Classification and Immunohistochemical Scoring of Breast Tissue Microarray Spots

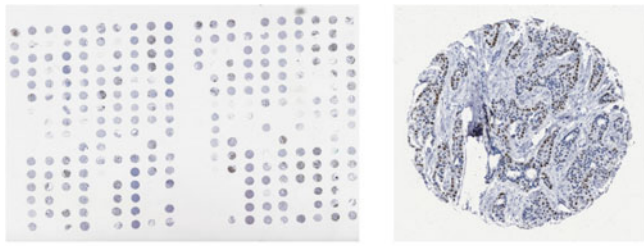
Telmo Amaral, Stephen J. McKenna*, Katherine Robertson, and Alastair Thompson

***Abstract*—Tissue microarrays (TMAs) facilitate the survey of very large numbers of tumors. However, the manual assessment of stained TMA sections constitutes a bottleneck in the pathologist's work flow. This paper presents a computational pipeline for automatically classifying and scoring breast cancer TMA spots that have been subjected to nuclear immunostaining. Spots are classified based on a bag of visual words approach. Immunohistochemical scoring is performed by computing spot features reflecting the proportion of epithelial nuclei that are stained and the strength of that staining. These are then mapped onto an ordinal scale used by pathologists. Multilayer perceptron classifiers are compared with latent topic models and support vector machines for spot classification, and with Gaussian process ordinal regression and linear models for scoring. Intraobserver variation is also reported. The use of posterior entropy to identify uncertain cases is demonstrated. Evaluation is performed using TMA images stained for progesterone receptor.**

***Index Terms*—Breast cancer, image analysis, immunohistochemical scoring, tissue microarrays (TMAs).**

I. Introduction

TISSUE microarray (TMA) technology facilitates high-throughput analysis of tissue specimens stained for one or more biological markers [1]. It is now extensively utilized in the study of cancers. Fig. 1 shows a breast TMA slide and an individual spot. At present, staining, capture, organization, and display of high-resolution composite TMA images are largely automated processes. However, analysis of the content of individual TMA spots remains laborious and time-consuming, constituting a bottleneck in pathology work flow. Furthermore, such analysis is subjective, exhibiting significant intra- and interobserver variability. Therefore, there is strong clinical and research-related motivation for the development of improved automated or semiautomated methods for quantitative analysis of TMA image data. This paper focuses on automated immunohistochemical assessment of breast cancer TMAs. Before outlining the scope of the rest of the paper, we describe TMA preparation, immunohistochemical staining, and analysis in some more detail.



(a) (b)

Fig. 1. (a) Breast TMA slide and (b) an individual spot.

A. TMAs and Immunohistochemistry

To create TMAs, pathologists typically identify up to six sites of interest on each *donor* block of formalin-fixed, wax-embedded tissue. Cylindrical biopsies, named *cores*, are then extracted from the identified sites in each donor block and inserted into a *recipient* wax block. This process is repeated for multiple donor blocks, in such a way that cores of known provenance are placed alongside each other in a mapped grid. The result is a grid arrangement of cores in the (single) recipient TMA block. Typical cores range from 2 to 4 mm in length and have a diameter of 0.6 mm. Sections of the TMA block, 4 to 8 μm in thickness, are then cut and mounted on microscope slides. Thus, each cylindrical core of tissue from the TMA block gives rise to disks of tissue referred to as *spots*.

Immunohistochemistry (IHC) is used to assess protein expression in TMA slides by staining them with a small aliquot of antibody. For example, antibodies directed against progesterone receptor (PR) can be used to detect nuclear expression of that antigen in epithelial cells of breast tumors. TMA slides are also counterstained, typically with haematoxylin, to render visible immunonegative structures. In Fig. 1(b), PR expression can be seen as brown staining, and immunonegative structures are visible as blue counterstaining. Cores in a TMA block typically originate from different patients. Thus, a single TMA slide can be used to test a given biological marker on tumors from multiple patients, whereas sequential slides cut from the same TMA block can be used to test multiple markers on the same specimens. Camp *et al*. [2] concluded that two TMA cores per patient

Manuscript received November 19, 2012; revised March 14, 2013 and April 26, 2013; accepted May 17, 2013. Date of publication May 24, 2013; date of current version September 14, 2013. This work was supported by the Breast Cancer Research Trust and by the Chief Scientist Office, U.K. under Grant CZB/4/761. *Asterisk indicates corresponding author.*

T. Amaral was with the School of Computing, University of Dundee, Dundee City, DD1 4HN, U.K. He is now with the Institute of Biomedical Engineering, University of Porto, Porto 4099-002, Portugal (e-mail: telmoamaral@sapo.pt).

*S. J. McKenna is with the School of Computing, University of Dundee, Dundee City, DD1 4HN, U.K. (e-mail: stephen@computing.dundee.ac.uk).

K. Robertson was with the Pathology and Neuroscience Department, Ninewells Hospital, Dundee, DD1 9SY, U.K. She is now with the Pathology Department, Royal Infirmary of Edinburgh, Edinburgh, EH16 4SA, U.K. (e-mail: katherine.robertson@nhs.net).

A. Thompson is with the Dundee Cancer Centre, University of Dundee, Dundee City, DD1 4HN, U.K. (e-mail: a.m.thompson@dundee.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2013.2264871

were sufficient to assess the expression of estrogen receptor and PR in invasive breast carcinoma specimens.

B. Spot Analysis

Cores in a TMA block are not homogeneous; for example, there may be a region of tumor in the top half of a core, while the bottom half contains only stroma. Therefore, pathology work flow typically includes a step in which each spot is labeled as to the type of tissue present. For example, a spot may be labeled as one of tumor, normal, stroma, fat, blood, or invalid (spot not present or not assessable). The first four of these are the most frequent. Note that the purpose here is not to segment tumor regions but rather to label spots in their entirety. Spots labeled as stroma or fat should not contain epithelial tissue. In spots labeled as tumor, at least some tumor tissue should be present, whereas epithelial tissue in spots labeled as normal should be benign.

After spot labeling, the degree of biological marker expression can be assessed in spots of interest (i.e., those labeled as containing tumor or benign epithelial tissue). This assessment is done in terms of estimated proportion of cells staining for the protein of interest and in terms of staining strength, resulting in a compound score for each spot. Several scoring methods are in use.[1] This study uses the Quickscore method [3], which does not require the pathologist to explicitly count cells, unlike the H-score [4]. Quickscore involves estimation of two ordinal quantities. The proportion of immunopositive (stained) nuclei within the tissue section is scored in the range 1–6, corresponding to proportions of 0–4%, 5–19%, 20–39%, 40–59%, 60–79%, and 80–100%, respectively. The average strength of staining is scored in the range 0–3, corresponding to negative (no staining), weak, intermediate, and strong staining, respectively.

C. Scope and Contributions

This paper presents a computational pipeline for analysis of breast cancer TMA spots that have been subjected to nuclear immunostaining, as well as an experimental evaluation using a set of PR-stained spots. Section III presents an overview of the proposed pipeline before describing its component parts. A first stage in the pipeline involves automatic classification of stained spots as belonging to the four categories tumor, normal, stroma, and fat. Spot classification using generative topic models (namely based on latent Dirichlet allocation) is compared to classifiers trained to compute estimates of class posterior probabilities directly [multilayer perceptrons (MLPs) and support vector machines (SVMs)]. The use of class posterior probabilities to assign confidence measures to decisions, enabling ambiguous spots to be flagged, is also described. Having identified spots likely to contain tumor and normal epithelial tissues, these can then be scored for IHC in the second stage of the pipeline. A method for computing continuous-valued image features analogous to the Quickscore quantities is described. These features are based on probabilistic labeling of pixels. An ordinal regression model (implemented using a Gaussian process model) is used to learn the nonlinear mapping from these features to the ordinal variables used by pathologists to form the Quickscore. Ordinal regression is compared to classification using linear and nonlinear neural networks.

Parts of the research presented in this paper build on earlier descriptions and experiments described in conference papers [5]–[7]. This paper proposes an overall processing pipeline, bringing these ideas together in a systematic way and extending their treatment. It also reports new quantitative results and comparisons, including an assessment of intraobserver variability.

[1]Immunohistochemical scoring should not be confused with cancer histological grading, which does not aim at scoring the reaction of breast tissue sections to IHC.

II. Related Work

Gurcan *et al*. [8] provided a review of histopathology image analysis more generally. Such a review is beyond the scope of this paper which instead mentions work of particular relevance.

Most published studies on classification of tissue sections have focused on discriminating between tumor and benign tissue (see, e.g., [9] and [10]) or between different types of tumor (see, e.g., [11] and [12]). This literature deals largely with tissue sections stained with haematoxylin and eosin (H&E), as opposed to sections subjected to some form of IHC. For example, Brook *et al*. [10] classified H&E stained tissue sections as benign, tumor *in situ*, or invasive carcinoma, based on histograms computed from a level sets representation. It should be noted that the presence of immunostaining does not necessarily help to distinguish benign tissue from tumor, given that both can exhibit IHC staining; the staining simply identifies an antigen that can be present in both normal and tumor cells.

Methods have been reported that estimate measures such as the proportion of nuclei that are immunopositive within tissue sections subjected to some form of nuclear immunostaining (see, e.g., [13]–[15]). However, typically, such methods do not involve any form of learning. Rather, features that characterize predetected structures within the tissue section are summarized by means of simple formulas. In most cases, no attempt is made to map the obtained measures onto actual discrete scores used by pathologists. For example, Kostopoulos *et al*. [13] analyzed breast carcinoma sections immunostained with diaminobenzidine to determine the percentage of epithelial nuclei that were stained. This value was computed from the previous segmentation and classification of epithelial nuclei, and allowed predicting the ER status of the tissue section (positive if the percentage was above 20%). Sont *et al*. [16] assessed inflammatory cell counts and cytokine expression in immunostained sections of bronchial tissue. Sanders *et al*. [17] scored immunostained head, neck, and prostate TMA images based on color statistics, thresholding, and locating brown and blue objects. Their use of pixel-level statistics to derive score values has a passing similarity to aspects of the formalized scores proposed below. Unfortunately, their method was not presented in reproducible form (e.g., “brown” and “blue” were undefined).

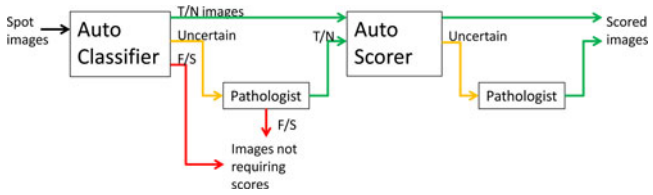


Fig. 2. Overview of the processing pipeline.

III. Method

A. Overview

Fig. 2 shows an overview of the proposed processing pipeline. Spots are first classified automatically. The classifier reports its uncertainty so that any spot with a highly uncertain classification can be referred for review by a human pathologist. Spots classified as tumor (T) or normal (N) are then IHC scored automatically. The scoring method also reports uncertainty so that spots with highly uncertain scores can be referred for review by a human pathologist.

B. Classification of TMA Spots

We adopted a *bag of visual words* approach. This general approach is now widely used in computer vision, e.g., for texture classification [18] and object and scene classification [19], and has been applied to histopathology images of skin cancer by [20]. Different applications vary in the manner in which they compute local features and sample feature locations.

We used K-means clustering to learn a visual word dictionary from a training set of normalized 15-D feature vectors. Each spot was then represented by extracting local image feature vectors from it, quantizing these to visual words using nearest neighbor matching with the learned dictionary, and forming a histogram of visual word frequencies. Each local feature vector consisted of a pixel's r, g, and b color values concatenated with 12 gray-level differential invariants denoted as $d_k, k \in \{1, 2, \ldots, 12\}$. Differential invariants were computed by convolving with a set of first- and second-order 2-D Gaussian derivative kernels at three scales through use of a Gaussian pyramid [21]. The results of these convolutions were then combined to obtain four differential invariants at each pixel location of the type proposed by Schmid and Mohr [22]. These features were chosen partly because they were invariant to image rotation, the orientation of a TMA spot being arbitrary. Gaussian derivative kernels had standard deviations of 8 pixels, and thus effectively 16 and 32 pixels at the second and third scales, respectively. Thus, they encompassed parts of nuclei, whole nuclei, and nuclei along with their immediate surroundings, respectively, given that the average epithelial nuclear radius was approximately 16 pixels. Kernels were three standard deviations in radius.

We previously reported a smaller spot classification experiment in which MLPs were found to give greater classification accuracy than either generalized linear models (GLMs) or nearest neighbor classifiers [5]. Here instead we compare spot classification results obtained using MLPs, latent Dirichlet allocation models (LDALs), and SVMs.

LDAL is a generative probabilistic model for discrete data [23], namely data such that each data instance can be represented as a bag of *codewords* belonging to a limited dictionary. The main assumption behind LDAL is that there is a given number of underlying *topics* associated with the data universe, where each topic is characterized by a distribution over codewords. It is further assumed that each data instance could be generated by, first, randomly choosing a distribution over topics, then, for each codeword needed in the data instance, sampling a topic from the chosen distribution over topics, and sampling the actual codeword from the distribution over codewords associated with the chosen topic. The codeword distributions that characterize the latent topics constitute the main parameters of the model, which need to be learned from training data.

Although commonly associated with the modeling of text collections, LDAL has applications to many types of data. In our case, visual words can be used as codewords and, because the order of codewords is irrelevant in LDAL, a visual word histogram is a sufficient representation of each data instance, i.e., each TMA spot. A motivation for using LDAL is that the learned visual topics provide an intermediate level description (between local image features and spot class label) that is potentially useful for other tissue analysis tasks. Being a latent description, topics do not have predefined semantics nor do they need to be annotated by pathologists.

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ denote a dataset where $\mathbf{x}_n = [x_1, \ldots, x_D]^{\mathrm{T}}$ denotes the vector of D visual word instances sampled from the nth spot. Let K denote the number of visual words in the dictionary. The visual word histogram computed from the nth spot has K bins and is a sufficient representation of $\mathbf{x}_n$. The topic mixture for a particular spot is defined by the frequencies of occurrence of the topics; thus, it has as many components as the chosen number of topics, Z. Topic mixtures are modeled as samples from a Dirichlet distribution, assumed because it has properties that facilitate the development of parameter estimation and inference algorithms. The parameters of an LDAL model that need to be estimated are denoted as α and β. The vector α consists of Z nonnegative elements that parametrize the Dirichlet distribution. The matrix β contains $Z \times K$ elements, storing one conditional visual word distribution $P(x|z)$ for each unique visual topic z. These parameters can be estimated via an alternating variational expectation maximization procedure [23]. We trained an LDAL model, M_c, for each of the four spot classes ($c \in \{1, 2, 3, 4\}$). Given the histogram of a test TMA spot, $\mathbf{x}$, the variational inference algorithm proposed by Blei *et al*. [23] was used to obtain a lower bound on $\log(P(\mathbf{x}|M_c))$. The spot was classified as the class associated with the largest lower bound.

Equal costs were assumed for all types of classification error, as we had no solid basis for the definition of an appropriate cost matrix (a problem that is far from trivial in this type of application).

C. Scoring of TMA Spots

Quickscoring [3] involves estimation of two ordinal values reflecting the perceived *proportion* of epithelial nuclei that are

immunopositive and the perceived *strength* of staining. Valuable information could be lost in this quantization process; nevertheless, pathologists are trained to base their decisions on such values. This was, therefore, the type of output required. Let q_p and q_s denote the proportion and strength values, respectively. Computation of features analogous to these values was based on probabilistic pixel labeling.

Manually annotated subregions of TMA training spots were used to estimate class-conditional probability distributions of local features for three classes of pixel, namely background (B), epithelial immunonegative (E_N), and epithelial immunopositive (E_P). We used the same local features as in the spot classification experiment, namely, r, g, and b color values and 12 gray-level differential invariants denoted as $d_k, k \in \{1, 2, \ldots, 12\}$. Denoting the pixel class as $v \in \{B, E_N, E_P\}$, the estimated distributions can be expressed as $P(r, g, b|v)$ for color features and $P(d_k|v)$ for each differential invariant feature d_k. Each of these distributions was estimated as a histogram. Although color components were considered interdependent, differential invariants were assumed to be conditionally independent given the pixel class. Thus, the likelihood function factored as $P(r, g, b|v) \prod_k P(d_k|v)$. A prior $P(v)$ over pixel classes was estimated from the frequencies of pixels belonging to each class observed in the annotated training data. Given a new image, posterior probabilities for each pixel n were estimated as in (1), where $\mathbf{u}^n$ denotes the local feature vector at that pixel, i.e., $\mathbf{u}^n = (r, g, b, d_1, \ldots, d_{12})^{\mathrm{T}}$:

$$P(v|\mathbf{u}^n) = \frac{P(v)P(r, g, b|v) \prod_k P(d_k|v)}{P(\mathbf{u}^n)} \tag{1}$$

Each pixel was labeled as belonging to the class with the highest posterior. A feature x_p, analogous to q_p, was computed as the number of pixels labeled as E_P divided by the total number of pixels labeled as epithelial (both E_N and E_P). This is shown in (2), where t_P^n is 1 if pixel n was labeled as E_P and 0 otherwise, and t_N^n is 1 if pixel n was labeled as E_N and 0 otherwise. A second feature, x_s, analogous to q_s, was computed as the average of the posterior probabilities for the E_P class, computed over all pixels assigned to that class, as shown in (3):

$$x_p = \frac{\sum_n t_P^n}{\sum_n (t_N^n + t_P^n)} \tag{2}$$

$$x_s = \frac{\sum_n P(v = E_P|\mathbf{u}^n) t_P^n}{\sum_n t_P^n}. \tag{3}$$

Predicting q_p and q_s from $(x_p, x_s)^{\mathrm{T}}$ is an ordinal regression task. We investigated the use of a state-of-the-art Gaussian process ordinal regression (GPOR) model [24]. Any Gaussian process is associated with a *random function* $f(\mathbf{x})$ such that, at any given point $\mathbf{x}'$, $f(\mathbf{x}')$ is a random variable that follows a Gaussian distribution [25]. In GPOR, the real codomain of $f(\mathbf{x})$ is divided into a series of contiguous intervals, which map real values of $f(\mathbf{x})$ into targets $t \in \{1, 2, \ldots, C\}$ while enforcing the ordering constraint. The boundaries between ordinal intervals are not assumed to be equally spaced. The prior distribution of $f(\mathbf{x})$ can be fully specified by the covariance matrix for the finite set of zero-mean random variables $\{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)\}$ associated with a training set of input vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. In turn, the covariance between any two such variables can be defined by Mercer kernel functions [26], [27], such as linear and Gaussian kernels. The parameters of the GPOR model include the ordinal interval boundaries, as well as constants defining the kernel prior. These parameters can be learned using maximum *a posteriori* (MAP) estimation or expectation propagation (EP). Given a test input $\mathbf{x}$, the predicted target can be inferred as that associated with the largest partial integral of the posterior distribution of $f(\mathbf{x})$ associated with the test input. GPOR was used to predict q_p and q_s from formalized scores $\mathbf{x} = (x_p, x_s)^{\mathrm{T}}$. In addition, classifiers were used to predict the Quickscore values in order to provide performance comparisons. These were a GLM and nonlinear MLP classifiers.

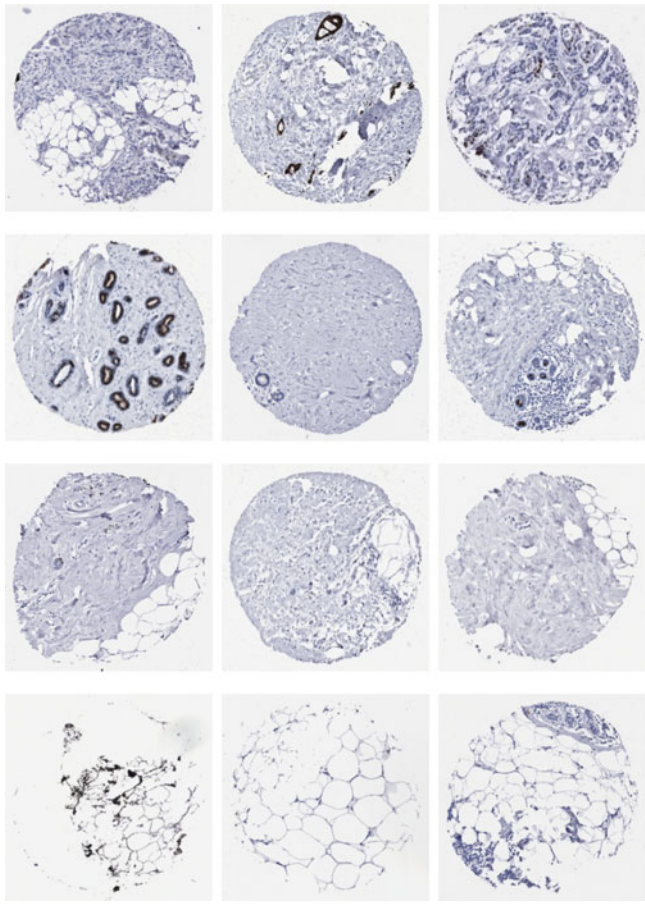


Fig. 3. Examples of breast TMA spots in each of the four categories considered. From top row to bottom row: tumor, normal, stroma, and fat.

IV. Experiments

Breast TMA spots subjected to PR IHC were used for evaluation. A total of 364 spots were randomly selected from four TMAs with the constraint that the data be approximately balanced across the four spot categories. These four TMAs contained spots originating from 112 different patients. Fig. 3 shows examples, illustrating the considerable inter- and intraclass variability. Each row shows spots from one of the four categories. An experienced pathologist, observing glass slides under a

microscope, classified all the spots and assigned Quickscores to tumor and normal spots. These assessments were fully carried out in one work session and repeated during a second session for the purpose of analyzing intraobserver variability.

A set of 40 spots (15 tumor, 15 normal, 5 stroma, and 5 fat) was used exclusively to train the visual word dictionary. Specifically, a dictionary of 160 words was learned from 400 000 feature vectors sampled randomly from these spots. For each of the remaining 324 spots, a visual word histogram was computed, based on a random sample of up to 400 000 local feature vectors per spot. Classification experiments were performed using tenfold cross validation on those 324 spots. Cross validation was repeated nine times when randomly initialized models were used (MLPs and LDAL).

On each of 20 spots, chosen at random from the data set, a circular subregion 500 pixels in diameter was randomly selected and all epithelial nuclei within that subregion were manually segmented and labeled as either immunonegative or immunopositive. Fig. 8 shows an example of an annotated subregion (although only the contouring of the nuclei is shown, not their labeling). The left third of this circular subregion is populated with epithelial cells, both stained and nonstained, whereas the remainder contains connective tissue. In total, approximately 700 epithelial nuclei were annotated in this way and these were subsequently reviewed by a pathologist. These annotated data were used to estimate the likelihood functions in (1). Formalized scores (2) and (3) were computed for 190 spots that had been classified as either tumor or normal in the first assessment session undertaken by the pathologist. Leave-one-out scoring experiments were performed so that at each fold 189 spots were used for training.

GLM and MLP classifiers with softmax functions were trained using conjugate gradients MAP optimization such that outputs were interpretable as posterior probability estimates. MLPs had three hidden units and their weight decay regularization constant was set to 0.1. The number of latent topics in each LDAL model was 60. Two parameter learning methods were compared for GPOR: MAP and EP. Furthermore, linear and Gaussian kernels were compared. SVMs used a radial basis function kernel. Their error penalty parameter C and kernel parameter γ were selected via fivefold cross validation over the training data.

We used the Netlab [28] implementations of k-means, GLM, and MLP, the C implementations of LDAL by Blei *et al*. [23] and of GPOR by Chu and Ghahramani [24], and LIBSVM [29]. Further code was implemented in MATLAB.

V. Results

Fig. 4 shows examples of spots for which the MLP agreed with the pathologist. Posterior probabilities computed by the classifier are shown for tumor (T), normal (N), stroma (S), and fat (F). Fig. 5 shows examples for which the classifier disagreed with the pathologist. Epithelial cells in the lower left region of the spot in Fig. 5(a) are unusually far apart, which may explain the low probability for class T. The spot in Fig. 5(b) contains relatively scattered epithelials which may have caused the probability of class T to be high. In the spot in Fig. 5(c), large regions of stroma and fat boosted the posterior probabilities of classes S and F, when what counted for the pathologist was the small portion of normal tissue in the upper right. The scattered but nonepithelial (inflammatory) cells in the spot in Fig. 5(d) seem to have been

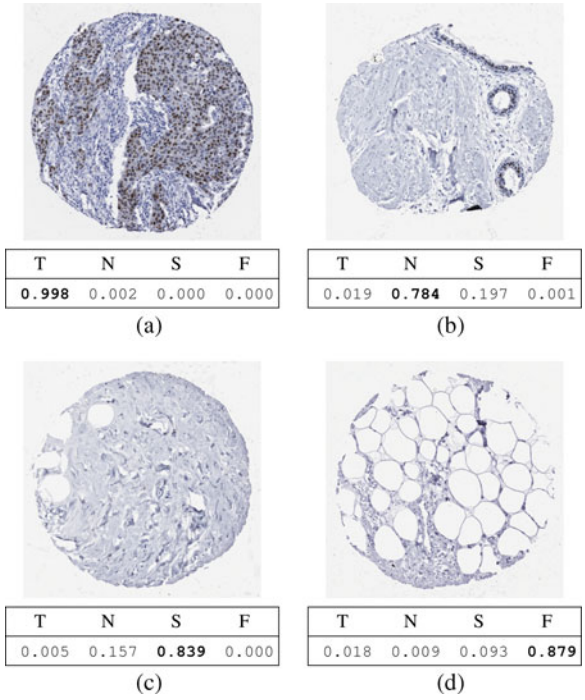


Fig. 4. Examples of spots correctly classified using MLPs, and corresponding softmax values for the four classes. (a) Tumor. (b) Normal. (c) Stroma. (d) Fat.

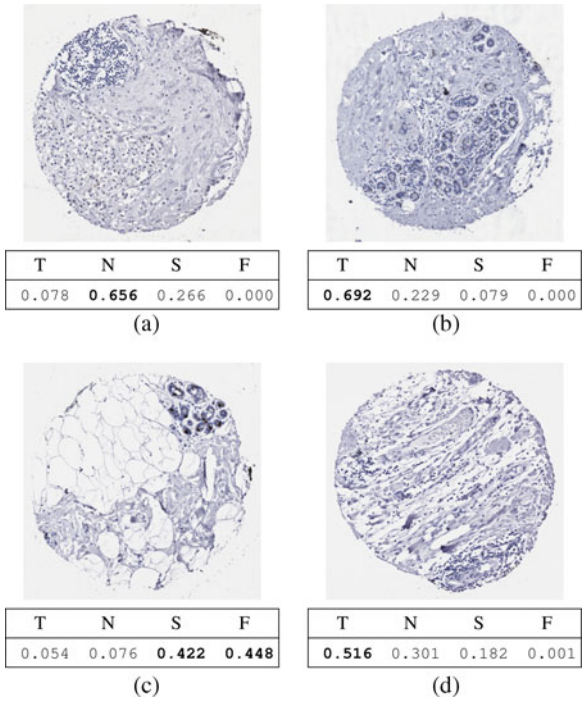


Fig. 5. Examples of spots misclassified using MLPs, and corresponding softmax values for the four classes. (a) T predicted as N. (b) N predicted as T. (c) N predicted as F. (d) S predicted as T.

TABLE I
CONFUSION MATRICES OF SPOT CLASSIFICATION EXPERIMENTS

(a) Using MLPs

Truth	Predicted			
(%)	T	N	S	F
T	79.7	6.5	11.6	2.2
N	15.7	61.0	21.9	1.4
S	12.2	5.7	73.2	8.9
F	0.0	0.0	3.8	96.2

(b) Using SVMs

Truth	Predicted			
(%)	T	N	S	F
T	81.5	5.4	10.9	2.2
N	11.4	72.9	15.7	0.0
S	9.8	2.4	79.3	8.5
F	0.0	1.2	3.8	95.0

TABLE II
CONTINGENCY TABLE OF THE INTRAOBSERVER CLASSIFICATION TRIAL

Session 1	Session 2				
	T	N	S	F	
T	82	1	4	0	87
N	9	56	2	1	68
S	3	0	66	3	72
F	1	1	3	73	78
	95	58	75	77	

misperceived as epithelial cells, leading to a very low posterior for the true class S. Table I(a) shows the confusion matrix obtained using MLPs, averaged over nine repetitions. Overall accuracy was $78.1 \pm 0.3\%$. SVMs yielded the confusion matrix shown in Table I(b) and an accuracy of 82.4%.

Of the 324 spots involved in the classification experiment, 305 were classified by the same pathologist on a second assessment session (the remaining 19 were deemed invalid). This allowed the estimation of an intraobserver agreement of 90.8%, corresponding to an unweighted Cohen's kappa coefficient of 0.877. Table II shows a contingency table for the pathologist's two classification sessions.

The entropy of the posterior distribution can be used as a measure of classification confidence; the lower the entropy, the higher the confidence. Fig. 6 shows the fractions of test spots that could be classified below different entropy thresholds using either MLPs or SVMs. Also plotted are classification accuracy and rate of misclassified tumor spots. In the case of MLPs, results were averaged over nine repetitions.

In order to compare classification based on MLPs, LDALs, and SVMs, the tenfold cross validation experiment was repeated while varying the number of spots used for training, from 10% to 100% of the 291 spots available for training at each fold. Fig. 7 plots the obtained classification accuracies. In the MLP and LDAL cases, error bars correspond to $\pm$ one standard deviation and were computed based on nine repetitions. MLP and LDAL obtained similar accuracy, whereas for larger training set sizes, SVM achieved the best results.

Scoring error was measured as the absolute difference between the predicted score and the pathologist-assigned score. Table III shows mean scoring errors and their standard deviations. Below each result, a normalized result (obtained by dividing by the number of targets) is given. Fig. 9(a) and (b) further details these results by showing the error distributions.

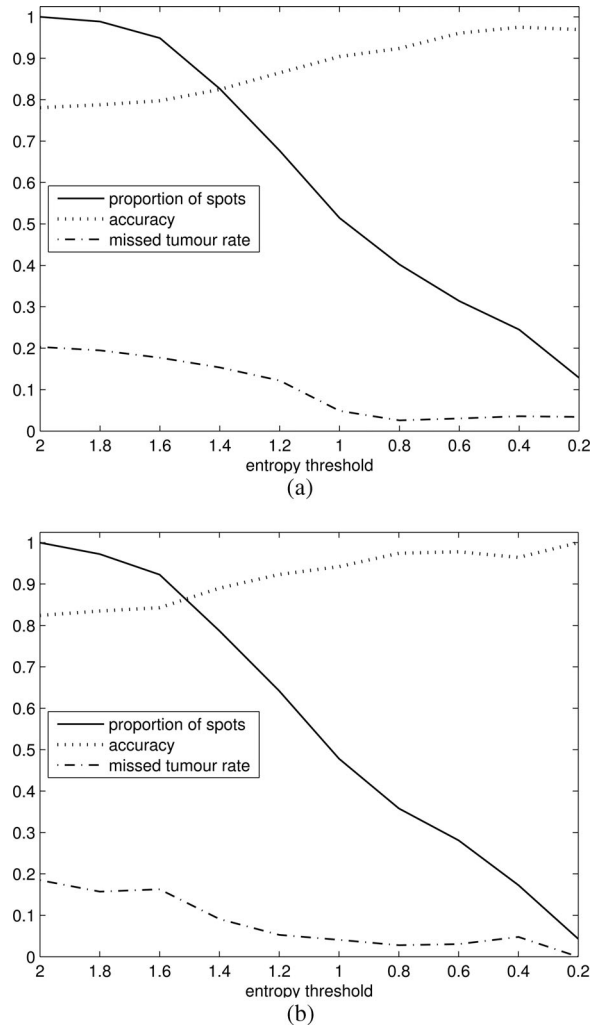


Fig. 6. Fraction of classified spots, correct classification rate, and rate of missed tumor spots, for different entropy thresholds, (a) using MLPs and (b) using SVMs.

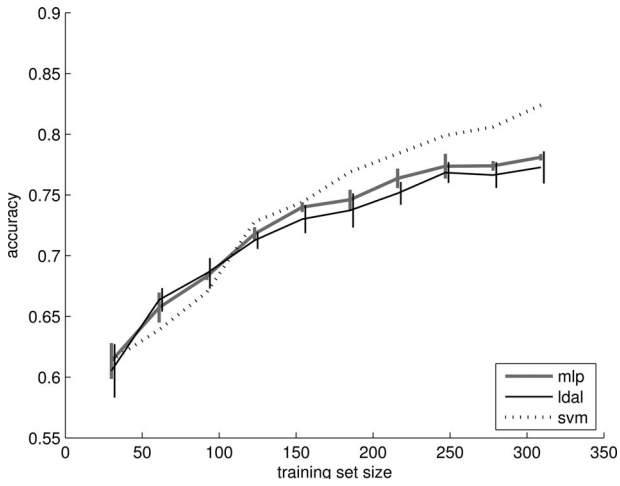


Fig. 7. Accuracy of classification using MLP, LDAL, and SVM models for different training set sizes.

TABLE III
MEAN AND STANDARD DEVIATION OF ABSOLUTE SCORING ERRORS

Predicted target	Algorithm					
	Classification		Gaussian process ordinal regression			
	GLM	MLP	MAP		EP	
			Linear	Gaussian	Linear	Gaussian
q_p	1.400 ±1.677	0.926 ±1.215	1.126 ±1.397	0.921 ±1.172	0.900 ±1.129	**0.888 ±1.175**
	0.200 ±0.240	0.132 ±0.174	0.161 ±0.200	0.132 ±0.167	0.129 ±0.161	0.127 ±0.168
q_s	0.937 ±1.097	**0.763 ±0.988**	0.937 ±1.106	0.784 ±1.003	0.800 ±1.025	0.779 ±0.994
	0.234 ±0.274	0.191 ±0.247	0.234 ±0.277	0.196 ±0.251	0.200 ±0.256	0.195 ±0.249

Normalized values are gray. Lowest errors are in bold.

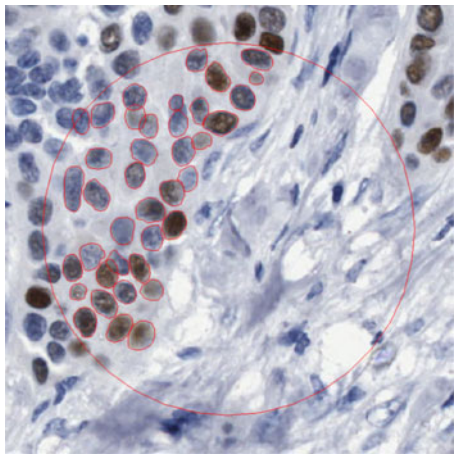


Fig. 8. Manual contouring (in red) of epithelial nuclei within a circular subregion of a TMA spot stained for PR.

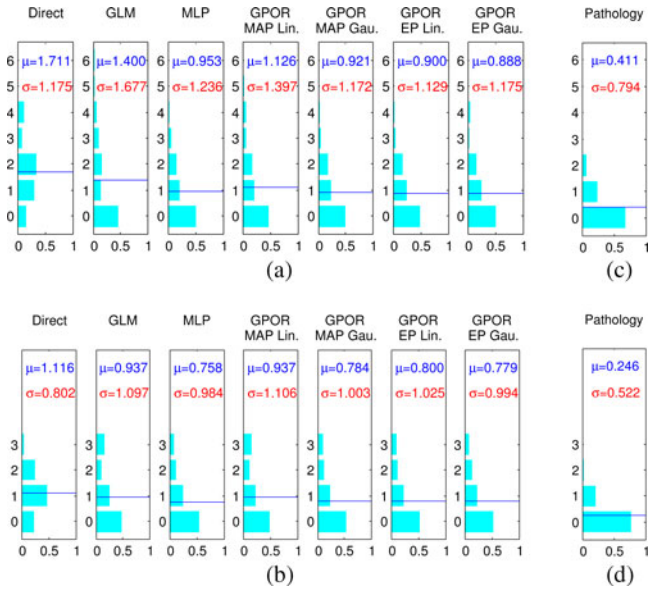


Fig. 9. (a) and (b) Distributions of q_p and q_s absolute prediction errors for each scoring approach. (c) and (d) Distributions of q_p and q_s absolute intraobserver disagreements between two scoring sessions.

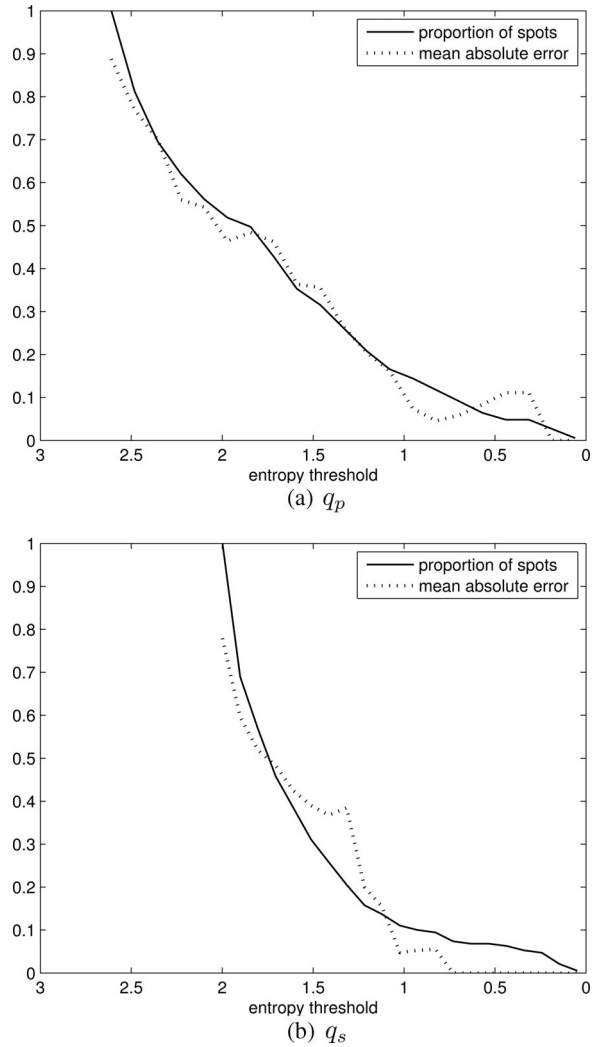


Fig. 10. Fraction of scored spots and associated mean absolute error, for different entropy thresholds and for both Quickscore values. From the experiments based on GPOR with EP and Gaussian kernels. (Lower entropy means higher confidence.)

The leftmost distributions were obtained by simply mapping the formalized scores (2) and (3) to the ordinal scores based on the Quickscore definition (see Section I-B).

The low accuracy of the GLM and direct mapping methods indicates that a linear mapping from formalized scores to ordinal Quickscores is suboptimal, supporting the use of nonlinear methods. The standard deviations associated with prediction errors make comparisons between MLP and GPOR inconclusive. GPOR with EP and Gaussian kernels obtained the lowest error for q_p and an error very close to the lowest for q_s. However, MLP performed surprisingly well given that it does not explicitly model the ordinal nature of the score variables. Furthermore, MLP was at least an order of magnitude faster at classifying each spot once trained (tenths of a second versus several seconds).

Fig. 10 shows the fraction of test spots whose q_p and q_s values can be predicted below a given entropy threshold. Also shown is the mean absolute error computed over each fraction of spots.

Of the 190 spots involved in the scoring experiment, 175 were scored by the same pathologist in a second assessment session (the remaining 15 were deemed not assessable). The average absolute interobserver disagreements were 0.411 and 0.246 and the linearly weighted Cohen's kappa coefficients were 0.832 and 0.817, for q_p and q_s, respectively. Fig. 9(c) and (d) shows the distributions of these disagreements. These intraobserver data suggest that further improvements are certainly possible. However, it should be noted that this is a tough benchmark since interobserver variability is expected to be higher than intraobserver variability [30].

VI. Conclusions and Recommendations

Spot classification accuracy was similar for MLP and LDAL classification methods, while SVMs achieved the best results. It should nevertheless be noted that, unlike SVMs, MLPs do not require an extra calibration step in order to output well-calibrated posterior probabilities [31]. The use of posterior entropy to reject spots enabled higher accuracies to be obtained for the remaining spots. Fig. 6 suggests that very low tumor misclassification rates can be achieved by setting the rejection threshold appropriately. Similarly, as the entropy threshold on scoring predictions was decreased, the mean absolute error tended to decrease [see Fig. 10(a) and (b)]. This suggests that it is possible to automatically process reasonable fractions of spots that are more unequivocal whilst identifying the more difficult spots for human assessment.

The generative topic modeling approach exemplified by LDAL is worth exploring further. In particular, sharing of topic models across classes [32] and use of hierarchical Dirichlet processes to automatically determine the number of latent topics [33] might be expected to yield improved accuracy. In LDAL models with a single layer of latent topics, the optimal number of topics tends to be considerably larger than the range of tissue types in a TMA spot. The addition of a second layer of latent topics (modeled as distributions over topics of the first layer) might, therefore, help to successfully represent different tissue types. The inclusion of additional layers of topics could also be used to model the relationship between epithelial regions and the immunopositivity of nuclei, in that both tumor and normal regions may contain both immunonegative and immunopositive nuclei.

A method for computing formalized Quickscore values from local likelihood functions was proposed and used to estimate ordinal Quickscores. The nonlinear GPOR and MLP methods gave better scoring results than did linear methods. Given the ordinal nature of the scores, GPOR might have been expected to be more accurate. Further research may be needed to take full advantage of the ordinal regression model. A possibility would be to investigate modifications that allowed us to accurately model the way in which pathologists mislabel spots, based on observer variability data. Currently, the GPOR method models noisy labeling through a single σ^2_{noise} hyperparameter. A set of variability parameters on which the user could set a prior might be more appropriate.

Spots were classified and scored without taking into account the differing costs of errors. The cost of misclassifying a tumor spot as normal will be greater than the cost of misclassifying a normal spot as a tumor spot, for example. Modeling such costs is left for future work.

The methods presented here were evaluated on spots immunostained for PR. Evaluation on other nuclear stains such as ER and tumor protein 53 (p53) is left for future work. However, analysis of these stains is probably easier than that of PR, due to their appearance.

Several commercial systems exist to assist in assessment of IHC in TMAs; published performance evaluations are limited [34], [35]. An attempt to assess the agreement between three such systems and pathologists was reported by Bolton *et al.* [36] and included PR IHC-stained breast tissue, but the manual tuning of various system parameters to the datasets used makes any direct comparison problematic.

Finally, we note that many other feature representations are possible and could be usefully compared. Errors on sparsely distributed epithelial and inflammatory tissue regions suggest that the local features should be combined with more contextual features. We are currently exploring the use of distribution-based descriptors such as intensity domain spin images [37], [38] and the use of methods for modeling spatial context for segmentation of tumor regions [38]–[40]. A system capable of automatically segmenting regions of tumor, benign tissue, fat, and stroma, and further segmenting tumor and normal regions into stained and unstained subregions, could in principle be used as a basis for both classification of TMA spots into types and their scoring.

Acknowledgment

TMA data were made available by the U.K. National Cancer Research Institute's Adjuvant Breast Cancer (ABC) Chemotherapy Trial [41].

References

[1] J. Kononen, L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. Mihatsch, G. Sauter, and O. Kallionimeni, "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Med.*, vol. 4, no. 7, pp. 844–847, 1998.

[2] R. Camp, L. Charette, and D. Rimm, "Validation of tissue microarray technology in breast carcinoma," *Lab. Invest.*, vol. 80, no. 12, pp. 1943–1949, 2000.

[3] S. Detre, G. Jotti, and M. Dowsett, "A "quickscore" method for immunohistochemical semiquantitation: Validation for oestrogen receptor in breast carcinomas," *J. Clin. Path.*, vol. 48, no. 9, pp. 876–878, 1995.

[4] K. McCarty, Jr., L. Miller, E. Cox, J. Konrath, and K. McCarty, Sr., "Estrogen receptor analyses correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies," *Arch. Pathol. Lab. Med.*, vol. 109, pp. 716–721, 1985.

[5] T. Amaral, S. J. McKenna, K. Robertson, and A. Thompson, "Classification of breast tissue microarray spots using texton histograms," in *Proc. Med. Image Understanding Anal.*, 2008, pp. 144–148.

[6] T. Amaral, S. J. McKenna, K. Robertson, and A. Thompson, "Scoring of breast tissue microarray spots through ordinal regression," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2009, vol. 2, pp. 243–248.

[7] T. Amaral, S. J. McKenna, K. Robertson, and A. Thompson, "Analysis of breast tissue microarrays using Latent Dirichlet Allocation," in *Proc.*

Opt. Tissue Image Anal. Microscopy, Histopathology Endoscopy, 2009, pp. 112–123.

[8] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng*., vol. 2, pp. 147–171, 2009.

[9] B. Karaçali and A. Tözeren, "Automated detection of regions of interest for tissue microarray experiments: An image texture analysis," *BMC Med. Imag*., vol. 7, no. 1, p. 2, 2007.

[10] A. Brook, R. El-Yaniv, E. Isler, R. Kimmel, R. Meir, and D. Peleg, "Breast cancer diagnosis from biopsy images using generic features and SVMs," Technion—Israel Institute of Technology, Kesalsaba, Israel, Tech. Rep. CS-2008-07, 2008.

[11] H. Qureshi, N. Rajpoot, R. Wilson, T. Nattkemper, and V. Hans, "Comparative analysis of discriminant wavelet packet features and raw image features for classification of meningioma subtypes," in *Proc. Med. Image Understanding Anal*., Aberystwyth, U.K., 2007, pp. 211–215.

[12] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *Proc. 5th IEEE Int. Symp. Biomed. Imag*, Paris, France, 2008, pp. 496–499.

[13] S. Kostopoulos, D. Cavouras, A. Daskalakis, P. Bougioukos, P. Georgiadis, G. Kagadis, I. Kalatzis, P. Ravazoula, and G. Nikiforidis, "Colour-texture based image analysis method for assessing the hormone receptors status in breast tissue sections," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc*., 2007, pp. 4985–4988.

[14] N. Elie, B. Plancoulaine, J. Signolle, and P. Herlin, "A simple way of quantifying immunostained cell nuclei on the whole histologic section," *Cytometry Part A*, vol. 56A, no. 1, pp. 37–45, 2003.

[15] J. Weaver and J. Au, "Comparative scoring by visual and image analysis of cells in human solid tumors labeled for proliferation markers," *Cytometry Part A*, vol. 27, no. 2, pp. 189–199, 1997.

[16] J. Sont, W. de Boer, W. van Schadewijk, K. Grünberg, J. van Krieken, P. Hiemstra, and P. Sterk, "Fully automated assessment of inflammatory cell counts and cytokine expression in bronchial tissue," *Amer. J. Respir. Crit. Care Med*., vol. 167, pp. 1496–1503, 2003.

[17] T. Sanders, T. Stokes, R. Moffitt, Q. Chaudry, R. Parry, and M. Wang, "Development of an automatic quantification method for cancer tissue microarray study," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc*., 2009, vol. 1, pp. 3665–3668.

[18] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis*., vol. 62, no. 1–2, pp. 61–81, 2005.

[19] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual categorization with bags of keypoints," in *Proc. ECCV Int. Workshop Statist. Learn. Comput. Vis*., 2004, pp. 1–22.

[20] J. Caicedo, A. Cruz, and F. González, "Histopathology image classification using bag of features and kernel functions," in *Proc. 12th Conf. Artif. Intell. Med*., 2009, pp. 126–135.

[21] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun*., vol. 31, no. 4, pp. 532–540, Apr. 1983.

[22] C. Schmid and R. Mohr, "Matching by local invariants," INRIA, Le Chesnay Cedex, France, Tech. Rep. RR-2644, 1995.

[23] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res*., vol. 3, pp. 993–1022, 2003.

[24] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *J. Mach. Learn. Res*., vol. 6, pp. 1019–1041, 2005.

[25] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[26] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA, USA: SIAM, 1990.

[27] B. Schölkopf, A. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2002.

[28] I. Nabney, *NETLAB: Algorithms for Pattern Recognition*. New York, NY, USA: Springer-Verlag, 2002.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol*., vol. 2, pp. 27:1–27:27, 2011.

[30] T. Kirkegaard, J. Edwards, S. Tovey, L. McGlynn, S. Krishna, R. Mukherjee, L. Tam, A. Munro, B. Dunne, and J. Bartlett, "Observer variation in immunohistochemical analysis of protein expression, time for a change?" *Histopathology*, vol. 48, no. 7, pp. 787–794, 2006.

[31] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. Int. Conf. Mach. Learn*., 2005, pp. 625–632.

[32] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2005, vol. 2, pp. 524–531.

[33] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc*., vol. 101, no. 476, pp. 1566–1581, 2006.

[34] S. Gokhale, D. Rosen, N. Sneige, E. Reserkova, A. Sahin, J. Liu, and C. Albarracin, "Assessment of two automated imaging systems in evaluating estrogen receptor status in breast carcinoma," *Appl. Immunohistochem. Mol. Morphol*., vol. 15, no. 4, pp. 451–455, 2007.

[35] G. Turashvili, S. Leung, D. Turbin, K. Montgomery, B. Gilks, R. West, M. Carrier, D. Huntsman, and S. Aparicio, "Inter-observer reproducibility of HER2 immunohistochemical assessment and concordance with fluorescent in situ hybridization (FISH): Pathologist assessment compared to quantitative image analysis," *BMC Cancer*, vol. 9, art. no. 165, 2009.

[36] K. Bolton, M. Garcia-Closas, R. Pfeiffer, M. Duggan, W. Howat, S. Hewitt, X. Yang, R. Cornelison, S. Anzick, P. Meltzer, S. Davis, P. Lenz, J. Figueroa, P. Pharoah, and M. Sherman, "Assessment of automated image analysis of breast cancer tissue microarrays for epidemiologic studies," *Cancer Epidemiol. Biomarkers Prev*., vol. 19, no. 4, pp. 992–999, 2010.

[37] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 27, no. 8, pp. 1265–1278, Aug. 2005.

[38] S. Akbar, T. Amaral, S. J. McKenna, L. Jordan, and A. Thompson, "Tumour segmentation in breast tissue microarray images using spin-context," in *Proc. Med. Image Understanding Anal*., Swansea, U.K., 2012, pp. 25–30.

[39] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 32, no. 10, pp. 1744–1757, Oct. 2010.

[40] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2012, pp. 964–971.

[41] Adjuvant Breast Cancer Trials Collaborative Group, "Polychemotherapy for early breast cancer: Results from the international adjuvant breast cancer chemotherapy randomized trial," *J. Nat. Cancer Inst*., vol. 99, no. 7, pp. 506–515, 2007.

Authors' photographs and biographies not available at the time of publication.