{y}} \ \min_{\mathbf{w}, b, \boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}^2\| + C \sum_{i=1}^{N} \xi_i,$$
$$s.t. \ \ y_i(\mathbf{w}^T \phi(\mathbf{x}_i)) \geq 1 - \xi_i, \qquad \forall i,$$
$$\xi_i \geq 0, \qquad \forall i,$$
$$\max(\mathbf{y}_b) = Y_b, \qquad \forall b.$$

At each iteration, the approximate solution can be found as follows: trains an instance-level standard SVM based on the current assignments of the latent variables, then update these variables by making predictions with the learned SVM.

4.6. Citation kNN [26]

This method extends the well-known k-nearest neighbors algorithm to the MIL setup by defining a bag similarity metric which is a robust variant of Hausdorff distance.

4.7. EMDD [31]

This method fits a Gaussian density kernel to the positive instances. In particular, it learns a hypothesis point $\mathbf{h}$ corresponding to the mean of the Gaussian (i.e. the centroid of the cloud of positive instances), and the standard deviation s_d of each data dimension d,

$$P(y_{bn} = +1|\mathbf{h}, \mathbf{x}_{bn}) = \exp\left(- \sum_{p=1}^{D} s_d^2 x_{bn}^{(p)^2} \right).$$

Similarly to MI-SVM, this method follows a two-step iterative algorithm that resembles *expectation maximization (EM)*. In the

E-step, one representative instance is chosen from each bag that has the largest probability among the instances within the bag

$$\mathbf{x}_b^* = \underset{\mathbf{x}_{bn} \in b}{\text{argmax}} \; P(y_{bn} = +1 | \mathbf{h}, \mathbf{x}_{bn}), \quad \forall b.$$

In the M-step, the kernel parameters $\mathbf{h}$ and $\mathbf{s} = [s_1, \cdots, s_D]$ are fit to the chosen instance set

$$\mathbf{h} = \underset{\mathbf{h}}{\text{argmax}} \; \prod_{b=1}^{B} \left(1 - \left| \mathbb{I}(Y_b = +1) - P(Y_b = +1 | \mathbf{h}, \mathbf{x}_b^*) \right| \right).$$

This optimization problem can simply be solved using gradient search.

4.8. Bag Key Instance SVM (B-KI-SVM) [17]

Differently from the methods above that do bag-level prediction, Liu et al.[17] introduce an MIL algorithm specifically tailored for instance-level prediction, called *Key Instance SVM (KI-SVM)*. Its central assumption is that there exists strictly one positive instance, called *key instance*, within each positive bag. This assumption is incorporated into the large margin SVM formulation by assigning each positive bag instance a latent variable $d_{bn} \in \{0, 1\}$, and imposing the key instance assumption as a constraint

$$\begin{aligned}
\min_{\mathbf{w},\rho,\boldsymbol{\xi},\mathbf{d}} \; & \frac{1}{2}\|\mathbf{w}^2\| + \frac{C}{2} \sum_{b \in \mathcal{B}^+} \xi_b^2 + \frac{\lambda C}{2} \sum_{b \in \mathcal{B}^-} \xi_b^2 \\
s.t. \; & \mathbf{w}^T \sum_{n=1}^{N_b} d_{bn} \phi(\mathbf{x}_{bn}) \geq \rho - \xi_b, \quad b \in \mathcal{B}^+ \\
& \sum_{n}^{N_b} d_{bn} = 1, \qquad b \in \mathcal{B}^+ \\
& \underbrace{-\mathbf{w}^T \sum_{n=1}^{N_b} \frac{\phi(\mathbf{x}_{bn})}{N_b} \geq \rho - \xi_b}_{One \;\; constraint \;\; per \;\; bag}, \qquad b \in \mathcal{B}^-
\end{aligned}$$

where $\mathbf{d}$ is the vector of all latent key instance variables, ρ is the width of the sparse margin, and λ is an additional regularization constant.

4.9. Instance Key Instance SVM (I-KI-SVM) [17]

Liu et al.[17] introduce a second variant of KI-SVM in their same work, called Instance KI-SVM. Its only difference from Bag KI-SVM is that as opposed to adding a single constraint per negative bag, Instance KI-SVM adds one constraint to the optimization problem per each negative bag instance

$$\begin{aligned}
\min_{\mathbf{w},\rho,\boldsymbol{\xi},\mathbf{d}} \; & \frac{1}{2}\|\mathbf{w}^2\| + \frac{C}{2} \sum_{b \in \mathcal{B}^+} \xi_b^2 + \frac{\lambda C}{2} \sum_{b \in \mathcal{B}^-} \sum_{n=1}^{N_b} \xi_{bn}^2 \\
s.t. \; & \mathbf{w}^T \sum_{n=1}^{N_b} d_{bn} \phi(\mathbf{x}_{bn}) \geq \rho - \xi_b, \quad b \in \mathcal{B}^+ \\
& \sum_{n}^{N_b} d_{bn} = 1, \qquad b \in \mathcal{B}^+ \\
& \underbrace{-\mathbf{w}^T \phi(\mathbf{x}_{bn}) \geq \rho - \xi_{bn}}_{One \;\; constraint \;\; per \;\; instance}, \qquad b \in \mathcal{B}^-, \\
& \qquad\qquad\qquad\qquad\qquad\quad n = 1, \cdots, N_b.
\end{aligned}$$

Here, ξ_b denotes the slack variable for a positive bag $b \in \mathcal{B}^+$, and ξ_{bn} the slack variable for instance n of a negative bag $b \in \mathcal{B}^-$. The rest of the notation is the same as for B-KI-SVM.

4.10. Iterative Axis-Parallel Rectangles (iAPR) [7]

This method builds on the central assumption that the positive instances lie in a close neighborhood, and can be isolated from negative instances by a hyperrectangle on the feature space. Authors introduce various heuristics about fitting the hyperrectangle to data such as stepwise growing.

4.11. SIL-SVM [3]

This is the standard supervised SVM that assigns each observation the label of the bag it belongs to. For cases where the average positive instance label ratio within positive bags is close to 1

$$\frac{1}{|\mathcal{B}^+|} \sum_{b \in \mathcal{B}^+} \frac{\sum_{n=1}^{N_b} \mathbb{I}(y_{bn} = +1)}{N_b} \approx 1,$$

the performance of SIL-SVM is comparable to MIL methods, since the bag labels have high correspondence to the instance labels. Hence, this naive method serves as a yardstick showing how weak the bag labels are in an MIL data set.

5. Results

We evaluate the MIL methods on two CAD applications:

1. **Histology**: This is a Barrett's cancer diagnosis data set provided by Institute of Pathology, Helmholtz Zentrum Munich, Germany [16]. The data set consists of 210 tissue core images (143 cancer and 67 healthy) taken from 97 patients. The average size of the images is 2179×1970 pixels. We split the images into a Cartesian grid of 200×200 pixel patches, and represent each patch with a 738-dimensional feature vector as described in Section 3. To maximize the covariance across feature dimensions, we reduce the feature dimensionality to 100 by standard principal component analysis (PCA). The resultant data set includes 14303 data points (patches). Among these patches, 58 % include cancer, and the rest are healthy. Figure 2

Figure 2: Three sample tissue core images from the Barrett's cancer histology data set. The image on the left shows a healthy case, and tumor regions are marked as green polygons on the other two.

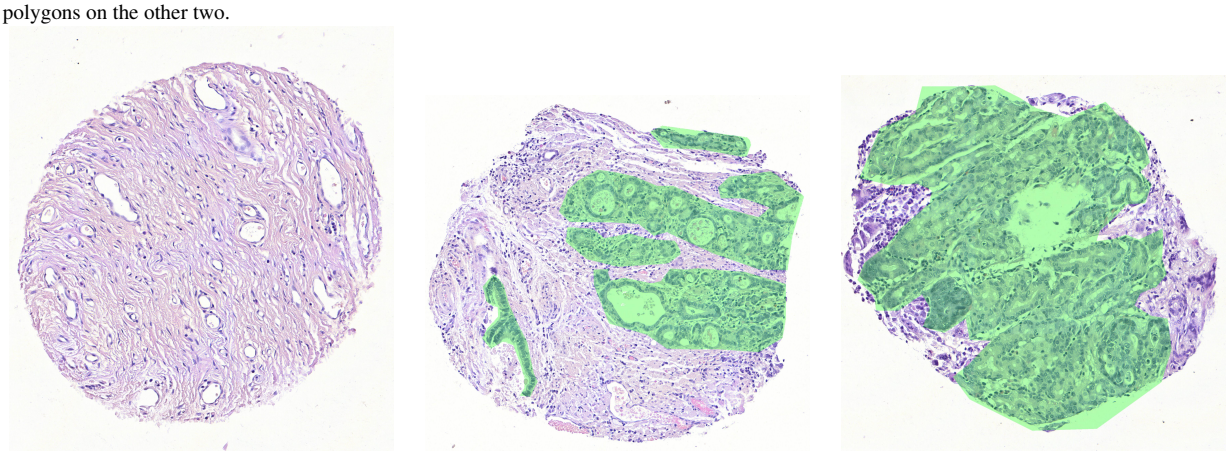
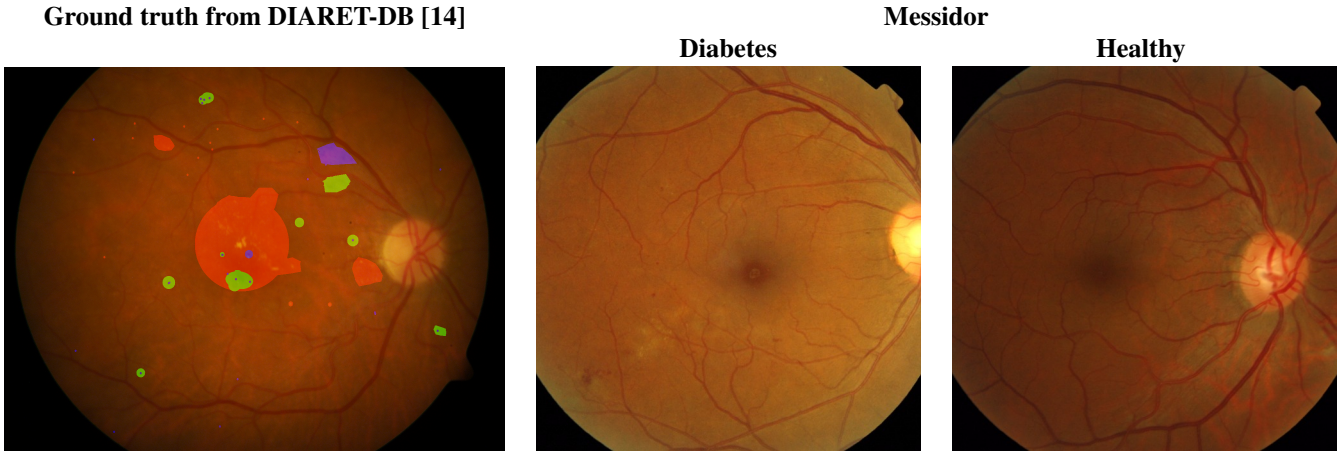


Figure 3: As a reference, ground-truth annotations of three different diabetic retinopathy lesions taken from the DIARET-DB [14] data set are shown on the left. **Red:** Exudates, **Green:** Hemorrhages, **Blue:** Microaneurysms. One diseased example and one healthy example taken from the Messidor data set are shown on the middle and on the right.



shows three sample images (bags) from this data set. For all images in this data set, tumor regions drawn by expert pathologists are available, which serve as pixel-level supervision.

2. **Retinopathy**: This is a public diabetic retinopathy screening data set, named as Messidor [3], collected by three universities in France. This data set contains 1200 eye fundus images (654 diseased and 546 healthy) taken by three hospitals in France. The original sizes of the images are between 1440×960 and 2304×1536 pixels. For standardization, we rescaled all images to 700×700 pixels, and applied contrast stretching. We split each image into patches of 135×135 pixels and represented each patch as described in Section 3. Figure 3 shows two sample images taken from this data set (middle and right). As a reference, we also give one example from the DIARET-DB [14] (left) along with pixel-level ground-truth annotations, since they are not available for the Messidor data set.

[3] http://messidor.crihan.fr/index-en.php.

We evaluate the generalization performance of all methods using 5 times 4-fold cross-validation for the histology data set, and 10 times 2-fold cross-validation (as in [21]), for the retinopathy data set. In all result tables below, we report the following four performance metrics averaged over all data splits and repetitions:

- **Accuracy :** Percentage of correctly classified test points,
- **F1 Score :** Harmonic mean of precision and recall,
- **AUC-ROC :** Area under Receiver Operating Characteristics (ROC) curve,
- **AUC-PR :** Area under precision-recall curve.

For all kernelizable methods in the list, we use the Radial Basis Function (RBF) kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$ with a length scale of $\sigma = \sqrt{D}$, following the heuristic of Chang et al.[4].

Table 4: **Bag level prediction :** Image prediction (cancer diagnosis) performance of MIL methods on the Barrett's cancer histology data set. mi-Graph gives higher performance than the other methods for all four metrics. Prediction performance of a kernel SVM under patch-level (strong) supervision is given in the bottom row. The highest score among MIL methods is shown in bold.

	Accuracy (%)	**F1 score**	**AUC-ROC**	**AUC-PR**
mi-Graph [33]	**86.4**	**0.90**	**0.93**	**0.97**
MILBoost [24]	83.0	0.88	0.91	0.96
B-KI-SVM [17]	82.6	0.88	0.91	0.95
GPMIL [15]	81.2	0.88	0.90	0.93
I-KI-SVM [17]	80.3	0.86	0.89	0.93
iAPR [7]	79.4	0.87	0.88	0.94
Citation k-NN [26]	74.5	0.83	0.72	0.82
EMDD [31]	72.2	0.83	0.72	0.82
mi-SVM [2]	68.4	0.81	0.86	0.76
MI-SVM [2]	68.1	0.81	0.89	0.94
SIL-SVM [3]	68.1	0.81	0.92	0.95
Fully Supervised SVM	85.0	0.90	0.92	0.96

Table 4 shows performance scores of the MIL methods for image-level prediction (i.e. cancer diagnosis) on the histology data set. Since patch-level expert labels are available for this data set, we also provide performance scores for the standard SVM trained by patch-level (i.e. strong) supervision (denoted as *Fully Supervised SVM*) for comparison. mi-Graph ranks as the best-performing MIL method in all four metrics, and is closely followed by MILBoost and B-KI-SVM. It is noteworthy that *Supervised SVM* performs marginally worse than mi-Graph, which reflects the annotation noise for high resolutions.

In addition to cancer diagnosis, MIL can also be used for cancer localization by making instance-level prediction from methods trained by bag-level supervision. Table 5 shows performance scores of MIL methods in instance prediction. Naturally, accuracies undergo a sharp drop due to the gap between supervision and prediction granularities. MILBoost gives the highest prediction accuracy, and mi-SVM outperforms other methods according to the remaining three metrics. High performance of mi-SVM can be attributed its semi-supervised nature (i.e. it simultaneously discovers missing positive bag instance labels and operates directly on instance-level data distributions).

Table 6 shows performance scores of the MIL methods on the retinopathy data set. As for the previous application, mi-Graph provides clearly the best peformance. On the same data set, Agurto et al.[1] report 0.84 AUC-ROC, and Quellec et al.[21] reach the globally highest score of 0.88 AUC-ROC. Our scores are not directly comparable to Agurto et al.[1] who focus their analysis on a subset of the Messidor data set. Thanks to its specialized feature set and multiscale patch representation, Quellec et al.[21] improves over our score by 6 percentage points. Note that both of these methods are specifically tailored for diabetic retinopathy screening, while we rely in this study on generic feature sets and learning algorithms for the sake of comparison across different CAD applications and MIL methods.

6. Training times

In addition to generalization performance, training time is also an essential metric in measuring the practical value of a machine learning method. I-KI-SVM appears as the method with the fastest training procedure on average, and MILBoost as the by far slowest. mi-Graph is fastest in the histology application. Its training time in retinopathy application is moderate, but still within the feasible boundary. The main reason for this huge speed difference between two data sets with comparable number of instances (14303 in histology versus 14400 in retinopathy) is the difference in feature sizes (100 in histology versus 657 in retinopathy). Extensive kernel computations in mi-Graph makes the feature dimensionality as the main bottleneck of its computational performance.

Table 7: Training times in seconds. I-KI-SVM is has the highest average training speed. mi-Graph is trained fastest on histology data set, and has a feasible training time on the retinopathy data set. The shortest training time is shown in bold for both applications.

	Histology	**Retinopathy**
mi-Graph	**10.6**	389.5
MISVM	10.8	168.3
I-KISVM	16.7	**9.1**
EMDD	20.0	36.0
SIL-SVM	106.2	232.5
B-KISVM	107.7	67.3
miSVM	126.6	742.1
Citation k-NN	943.0	799.4
iAPR	949.6	840.1
GPMIL	1491.7	149.4
MILBoost	2896.6	5992.3

7. Discussion and conclusion

In this study, we estimate the predictive power of multiple instance learning in two very different CAD applications: histology cancer diagnosis and diabetic retinopathy screening. The main outcome of the study is that the mi-Graph method generalizes best across application domains in bag label prediction. This is likely to be due to that mi-Graph directly models within-bag instance relationships, which is a rich informa-

Table 5: **Instance level prediction :** Patch prediction (tumor localization) performance of MIL methods on the Barrett's cancer histology data set. MILBoost gives the highest prediction accuracy. mi-SVM, on the other hand, ranks the first according to the remaining three performance metrics. Entries for mi-Graph are left blank, since this method does not allow instance-level prediction. Prediction performance of a kernel SVM under patch-level (strong) supervision is given in the bottom row. The highest score among MIL methods is shown in bold.

	Accuracy (%)	**F1 score**	**AUC-ROC**	**AUC-PR**
MILBoost [24]	**66.7**	0.70	0.75	0.71
GPMIL [15]	65.8	0.54	0.77	0.69
B-KI-SVM [17]	64.7	0.48	0.67	0.67
I-KI-SVM [17]	63.0	0.37	0.69	0.68
mi-SVM [2]	62.7	**0.71**	**0.84**	**0.82**
iAPR [7]	57.8	0.34	0.50	0.47
Citation k-NN [26]	54.3	0.67	0.69	0.76
EMDD [31]	54.1	0.33	0.56	0.52
MI-SVM [2]	46.9	0.64	0.74	0.71
SIL-SVM [3]	46.9	0.64	0.80	0.75
mi-Graph [33]	-	-	-	-
Fully Supervised SVM	83.5	0.82	0.91	0.90

Table 6: **Bag level prediction :** Diabetes detection performance of MIL methods in comparison. mi-Graph gives higher performance than the other methods by all four metrics. Fully Supervised SVM results are not given since instance level supervision is not available for this data set. The highest score among MIL methods is shown in bold.

	Accuracy (%)	**F1 score**	**AUC-ROC**	**AUC-PR**
mi-Graph [33]	**72.5**	**0.75**	**0.81**	**0.85**
MILBoost[24]	64.1	0.66	0.70	0.73
Citation k-NN [26]	62.8	0.68	0.65	0.69
GPMIL [15]	59.2	0.43	0.76	0.80
SIL-SVM [3]	58.4	0.72	0.78	0.82
B-KI-SVM [17]	55.9	0.68	0.60	0.64
I-KI-SVM [17]	55.5	0.44	0.61	0.65
EMDD [31]	55.1	0.69	0.58	0.61
MI-SVM [2]	54.5	0.70	0.68	0.73
mi-SVM [2]	54.5	0.71	0.58	0.62
iAPR [7]	54.4	0.70	0.53	0.60
Fully Supervised SVM	-	-	-	-

tion source in CAD applications, since instances are spatially-correlated: neighboring patches are expected to be more similar to each other than non-neighboring ones. The uppermost performance of mi-Graph motivates future research on specializations of mi-Graph to various CAD applications, for instance, by application-specific kernels. In instance label prediction, mi-SVM appears as the best-performing method benefiting from its semi-supervised nature. The fact that SIL-SVM (for each bag, bag label is assigned to all of its instances, and a standard SVM is trained on the resultant data set) is not drastically worse than MIL methods indicates that the positive class ratio is not very low. However, the constant improvement of at least one of the MIL methods over SIL-SVM in all cases motivates use of the MIL setup in CAD applications.

References

[1] Carla Agurto, Victor Murray, Eduardo Barriga, Sergio Murillo, Marios Pattichis, Herbert Davis, Stephen Russell, Michael Abràmoff, and Peter Soliz. Multiscale am-fm methods for diabetic retinopathy lesion detection. *Medical Imaging, IEEE Transactions on*, 29(2):502–512, 2010.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in NIPS*, 2003.

[3] Razvan C Bunescu and Raymond J Mooney. Multiple instance learning for sparse positive bags. In *Proceedings of the 24th international conference on Machine learning*, pages 105–112. ACM, 2007.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[5] Hang Chang, Alexander Borowsky, Paul Spellman, and Bahram Parvin. Classification of tumor histology via morphometric context. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2203–2210. IEEE, 2013.

[6] Yixin Chen, Jinbo Bi, and James Ze Wang. Miles: Multiple-instance learning via embedded instance selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):1931–1947, 2006.

[7] T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

[8] Scott Doyle, Mark Hwang, Kinsuk Shah, Anant Madabhushi, Michael Feldman, and John Tomaszeweski. Automated grading of prostate cancer using architectural and textural image features. In *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pages 1284–1287. IEEE, 2007.

[9] Luca Giancardo, TP Karnowski, KW Tobin, F Meriaudeau, and Edward Chaum. Validation of microaneurysm-based diabetic retinopathy screening across retina fundus datasets. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pages 125–130. IEEE, 2013.

[10] Cigdem Gunduz, Bülent Yener, and S Humayun Gultekin. The cell graphs of cancer. *Bioinformatics*, 20(suppl 1):i145–i151, 2004.

[11] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[12] Po-Whei Huang and Cheng-Hsiung Lee. Automatic classification for pathological prostate images based on fractal analysis. *Medical Imaging, IEEE Transactions on*, 28(7):1037–1050, 2009.

[13] M. Kandemir, A. Feuchtinger, A. Walch, and F. A. Hamprecht. Digital Pathology: Multiple instance learning can detect Barrett's cancer. In *ISBI. Proceedings, in press*, 2014.

[14] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Asta Raninen, Raija Voutilainen, Hannu Uusitalo, Heikki Kälviäinen, and Juhani Pietilä. The diaretdb1 diabetic retinopathy database and evaluation protocol. In *BMVC*, pages 1–10, 2007.

[15] M. Kim and F. De La Torre. Gaussian process multiple instance learning. In *Proc. of ICML*, 2010.

[16] Rupert Langer, Sandra Rauser, Marcus Feith, Jörg M Nährig, Annette Feuchtinger, Helmut Friess, Heinz Höfler, and Axel Walch. Assessment of erbb2 (her2) in oesophageal adenocarcinomas: summary of a revised immunohistochemical evaluation system, bright field double in situ hybridisation and fluorescence in situ hybridisation. *Modern Pathology*, 24(7):908–916, 2011.

[17] Y.-F. Li, J.T. Kwok, I.W. Tsang, and Z-H. Zhou. A convex method for locating regions of interest with multi-instance learning. In *Machine learning and knowledge discovery in databases*, pages 15–30. Springer, 2009.

[18] C.G. Loukas, G.D. Wilson, B. Vojnovic, and A. Linney. An image analysis-based approach for automated counting of cancer cell nuclei in tissue sections. *Cytometry A*, 55(1):30–42.

[19] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Advances in NIPS*, pages 570–576, 1998.

[20] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Advances in NIPS*, pages 512–518. MIT Press, 2000.

[21] Gwénolé Quellec, Mathieu Lamard, Michael D Abràmoff, Etienne Decencière, Bruno Lay, Ali Erginay, Béatrice Cochener, and Guy Cazuguel. A multiple-instance learning framework for diabetic retinopathy screening. *Medical Image Analysis*, 2012.

[22] Gwénolé Quellec, Mathieu Lamard, Guy Cazuguel, Béatrice Cochener, and Christian Roux. Adaptive nonseparable wavelet transform via lifting and its application to content-based image retrieval. *Image Processing, IEEE Transactions on*, 19(1):25–35, 2010.

[23] Bernhard Scholkopf and Alex Smola. Learning with kernels, 2002.

[24] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. *Advances in NIPS*, 2006.

[25] Ching-Wei Wang. Robust automated tumour segmentation on histological and immunohistochemical tissue images. *PloS one*, 6(2):e15818, 2011.

[26] J. Wang and J.D. Zucker. Solving multiple-instance problem: A lazy learning approach. 2000.

[27] Bangxian Wu. Clinical applications of imaging informatics. *International J of Computer Assisted Radiology and Surgery*, 7(4):635–646, 2012.

[28] Y. Xu, J. Zhang, E.-C. Chang, M. Lai, and Z. Tu. Context-constrained multiple instance learning for histopathology image segmentation. volume 7512 of *Lecture Notes in Computer Science*, pages 623–630. Springer Berlin Heidelberg, 2012.

[29] Y. Xu, J.Y. Zhu, E. Chang, and Z. Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *Int'l Conf. CVPR*, pages 964–971. IEEE, 2012.

[30] Gang Zhang, Jian Yin, Ziping Li, Xiangyang Su, Guozheng Li, and Honglai Zhang. Automated skin biopsy histopathological image annotation using multi-instance representation and learning. *BMC Medical Genomics*, 6(Suppl 3):S10, 2013.

[31] Q. Zhang and S.A. Goldman. EM-DD: An improved multiple-instance learning technique. *Advances in NIPS*, 14:1073–1080, 2001.

[32] Dehua Zhao, Yixin Chen, and N Correa. Automated classification of human histological images, a multiple-instance learning approach. In *Life Science Systems and Applications Workshop, 2006. IEEE/NLM*, pages 1–2. IEEE, 2006.

[33] Z.H. Zhou, Y.Y. Sun, and Y.F. Li. Multi-instance learning by treating instances as non-iid samples. In *Proc. ICML*, pages 1249–1256. ACM, 2009.