DIGITAL PATHOLOGY: MULTIPLE INSTANCE LEARNING CAN DETECT BARRETT'S CANCER

Melih Kandemir[1]*, Annette Feuchtinger*[2]*, Axel Walch*[2] *and Fred A. Hamprecht*[1]

[1]University of Heidelberg, HCI/IWR, Germany,
[2]Institute of Pathology, Helmholtz Zentrum München, Germany

ABSTRACT

We study diagnosis of Barrett's cancer from hematoxylin & eosin (H & E) stained histopathological biopsy images using multiple instance learning (MIL). We partition tissue cores into rectangular patches, and construct a feature vector consisting of a large set of cell-level and patch-level features for each patch. In MIL terms, we treat each tissue core as a bag (group of instances with a single group-level ground-truth label) and each patch an instance. After a benchmarking study on several MIL approaches, we find that a graph-based MIL algorithm, mi-Graph [1], gives the best performance (87% accuracy, 0.93 AUC), due to its inherent suitability to bags with spatially-correlated instances. In patch-level diagnosis, we reach 82% accuracy and 0.89 AUC using Bayesian logistic regression. We also pursue a study on feature importance, which shows that patch-level color and texture features and cell-level features all have significant contribution to prediction.

***Index Terms*—** Cancer diagnosis, multiple instance learning, histopathological tissue imaging

1. INTRODUCTION

Computer-assisted diagnosis (CAD) is a very useful application of pattern recognition to medicine. In CAD, computerized data analysis results are used in making clinical decisions. An application of CAD is automated analysis of histopathological images gathered from tissue biopsies (see [2] for a comprehensive survey). Automation of tissue analysis is beneficial from several perspectives. It alerts the pathologist to suspicious regions, hence helps her not to miss any important information in the tissue image. It also gives way to standardization of cancer diagnosis and grading. Although protocols such as Gleason grading system [3] have been introduced, the subjectivity problem still persists, and can only be avoided by automatization.

Recent machine learning and image analysis techniques allow highly robust recognition of cancerous tissues from image content. Wang et al. [4] report 80% accuracy in pixel-level classification of lung cancer. This sort of finest-level prediction with acceptable accuracy is possible only in discriminating high-grade cancer from healthy cases. Besides, this accuracy is an overestimation since classification of stroma pixels is also taken into account, which is essentially a rather simple background subtraction task. Huang et al. [5] show that fractal features are very discriminative in image-level classification of prostate tissues into five grades. In their comprehensive survey, Gurcan et al. [2] report 62% accuracy in brain tumor detection at the image level when only intensity information is used, 89% when textural info is used, and accuracies over 90% when graph-based techniques are used.

In this paper, we report a feasibility study on automated diagnosis of *Barrett's cancer* from H & E stained tissue images [6]. The particular difficulty of the data set is that it includes images from both low and high-grade cases. In low-grade cases, glandular structures are not severely distorted, and cell density has not grown drastically. Hence, the required predictor has to be very sensitive to slight changes in various levels of detail. We overcome this problem by constructing a comprehensive feature set that includes textural, color, and cell-level features altogether. We use an interactive segmentation software, ilastik [7], for cell segmentation, which dramatically facilitates the annotation process.

We construct our data set by partitioning tissue core images into a grid of patches, and representing each patch by a feature vector. For core-level diagnosis, we formulate the problem in the multiple instance learning (MIL) framework [8], which stands for the machine learning setup where the data set $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_B\}$ is partitioned into groups of observations $\mathbf{X}_1 = \{\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i}\}$, called bags, and only one ground-truth label $l_i \in \{-1, +1\}$ is provided for each bag. While a positively labeled bag contains *at least* one positively labeled instance, all instances in a negatively labeled bag have a negative label. We perform a benchmarking study on several existing MIL algorithms, and find that mi-Graph [1] gives the best performance with significant improvement over several other approaches (87% accuracy, 0.93 AUC). The mi-Graph represents each bag as a graph with edges between similar instances. This provides information about spatial relationships of within-bag instances, making mi-Graph an inherently suitable algorithm for cancer diagnosis tasks.

We also invesigate Barrett's cancer diagnosis at higher resolutions. Using Bayesian logistic regression, we reach an accuracy of 82% in classification of the patches as healthy and cancer. The fact that the patch-level performance is slightly lower than core-level performance can be attributed to the noise in ground-truth labels (cancer regions drawn by pathologists) at higher levels of detail.

Finally, we provide a quantitative analysis of feature importance, which shows that both patch-level summary features and cell-level features contribute to diagnosis significantly.

2. MATERIALS AND METHODS

2.1. Cell Segmentation

Our data set consists of biopsy images of H & E stained tissues taken from patients at both early and late stages of the cancer. Hence, the data set contains many cases where Barrett's cancer did not cause a drastical change in the appearance of the tissue. Glandular structures remain preserved, and the increase in cell density is small. The most informative visual cues in such cases are slight changes in size and color of the nuclei. Capturing this information requires a successful segmentation of nuclear regions. We achieve this segmentation in three steps as described below.

Step 1: We train a pixel-level classifier using an interactive learning software: ilastik [7]. The expert annotates the image by a few brush strokes indicating ground-truth labels, and trains the classifier. Then she keeps annotating only the misclassified pixels. As seen in Figure 1 (a), very little ground-truth data is sufficient for a reasonable segmentation. In our case, we segment the raw image into the following five classes (corresponding color codes of classes used in the figures are given in parentheses): i) cancer cell (red), ii) healthy cell (green), ii) lymphocyte cell (pink), iv) stroma (blue), background (yellow). We introduce the first two classes to discriminate cancer and healthy cells right at the pixel classification stage. The third cell class, *lymphocyte*, has been introduced to discriminate lymphocytes from cancer cells, since both look faded. Notice that the cancer regions are dominated by red pixels, and the healthy ones are dominated by green and pink ones.

Pixel classification has been done using a standard random forest classifier along with a set of filter response features such as gaussian smoothing and laplacian of gaussian. Figures 1 (b) and (c) show a tissue with a cancer region marked in green, and its segmentation, respectively.

Step 2: We detect the local maxima in the probability map (matrix constructed by the probabilistic decision output of the decision tree classifier for each pixel) of healthy and cancer cells using extended maxima transform [9].

Step 3: We segment the cells by watershed transform using the detected local maxima as seed points. Figure 1 (d) shows the end-result of the segmentation for an example core.

Table 1. Object-level features extracted from each healthy and cancer 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 2. Features extracted from each patch.

	Color features
1	Color histograms of the entire patch, healthy, cell pixels, cancer cell pixels, lymphocyte cell pixels, and stroma pixels
	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
	Object features
5	Minimum, maximum, mean, standard deviation, skewness, and kurtosis of features (given in Table 1) of all healthy and cancer cells in a patch

Note that here our goal is not to segment all the cells as accurately as possible, but to extract useful markers for diagnosis.

2.2. Data Preparation

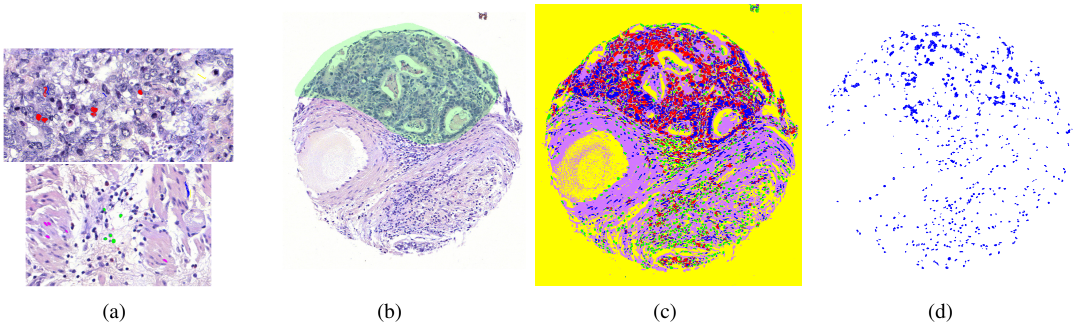
Our data consists of 214 tissue cores (145 cancer and 69 healthy) taken from 97 patients. We manually annotated 5.8% of the pixels of two cancer and two healthy core images in ilastik to train the pixel classifier. We discarded these cores in further analysis.

We split each tissue core (with avg. size of 2179x1970 pixels) into a grid of 200x200 pixel patches. For each healthy and cancer cell within each patch, we extracted the object-level features listed in Table 1. In addition to several statistics of these features within a patch, we extracted a large set of color and texture features, as listed in Table 2. The resultant data set includes 16698 data points (one per each patch) and 1445 features. The patch-level base rate (percentage of the dominating class, cancer in our case) of the data set is 58%.

2.3. Barrett's Cancer Diagnosis by Graph-Based Multiple Instance Learning

We formulate the cancer diagnosis problem within the MIL framework as follows. Each tissue core is a bag, and each patch within a tissue core is an instance. A bag with positive label denotes existence of a cancer region within the core.

Fig. 1. The segmentation process. a) Manual annotations using ilastik. b) Raw image (green region is cancer). c) Pixel classification result of ilastik. **Red:** cancer cells, **Green:** healthy cells, **Blue:** lymphocyte cells, **Pink:** stroma, **Yellow:** background. More red pixels in the north, more green and pink pixels in the south. d) Resultant segmentation of healthy and cancer cells after watershed transform.



Otherwise, the tissue is healthy. Among several solutions to MIL, we prefer mi-Graph [1], since its graph-based structure fits nicely to the spatial relationships of patches within a core. As discussed in [2], topological properties of tissue structures are very informative indicators of cancer.

The algorithm represents each bag by a similarity graph. In particular, it calculates the cross-similarities of bag instances by a kernel function $k(\mathbf{x}_{ia}, \mathbf{x}_{jb})$. Then it constructs a graph by drawing a link between two instances if their similarity is above a predefined threshold δ. Let W_i be the affinity matrix of bag i, whose entry is $w^i_{au} = 1$ if there is a link between instances a and u, and $w^i_{au} = 0$ otherwise. Given the instance-level kernel function, and the affinity matrices of bags, mi-Graph constructs the following bag-level kernel function:

$$k(X_i, X_j) = \frac{\sum_{a=1}^{n_i} \sum_{b=1}^{n_j} W_{ia} W_{jb} k(\mathbf{x}_{ia}, \mathbf{x}_{jb})}{\sum_{a=1}^{n_i} W_{ia} \sum_{b=1}^{n_j} W_{jb}}$$

where $W_{ia} = 1/\sum_{u=1}^{n_i} w^i_{au}$, $W_{jb} = 1/\sum_{v=1}^{n_j} w^j_{bv}$. A standard SVM is then trained with the bag-level kernel matrix calculated by this function. Here, W_{ia} has a smaller value for instances that are similar to more number of other instances within the same bag, and a larger value for instances more different from the rest. Resultantly, odd instances within bags are made more influential on the bag-level similarity metric.

3. RESULTS AND DISCUSSION

We evaluate the feasibility of automated diagnosis of Barrett's cancer in two experiments, one at the core level, and one at the patch level. While the core-level classification experiment stands for *diagnosis* of cancer for a patient, the patch-level experiment corresponds to *locating* it within the tissue. In both experiments, we use randomly selected 75% of the cores for training, and the rest for testing. We repeat this procedure 20 times and averaged all numbers reported below across repetitions. Core-level splitting of training and test data also for patch-level analysis prevents the potential bias from cross-similarities of patches belonging to the same core. To eliminate noisy features from the large feature set, we reduce the dimensionality of our data set to 200 using principal component analysis (PCA).

Table 3. Bayesian logistic regression (B. Log. Reg.) outperforms SVM with linear and RBF kernels in cancer location.

Method	**Acc.**	**AUC**	**F1 Score**	**Time (sec)**
B. Log. Reg.	**0.82**	**0.89**	**0.81**	**3.7**
SVM (Linear)	0.77	0.85	0.77	28.3
SVM (RBF)	0.78	0.84	0.76	32.6

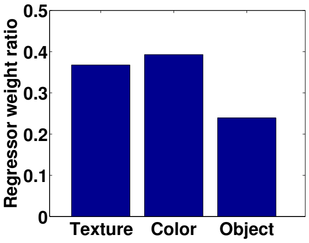
We evaluate the performance of the learning models in comparison with respect to the following three metrics:

- **Accuracy (Acc.):** Correct classification ratio.
- **AUC:** Area under Receiver Operating Characteristics (ROC) curve.
- **F1 Score:** Harmonic mean of precision and recall.

For cancer location, we compare three commonly used algorithms: i) Bayesian logistic regression with automatic relevance determination (ARD) prior on regressor weights, ii) SVM with linear kernel, and iii) SVM with the nonlinear radial basis function (RBF) kernel. Test performance of these algorithms is given in Table 3. The fact that the nonlinear SVM does not bring any improvement in performance over the linear models is due to that the relationship between our feature space and the diagnosis output is *linear*.

Figure 2 shows the ratio of the absolute sum of the regressor weights of the three feature categories (textural, color, and cell) when logistic regression is trained on the entire feature set. The fact that each category has at least 20% contribution serves as an evidence to that each category plays an important role in prediction (one-way ANOVA test between the contribution of each category in 20 replications and a zero column vector gives $p < 3.46 \times 10^{-8}$).

Fig. 2. Total contributions of feature categories to prediction. All three feature categories contribute to diagnosis significantly.



For core-level diagnosis, we compare the following well-known MIL algorithms to mi-Graph:

- **SIL-SVM**: The bag label is assigned to all positive bag instances, and a standard SVM is trained in a single-instance fashion.
- **MI-SVM** [10]: Standard SVM is trained on the most representative instances of bags. The SVM and the instances are inferred in an iterative EM-like (expectation maximization) algorithm. In the E-step, given the current SVM model, the instance with the largest distance from the margin is chosen from each bag as its representative. And in the M-step, SVM is trained on the selected set of representative instances.
- **mi-SVM** [10]: This model treats MIL as a semi-supervised learning problem. It infers the *missing* labels of positive bag instances in iterations. In the E-step, missing labels are predicted from the current SVM model. And in the M-step, SVM is trained on the data set with the inferred labels.

As shown in Table 4, mi-Graph clearly outperforms the other three models in all three metrics. This is because mi-Graph uses a rich information of within-bag relationships of instances. This information is particularly important in setups like ours where instances are spatially-related: neighboring patches are expected to be similar to each other. The other three models above are ignorant to this source of information. In terms of training time, mi-Graph ranks as second after MI-SVM with marginal difference.

The fact that core-level diagnosis performance is better than patch-level performance is due to the increase in ground-truth noise as spatial resolution gets larger. For the purpose of pathology, it is sufficient to provide labels as *regions*. However, the ground truth in these regions is not necessarily homogeneous. A region marked as cancer may, and often does, include healthy subregions, or cells. Hence, more robust diagnosis is possible at larger scales, at the expense of being spatially less accurate. The MIL framework exactly suits to this problem setup.

Table 4. Cancer diagnosis performance of MIL models at the core level. mi-Graph gives the best performance, since it exploits the spatial correlations of instances belonging to the same bag.

Method	Acc.	AUC	F1 Score	Time (sec)
SIL-SVM	0.68	0.89	0.40	183.3
MI-SVM	0.68	0.79	0.41	**20.9**
mi-SVM	0.69	0.88	0.43	731.1
miGraph	**0.87**	**0.93**	**0.84**	24.9

4. REFERENCES

[1] Z.H. Zhou et al., “Multi-instance learning by treating instances as non-I.I.D. samples,” *Proc. Int’l Conf. Machine Learning*, pp. 1–8, 2009.

[2] M.N. Gurcan et al., “Histopathological image analysis: a review.,” *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–71, Jan. 2009.

[3] D.F. Gleason, “Histologic grading of prostate cancer: a perspective,” *Human pathology*, vol. 23, no. 3, pp. 273–279, 1992.

[4] C.W. Wang, “Robust automated tumour segmentation on histological and immunohistochemical tissue images,” *PLoS ONE*, vol. 6, no. 2, pp. e15818, 2011.

[5] P.W. Huang and C.H. Lee, “Automatic classification for pathological prostate images based on fractal analysis.,” *IEEE Trans. Medical Imaging*, vol. 28, no. 7, pp. 1037–50, 2009.

[6] R. Langer et al., “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*, vol. 24, no. 7, pp. 908–916, 2011.

[7] C. Sommer et al., “”ilastik: Interactive learning and segmentation toolkit”,” in *Int’l Symposium on Biomedical Imaging*, 2011.

[8] O. Maron et al., “A framework for multiple-instance learning,” *Advances in Neural Information Processing Systems*, pp. 570–576, 1998.

[9] P. Soille, *Morphological image analysis: principles and applications*, Springer-Verlag New York, Inc., 2003.

[10] S. Andrews et al., “Support vector machines for multiple-instance learning,” in *Advances in Neural Information Processing Systems*, 2003.