Automated gastric cancer diagnosis on H&E-stained sections; training a classifier on a large scale with multiple instance machine learning

Eric Cosatto[a], Pierre-François Laquerre[a], Christopher Malon[a], Hans Peter Graf[a], Akira Saito[b], Tomoharu Kiyuna[b], Atsushi Marugame[b], Ken'ichi Kamijo[b]

[a] *Dept. of Machine Learning, NEC Laboratories America, 4 Independence Way, Suite 200, Princeton, NJ 08540, USA*

[b] *Medical Solutions Division, NEC Corporation, Daito Tamachi Building 2nd floor,4-14-22, Shibaura, Minato-ku, Tokyo 108-8558, JAPAN*

Abstract

We present a system that detects cancer on gastric tissue sections stained with hematoxylin and eosin (H&E). We train a tissue classifier using the semi-supervised multi instance learning framework (MIL) where, each tissue is represented by a collection of small rectangular regions-of-interest (ROI) and a single, tissue-level label. Such labels are readily obtained because pathologists diagnose each tissue independently as part of the normal clinical workflow. From a large dataset of 31'406 gastric tissue sections from 12'745 slides (12'726 patients) obtained from a regular 2-month clinical load, we train a classifier on a patient-level partition of the dataset (10'648 patients) and obtain 90% sensitivity for a 90% specificity on the remaining 2'078 never-seen before patients (4'422 tissues). This performance equals the traditional supervised approach where individual ROIs need to be manually labeled which is very costly. The large amount of data used to train our system gives us confidence in its robustness and that it can be safely used in a clinical setting. We demonstrate how it can improve the clinical workflow when it is used for pre-screening or quality control. For pre-screening, the system can diagnose 50% of the tissues with a 0.6% false negative rate, thus halving the case load. For quality control, we compare it to the standard setup and claim a five-fold reduction in the re-examination rate while keeping the same error. The system is currently being validated at a large Japanese laboratory where it is used to double-check clinician’s diagnoses.

Keywords: computer-assisted diagnosis, gastric cancer, histo-pathology, image analysis, machine learning, multi-instance learning, semi-supervised learning, whole-slide imaging

1. INTRODUCTION

While recent studies in molecular biology have provided great advances for diagnostic molecular pathology, traditional histological diagnosis is still the most powerful method for diagnosing diseases. Although still mostly performed by pathologists using optical microscopes, histological diagnosis is currently undergoing the ’digital revolution’ that occurred in the fields of radiology and cytology [Ronco et al., 2003]. This revolution was sparked by the advent of high-resolution whole slide imaging (WSI) scanners and applications in remote diagnosis, teaching and archival systems are already taking advantage of the convenience afforded by digital files over glass slides. Next in line are automated or assisted diagnosis systems, where the computer analyzes the image to provide increased accuracy and speed to the clinical workflow.

However, automated analysis of H&E tissue sections by computer image analysis is extremely difficult for two main reasons. First, at high-power magnification, the segmentation of cells from the structures in which they are embedded is hard, making cell-based diagnosis very challenging. Second, many tumors manifest themselves as subtle changes in the structural fabric of the tissue, making it necessary to develop additional structural analysis algorithms at low-to-medium magnification. Those two types of analysis, taking place at different magnifications, must be combined to produce accurate diagnosis. Those difficulties are compounded by the presence of various histological conditions such as necrosis, hyperplasia, inflammation, etc. Furthermore, structural abnormalities of tissues and benign tumors may complicate the task. For these reasons, automated analysis of histological H&E tissue sections has so far had limited acceptance in the clinical workflow.

Machine learning has recently become the de-facto method to tackle the automated analysis of histological H&E tissue sections. While clinical use of such systems is still in its infancy, there has been a large amount of research in this field. Among the more mature systems we note the prostate cancer detection of [Monaco et al., 2010]. While the majority of computer-assisted diagnosis (CAD) systems use supervised learning, a key aspect of whole tissue classification makes this approach inefficient. While a negative-labeled tissue shows no sign of malignancies on its entire area, a positively-labeled tissue only shows malignancies on parts of the tissue. This problem has been generally addressed by having pathologists manually trace the tumor areas, thus providing definite positive labels. Unfortunately, this approach is labor intensive and cannot be scaled to large training sets, which, in turn, are essential to capture the wide range of conditions encountered in a typical clinical setting. Furthermore, pathologists are often loath to assign a label to small regions without taking into account a larger contextual area. Yet, the key to attaining adequate performance is the ability of a classifier to be trained on a large scale with real day-to-day data samples.

A solution to this problem is provided by the multi-instance learning framework (MIL) [Dietterich, 1997]. Typical supervised learning algorithms deal with instances represented by a single, fixed dimensionality feature vector, to which a label is assigned. In MIL, the input is instead a set of multiple vectors with a single label for the set. A positive label means that at least one instance in the set is labeled positive, while a negative label means that all instances in the set are labeled negative. Hence a tissue sample is segmented into a set of regions of interest (ROI). For positive tissues, one or more ROIs will contain evidence of cancer, while for negative tissues, no ROI will contain any sign of cancer. MIL has been successfully used in a wide range of applications, from drug activity prediction where it was first formalized by [Dietterich, 1997] to content-based image retrieval [Zhang et al., 2005] and face detection [Viola et al., 2005]. Previous uses of MIL in histological sample analysis include [Dundar et al., 2010] where it was used to train support vector machine (SVM) classifiers to differentiate between atypical ductal hyperplasia and ductal carcinoma in-situ in a small dataset of breast biopsy samples. More recently, the work of [Xu et al., 2012] has shown the advantages of MIL for classification of histological tissues, albeit on a very small dataset of colon tissues.

While most academic studies report the performance of classifiers on a given dataset, the modalities and impact of inserting such a classifier in a real clinical workflow are rarely discussed. A common use of classifiers has been in pre-screening, where the task of the classifier is to 'weed-out' easy-to-classify, high-confidence cases (either positive or negative) [Vassilakos et al., 2002]. Another possible use is in quality control, where the classifier rechecks all cases to reduce the incidence of errors.

In this study, we describe how a MIL classifier has been successfully trained to detect cancer on a large dataset of hematoxylin and eosin (H&E) stained gastric tissue sections. We also discuss the impact of the MIL classifier on quality control and pre-screening.

2. MATERIALS AND METHODS

2.1. Data acquisition

We have obtained from a Japanese lab a large set of H&E-stained gastric tissue section on standard glass slides made of a 2-month run, totaling 12'745 slides from 12'726 distinct patients (some patients had several slides). Each slide was scanned, resulting in a single whole-slide file, which is the input to our system. Because accurate analysis of pathological images demands high resolution and the maximum possible amount of color information, we used the NDP NanoZoomer (Hamamatsu Photonics K. K., Shizuoka, Japan) with a resolution of 0.23 microns per pixel at 400X magnification (corresponding to 111'493 dpi). Each slide contains between 1 and 8 tissues, their position automatically detected by software. From the lab pathology reports, binary labels were imputed to each tissue (negative for normal and benign lesion, including adenoma, positive for carcinomas), resulting in a set of 26'879 individually labeled tissues. The set includes 25'087 negative tissues and 1'792 positive tissues showing the typical gastric cancer prevalence in Japan of about 7%. To assess the reliability of the labels, we had them re-examined by an independent pathologist and obtained a relatively high agreement between both pathologists ($AC1 = 97\%, \kappa = 80\%, \pi = 80\%$). These agreement measures are obtained from the following confusion matrix (PP=1'431; NN=24'139; NP=275; PN=361).

2.2. Feature extraction

While many image classification tasks can easily be solved with generic local descriptors such as SIFT [Mikolajczyk and Schmid, 2003], the analysis of histo-pathological samples often relies on counts of particular objects such as nuclei and glands. Hence we program our system to first identify and segment such objects and then extract high-level, medically-relevant features to represent a ROI for classification tasks. As pathologists examine a tissue on a slide under a microscope, they typically identify areas of interest at low magnification and then zoom in on those areas to analyze them in more details. Often, entire areas of the tissue can be safely ignored because they do not contain any object of interest for diagnosis. Furthermore, at the native magnification of 400X, a tissue would be too large to be analyzed efficiently for complex features. Instead, we choose to segment tissue units into regions of interest (ROI) that can be analyzed independently on a single CPU with 1GB of memory and thus can be easily parallelized on today's multi-core CPUs.

A first step in analyzing images of H&E-stained sections is to identify the exact colors of the stains as they are imaged. Specimens stained at different labs exhibit color changes due to the slight variations in concentration of the dyes. Other factors such as staining time, temperature and pH of the solution also affect the colors [Abe et al., 2004]. To robustly identify the color of the dyes in an image, we train a support vector regressor (SVR) to predict the intensities of the R,G and B components of the hematoxylin (H) and eosin (E) colors from color histograms of the input image. On a dataset of 473 training images and 255 validation images, the regressor produces an average error of 4% per channel, so we expect it will provide good generalization. From a new input image, hematoxylin and eosin maps are then obtained by projection of the pixels onto H and E color vectors predicted by the regressor. Other methods, such as color deconvolution, principal component analysis, linear discriminant analysis [Rabinovich et al., 2003], expectation minimization and hierachical self-organizing maps [Datar et al., 2008] have been proposed. However this approach is more robust as it learns from a set of representative samples and is also computationally efficient. We will use it as a first step in the analysis of our tissue samples at both low and high resolution.

We start our analysis workflow at a very low magnification of 10X where non-white areas of the slide are first identified as tissue units that can be analyzed separately. Then, to locate ROIs on each of these tissue units, we obtain their H and E color maps at a low 20X magnification (this easily fits into memory) and compute a smoothed aggregate pixel map A:

$$A = \frac{H + H \cdot E}{2} \tag{1}$$

Intensity peaks on this pixel map provide the center of ROIs. This is a simple and efficient approach to quickly locate areas containing H color, still favoring areas also containing E color to make sure we do not overly emphasize areas of very dense H color such as lymphocyte clusters. The space between peaks is determined by the size of ROIs in order to avoid excessive overlap and the threshold for peak detection is set such as to avoid analysis of areas with few nuclei present. In practice, we obtain an average of 20 ROIs per tissue. We analyze each ROI at two different magnification: 200X and 100X.

At half the native resolution of the scanner (200X), we analyze ROIs of 230 by 230 microns for individual nuclei. We aim to segment individual nuclei from their surroundings in order to assess their number, shape and size. Indeed, nuclei exhibiting increased size are often indicative of malignancies [JGCA, 1998]. The processing steps for segmenting nuclei from the image are shown on figure 1. An adaptive ridge filter is first convolved over the hematoxylin map, detecting small ridges present between two adjacent nuclei. Subtracting its response from the hematoxylin map has the effect of separating touching nuclei. A morphological closing operation is then performed to further separate touching elements. Centroids of connected pixel blobs can now be taken as the center of nuclei. The resulting segmentation is fast and reliable. We assess the performance of the nuclei segmentation on a set of 9000 manually traced nuclei contours and show that the average difference in the long axis length is less than half a micron (or 5%). An example is shown on figure 2.

We then separate the resulting segmented nuclei into 2 bins based on their area. For each bin, we obtain the number of nuclei and statistics

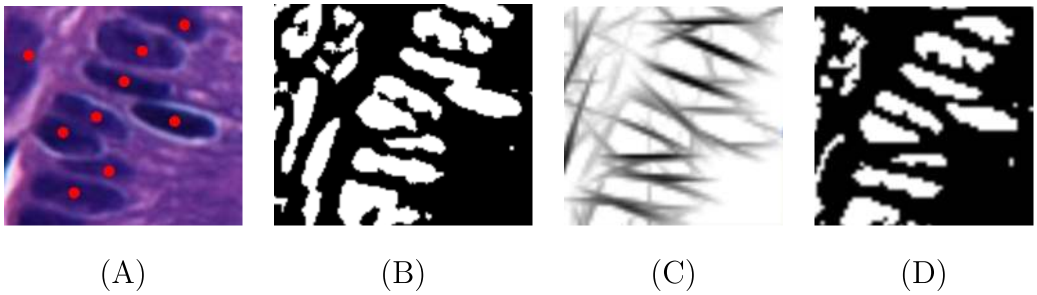

Figure 1: Nuclei extraction. The color image (A) is first binarized by projection onto the predicted hematoxylin color (B). Ridge kernels are convolved over (B), producing map (C) which is multiplied by (B) to produce the resulting (C) map. After morphological closing (D), centroids of connected pixel blobs are marking the center of nuclei in (A).

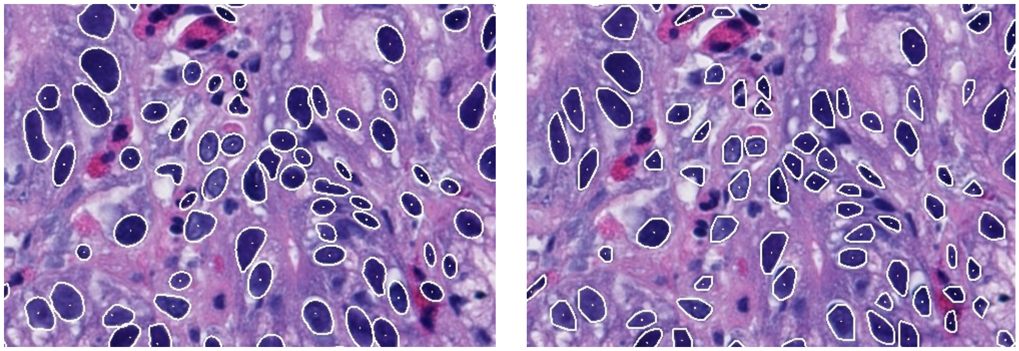

Figure 2: Comparison between manually labeled nuclei contours (left) and automatically generated contours (right).

(mean, standard deviation and percentiles) on their area, long-axis length and hematoxylin content.

At a quarter of the native resolution of the scanner (100X), we analyze ROIs of 460 by 460 microns to quantify the degree of arrangement of nuclei into glands. The function of a tissue is determined to a large extent by the arrangement of the cells, and pathologists can obtain a lot of information about the functionality and health of a tissue by looking at the structure of

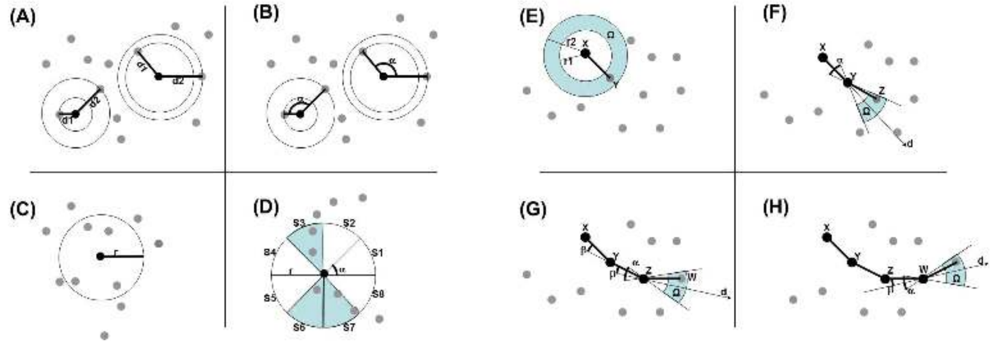

Figure 3: Illustration of structural features.

cell arrangements. In particular, in the presence of cancer, the cells lose their ability to grow in well-organized structures such as epithelial layers and their arrangements tend to become more random. Such randomness (or structural entropy) is an important diagnostic measure and pathologists are trained to identify cell arrangements as normal, functional tissue or as non-typical and indicative of a disease.

An initial estimate of the presence of such epithelial layers is given by a ridge filter tuned to the width of single nucleus layer. However, the proximity of neighboring glands makes it difficult for this approach to produce reliable results. Hence, we need to identify nuclei and verify whether they are arranged in structures. The algorithm first locates the center position of nuclei on the H map using difference of Gaussian filters tuned to detect disks of 3 different sizes. Because it operates at a lower magnification and because it does not need to evaluate the shape of nuclei, this approach is more efficient than the one used at 200X magnification. Using the center points of only medium and large detections, the algorithm proceeds to identify cliques and paths of cells. A clique is a small group of neighboring cells that are joined in a graph where the vertices represent nuclei and the edges the distance between the nuclei. Within cliques of 3 immediate neighbors, we measure alignment and average distance (figure 3 (A,B)). Then, within the ROI, we calculate the mean and standard deviation of those 2 measures. Another set of features is extracted from larger cliques formed around a radius of 40 microns around a center cell (figure 3 (C)). Within these cliques, the number of nuclei as well as the proportion of empty circular sections is obtained (figure 3 (D)). Paths are groups of cells joined by hopping from cell to cell, following a curved trajectory. Angle and distance constraints are defined such that these trajectories match those of typical epithelial layers of glands. Figure 3 (E-H) show how paths can be formed iteratively to locate structured epithelium. Within paths, we measure the number of cells, the average distance to the next cell, the standard deviation of the distances and angles to the next cell. Then, within the ROI, mean and standard deviation of these measure are obtained.

2.3. MIL training

We use the terms 'set' to denote a tissue and 'instance' to denote a region of interest (ROI) belonging to the tissue. We have obtained pathologists' labels for the sets (the tissues), but the instances (the ROIs) do not have labels. Using the set labels and the instances' feature vectors, we train a multi-layer perceptron (MLP) to predict instance labels using the classic backpropagation approach [Rumelhart et al., 1986]. Note that since we do not have labels for the individual instances, we cannot use the error on a single instance for backpropagation. Instead, the predicted label for the set is determined through a one-positive rule, i.e. a set of instances is predicted positive if and only if at least one of its instances is positive. Therefore, we train an instance-level classifier $g : \mathbb{R}^n \rightarrow \mathbb{R}$ and classify sets of instances with:

$$f(X) = \max_{x_i \in X} g(x_i) \tag{2}$$

which amounts to taking the maximum response across all instances. The success of this approach depends on whether g can properly generalize despite the noisy labels caused by the presence of negative instances in positive sets. The training procedure then aims to minimize the loss on $f(X)$ which results in the backpropagation of the error only on the instance that had the maximal response $g(x_i)$. For this purpose, we use a fully connected MLP whose two outputs (one per class) are transformed to represent the probability $P(Y = y|x)$ by applying the function:

$$\text{softmax}(\text{output}_i) = \frac{\exp(\text{output}_i)}{\sum_j \exp(\text{output}_j)} \tag{3}$$

We optimize the MLP weights by minimizing the cross-entropy error:

$$E = -t \log y - (1 - t) \log(1 - y) \tag{4}$$

where t is the true label and y is the estimated probability of the instance being positive. We use two hidden layers with an *hardtanh* activation function, a fast approximation of the hyperbolic tangent function which generally performs as well for a similar number of hidden units[Collobert, 2004]:

$$\mathit{hardtanh}(x) = \begin{cases} -1, & \text{if } x < -1 \\ x, & \text{if } -1 \leq x \leq 1 \\ 1, & \text{if } x > 1 \end{cases} \tag{5}$$

Backpropagating over the one-positive rule has the effect of pushing down on the instance with the highest probability when the set is negative, and pushing it up when it is positive.

Using stochastic gradient descent, we train the classifier presenting the sets in a randomized order every iteration. The MLP weights are initialized randomly and uniformly around $\pm\frac{1}{\sqrt{w}}$, where w is the number of incoming connections at that particular node. This makes the weights small enough to avoid having a saturated output, which would lead to derivatives that are much too small to backpropagate on a computer with finite precision [LeCun et al., 1998]. Similarly, the learning rate η is scaled at a given node by the number of incoming connections. We introduce a learning rate decay λ so that the learning rate at time t is $\eta_t = \frac{\eta}{1+\lambda t}$. This allows us to use a large learning rate at the beginning to make gains more quickly and then progressively fine tune using smaller and smaller updates. We then perform a 3-fold cross-validation to find the most appropriate hyper-parameters:

- η, the learning rate
- the number of hidden units in the first and second layer

The instance (ROI) classifier providing the smallest balanced error rate computed as:

$$e = \frac{1}{2}(\frac{\text{false negatives}}{\text{positives}} + \frac{\text{false positives}}{\text{negatives}}) \tag{6}$$

on the validation set is chosen and integrated into a tissue classifier. The tissue classifier takes as input the classification outputs of the ROI classifier for all ROIs on the tissue and outputs a final decision for the entire tissue.

The architecture of the tissue classifier follows the rule:

$$t(X) = \begin{cases} cancer, & \text{if } \frac{|\{x_i \in X | x_i > t_1\}|}{|S|} > t_2 \\ normal, & \text{otherwise} \end{cases} \tag{7}$$

which essentially means that, if the proportion of positive ROIs (for which the ROI classifier output is $> t_1$) is $> t_2$, then the tissue is positive. We apply the tissue classifier on the training set for a range of possible thresholds points $\{t_1 \in \mathbb{R}, t_2 \in \mathbb{R}\}$ and obtain a cloud of points in the ROC space (false positive rate, true positive rate). Since a classifier is potentially optimal if and only if it lies on the convex hull of the set of points in ROC space [**?**], we declare the set of thresholds points on the ROC convex-hull our tissue classifier.

2.4. Quality control

One possible application of our system is to improve upon traditional quality control in the clinical workflow. In a typical quality control setup, after each tissue has been examined by a primary pathologist, a proportion r of the tissues are randomly picked and re-examined by a second pathologist, who takes the final decision. In contrast, we propose a setup where the MIL classifier re-examines all tissues and when a discrepancy is seen between the machine's diagnosis and the primary pathologist's diagnosis, the tissue is re-examined by a second pathologist who takes the final decision. To quantify the advantage that our system can provide over the existing setup, we perform a comparative error analysis. We make the following assumptions to make this calculation possible: 1) the two pathologists make independent errors and have a similar error rate, 2) the machine and the pathologists make independent errors. The probability of the (*old*) system making an error is:

$$\begin{aligned} P_{old}(err) &= P(S_{old} - |C+) \cdot pv + P(S_{old} + |C-) \cdot (1 - pv) \\ &= (fnr_h \cdot (1 - r) + fnr_h^2 \cdot r) \cdot pv \\ &+ (fpr_h \cdot (1 - r) + fpr_h^2 \cdot r) \cdot (1 - pv) \end{aligned} \tag{8}$$

where r is the recheck rate, pv the prevalence of cancer (C+) and fpr_h, fnr_h denote the performance of a human pathologist (type I and type II errors). In the (*new*) setup, the possible outcomes are summarized in table 1. The probability of error becomes:

$$\begin{aligned} P_{new}(err) &= P(S_{new} - |C+) \cdot pv + P(S_{new} + |C-) \cdot (1 - pv) \\ &= ((1 - fnr_h) \cdot fnr_m \cdot fnr_h + fnr_h^2 \cdot (1 - fnr_m) + fnr_h \cdot fnr_m) \cdot pv \\ &+ (fpr_h \cdot fpr_m + fpr_h^2 \cdot (1 - fpr_m) + (1 - fpr_h) \cdot fpr_m \cdot fpr_h) \cdot (1 - pv) \end{aligned} \tag{9}$$

where fpr_m, fnr_m denote the performance of the MIL classifier (type I and type II errors). The probability that a tissue has to be re-examined by the second pathologist is:

$$\begin{aligned} P_{new}(recheck) &= P(Pa1+, MIL - |C+) + P(Pa1-, MIL + |C+) \\ &+ P(Pa1+, MIL - |C-) + P(Pa1-, MIL + |C-) \\ &= (1 - fnr_h) \cdot fnr_m + fnr_h \cdot (1 - fnr_m) \\ &+ fpr_h \cdot (1 - fpr_m) + (1 - fpr_h) \cdot fpr_m \end{aligned} \tag{10}$$

3. RESULTS

We partitioned the data set of 12’726 patients into three subsets, separating them by patient and keeping the proportion of positive to negative

Table 1: Outcomes of the new (S_{new}) quality control system. Pa1 is the primary pathologist's diagnosis, MIL is the diagnosis of the machine and Pa2 is the secondary diagnosis in case of discrepency between Pa1 and MIL, in which case Pa2 takes the final decision.

Pa1	MIL	Pa2	S_{new}
+	+	X	+
+	-	+	+
+	-	-	-
-	+	+	+
-	+	-	-
-	-	X	-

tissues balanced across the sets: a training set of 8'558 patients (1'257 positive tissues, 19'857 negative tissues), a validation set of 2'090 patients (308 positive tissues, 4830 negative tissues) and a test set of 2'078 patients (291 positive tissues, 4'131 negative tissues). At an average of 20 ROIs per tissues, we extracted features from over 600'000 ROIs. The current time spent for extracting features on one ROI is about 30 seconds, resulting in a total processing time of 20hrs on a 256 cores cluster. We trained the MIL classifier and tissue classifiers on the training set and we evaluated its performance on the test set. We report the performance of our system on figure 4. The curves show the convex hull of the performance points obtained by varying thresholds of the tissue classifier.

For comparison, we also show the ROC curve obtained from the more conventional approach of training a classifier with a set of labeled ROIs. 8000 positive ROIs were labeled manually by a human expert and 8000 negative

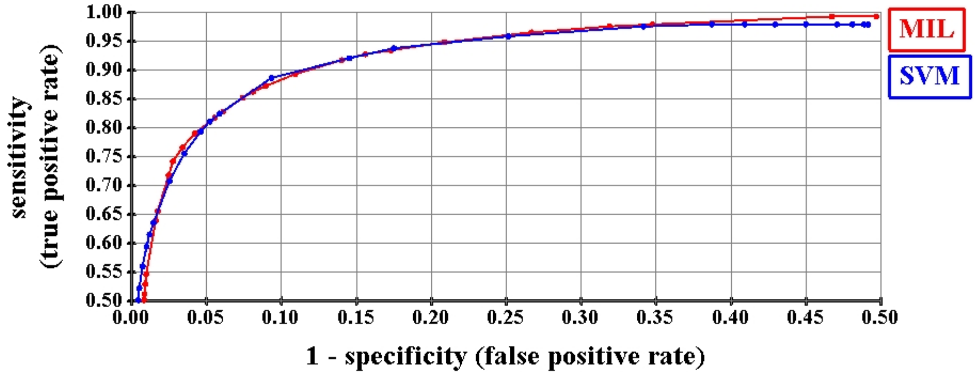

Figure 4: ROC curve of the MIL and SVM gastric tissue classifiers measured on the test set. The area under the curve (AUC) characterizes the overall performance of a classifier. Note that only the upper left quadrant is shown.

ROIs were selected automatically from negative tissues. A support vector machine (SVM) classifier was then trained using the same ROI features. Also, the same rule-based classifier was trained on the same validation set using the SVM classifier and tested on the same test set. Although the SVM classifier was trained with more precise labels (each ROI was individually labeled by a human), the MIL classifier is able to take advantage of the large number of imprecise labels and match the performance of the SVM classifier.

Once such a classifier is deployed, it is necessary to freeze its thresholds at a given operating point. This point is chosen based on the ROC curve and is a trade-off between false positive errors and false negative errors. For the quality control setup, we consider two operating points: one that produces the smallest possible error rate, the other that keeps the error rate the same but maximally reduces the number of tissues that have to be re-examined by a pathologist. Figure 5 shows the plot of the old/new error ratio (given by

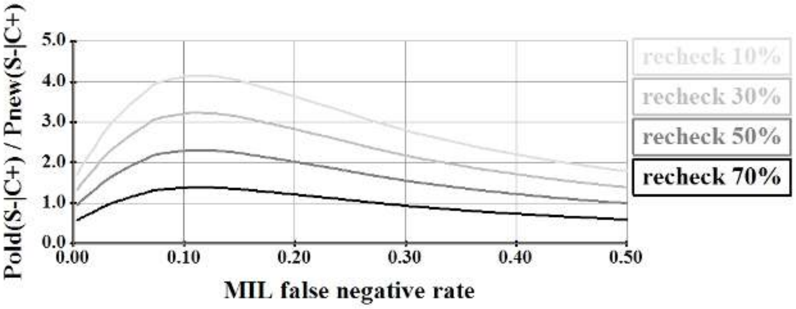

Figure 5: error analysis in a quality control scenario. The error ratio of the old to the new system is plotted as a function of the MIL classifier's operating point (false negative rate).

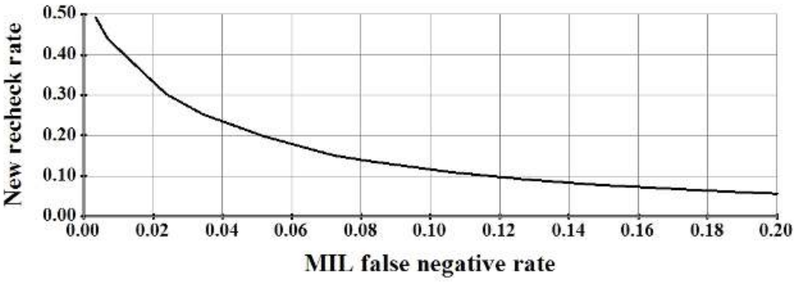

Figure 6: error analysis in a quality control scenario. The new recheck rate is plotted as a function of the MIL classifier's operating point (false negative rate).

equations 8 and 9 respectively) as a function of the machine false negative rate fnr_m. At $fnr_m = 0.109$, the curve reaches its optimal operating point that maximizes the benefit of using the new system. Four curves are shown for different values of the recheck rate r (eq. 8). Assuming that human pathologists make type I and type II error at a similar rate, varying that rate has little effect on the curves, as long as it is kept small, below 1%. We observe that for rates of recheck lower than 70%, the new system substantially improves upon the old system. For example, at a 30% recheck rate, the error

Table 2: Operating point for quality control. The new recheck rate is indicated as well as the avergae agreement between the MIL classifier and the two pathologists that provided groun-truth labels for the tissues.

TP	FP	TN	FN	FNR	FPR	recheck	AC1
260	451	3680	31	10.9%	10.6%	10.9%	87%

Table 3: Operating point for pre-screening. The classifier screens the negative samples (TN+FN) leaving only half of them to be examined by a pathologist at a cost of less than 1% missed cancers.

TP	FP	TN	FN	FNR	FPR	savings
289	1930	2201	2	0.6%	46%	50%

rate is reduced by more than 3 times.

Figure 6 plots the recheck rate of the new system given by eq. 10. At the optimal point found on figure 5, we obtain a new recheck rate of about 10%. The actual operating points are summarized in table 2.

Pre-screening of high-confidence negative cases can also be achieved with the MIL classifier. The ROC curve from figure 4 shows that very low false negative rates (< 1%) are achieved for false positive rate above 40%. One such actual operating point is given by table 3 and shows that after the MIL classifier screens negative tissues, only about half the tissue need to be examined by a pathologists (2219 of 4422).

4. DISCUSSION

We have presented a machine-learning-based computer system capable of detecting cancer on H&E-stained gastric tissue samples. We trained and tested the system on a real large scale dataset of over 30’000 tissue collected from a 2 month run at a Japanese lab and showed a performance level sufficient to justify its use in a clinical workflow for pre-screening or quality control. We contrasted two machine-learning approaches; one, using the classical supervised learning framework requires the manual labeling of a large number of regions on the tissue images, the other, using multiple-instance learning relaxes that requirement allowing large-scale training of classifiers with only tissue-level labels. We showed that the multiple-instance learning approach matches the classical supervised approach, while avoiding the need for expensive region-level labels.

The accuracy of the system (90% specificity for a sensitivity of 90%) makes it possible to be used effectively in a lab setting for quality control. In existing settings, hospitals or labs typically perform quality control by randomly sampling a certain percentage of the diagnosed cases, sending them for re-inspection to a second pathologist. With our system, two competing goals may be achieved. One is to reduce the workload of pathologists assigned to quality control, the other is to reduce the number of missed cancers. Both goals cannot be maximized at the same time and the trade-off is matter of choice by the institution. When reducing the workload of pathologists assigned for re-inspection and keeping the error the same, our system reduces the re-inspection load 5-fold, from 50% to 10%. Using the system’s optimal operating point of 10% false negative and 90% true positive and comparing to a recheck rate of 30%, the pathologist' workload is reduced to 20%, while the error rate is cut by more than half.

Our system has applicability beyond that of quality control or pre-screening. We are researching ways to adapt it to a an interactive setup, where strong classifiers coupled with an advanced graphical user interface would improve the efficiency of pathologists in their daily examination of slides. For example, as part of the lab preparation of slides, such as system would pre-analyze tissues and, as the pathologist later prepares the diagnosis, he or she would be presented with objective measurements of cancerous features, such as size of nuclei, loss of polarity in glands, and eventually a full histological grading system. Regions of interest may also be preselected by the computer for examination by the pathologist saving her precious time.

We have also started to modify our system so that it can be applied to other types of cancer. While our nuclei-level analysis can be easily transfered to other cancer types (measuring the size of a nuclei can be done in the same manner in gastric samples as in breast samples), structural elements (glands) tend to exhibit different shapes and forms and prevalent features of malignancy vary quite widely across different types of cancer. Therefore exisiting features may have to be adjusted or new ones developed to adapt the module to other types of cancer. We are actively investigating colorectal, breast and prostate cancers . In breast cancer, for example, pathological factors such as lymph node status, tumor size, histological type and histological grade are the most useful prognostic factors [Fitzgibbons et al., 2000; Mansour et al., 1994].

References

Abe T, Yamaguchi M, Murakami Y, Ohyama N, Yagi Y. Color correction of pathological images for different staining-condition slides. In: Enterprise Networking and Computing in Healthcare Industry. 2004. p. 218 –23.

Collobert R. Large Scale Machine Learning 2004;.

Datar M, Padfield D, Cline H. Color and texture based segmentation of molecular pathology images using hsoms. In: IEEE International Symposium on Biomedical Imaging. 2008. p. 292–5.

Dietterich T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 1997;89(1-2):31–71.

Dundar M, Badve S, Raykar V, Jain R, Sertel O, Gurcan M. A multiple instance learning approach toward optimal classification of pathology slides. In: International Conference on Pattern Recognition. IEEE; 2010. p. 2732–5.

Fitzgibbons P, Page D, Weaver D, Thor A, Allred D, Clark G, Ruby S, O'Malley F, Simpson J, Connolly J, et al. Prognostic factors in breast cancer. Archives of pathology & laboratory medicine 2000;124(7):966–78.

JGCA . Japanese gastric cancer association: Japanese classification of gastric carcinoma. Gastric cancer 1998;1(1):10–24.

LeCun Y, Bottou L, Orr G, Müller K. Efficient backprop. Neural networks: Tricks of the trade 1998;:546–7.

Mansour E, Ravdin P, Dressier L. Prognostic factors in early breast carcinoma. Cancer 1994;74(S1):381–400.

Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. In: Computer Vision and Pattern Recognition. volume 2; 2003. p. II–257.

Monaco J, Tomaszewski J, Feldman M, Hagemann I, Moradi M, Mousavi P, Boag A, Davidson C, Abolmaesumi P, Madabhushi A. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models. Medical Image Analysis 2010;14(4):617–29.

Rabinovich A, Agarwal S, Laris C, Price J, Belongie S. Unsupervised color decomposition of histologically stained tissue samples. Advances in Neural Information Processing Systems 2003;.

Ronco G, Vineis C, Montanari G, Orlassino R, Parisio F, Arnaud S, Berardengo E, Fabbrini T, Segnan N. Impact of the AutoPap (currently Focalpoint) primary screening system location guide use on interpretation time and diagnosis. Cancer Cytopathology) 2003;99(2).

Rumelhart D, Hintont G, Williams R. Learning representations by back-propagating errors. Nature 1986;323(6088):533–6.

Vassilakos P, Carrel S, Petignat P, Boulvain M, Campana A. Use of automated primary screening on liquid-based, thin-layer preparations. Acta cytologica 2002;46(2):291–5.

Viola P, Platt J, Zhang C. Multiple instance boosting for object detection. In: Advances in Neural Information Processing Systems. 2005. .

Xu Y, Zhu J, Chang E, Tu Z. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In: International Conference on Pattern Recognition. 2012. .

Zhang C, Chen X, Chen M, Chen SC, Shyu ML. A multiple instance learning approach for content based image retrieval using one-class support vector machine. In: Proc. of IEEE International Conference on Multimedia & Expo (ICME). 2005. .