# STATISTICAL METHODS FOR TISSUE ARRAY IMAGES—ALGORITHMIC SCORING AND CO-TRAINING

BY DONGHUI YAN, PEI WANG[1], MICHAEL LINDEN, BEATRICE KNUDSEN AND TIMOTHY RANDOLPH[1]

*Fred Hutchinson Cancer Research Center, Fred Hutchinson Cancer Research Center, University of Minnesota Medical School, Cedars-Sinai Medical Center and Fred Hutchinson Cancer Research Center*

Recent advances in tissue microarray technology have allowed immuno-histochemistry to become a powerful medium-to-high throughput analysis tool, particularly for the validation of diagnostic and prognostic biomarkers. However, as study size grows, the manual evaluation of these assays becomes a prohibitive limitation; it vastly reduces throughput and greatly increases variability and expense. We propose an algorithm—Tissue Array Co-Occurrence Matrix Analysis (TACOMA)—for quantifying cellular phenotypes based on textural regularity summarized by local inter-pixel relationships. The algorithm can be easily trained for any staining pattern, is absent of sensitive tuning parameters and has the ability to report salient pixels in an image that contribute to its score. Pathologists' input via informative training patches is an important aspect of the algorithm that allows the training for any specific marker or cell type. With co-training, the error rate of TACOMA can be reduced substantially for a very small training sample (e.g., with size 30). We give theoretical insights into the success of co-training via thinning of the feature set in a high-dimensional setting when there is "sufficient" redundancy among the features. TACOMA is flexible, transparent and provides a scoring process that can be evaluated with clarity and confidence. In a study based on an estrogen receptor (ER) marker, we show that TACOMA is comparable to, or outperforms, pathologists' performance in terms of accuracy and repeatability.

**1. Introduction.** Tissue microarray (TMA) technology was first described by Wan, Fortuna and Furmanski (1987) and substantially improved by Kononen et al. (1998) as a high-throughput technology for the assessment of protein expression in tissue samples. As shown in the top panel of Figure 1, the construction of a TMA begins with cylindrical cores extracted from a donor block of formalin-fixed and paraffin-embedded tissues. The cores are transferred to the grid of the recipient block. This grid is generated by punching cylindrical holes at equal distance into a precast rectangular mold of solid paraffin wax. Once all the holes are filled with
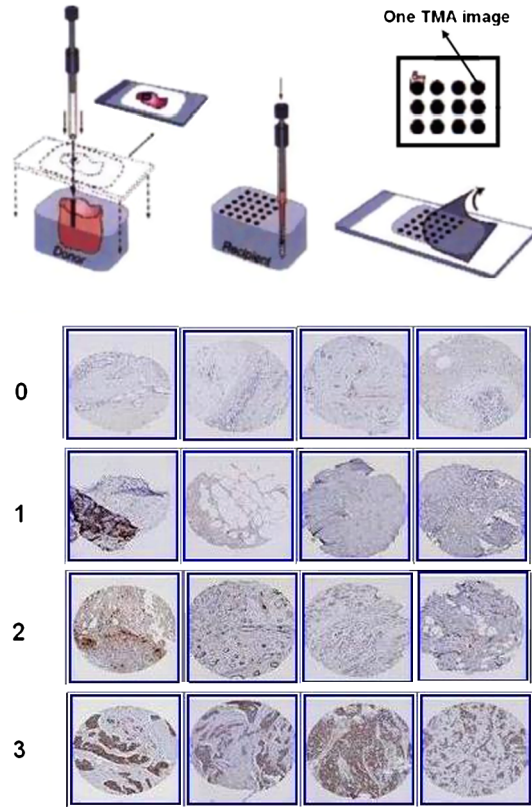
FIG. 1. *Illustration of the TMA technology and example TMA images. The top panel is reprinted from Kononen et al.* (1998) *by courtesy of the Nature Publishing Group and it shows the steps involved in how a TMA image may be produced. The bottom panel displays example TMA images with the numbers in the left indicating the score of images in the same row.*

donor cores, the block is heated to fuse the cores to the wax of the block. Normally, recipient blocks contain 360 to 480 tissue cores from donor blocks, often in triplate samples from each block and are thus called tissue micro arrays (TMA). They are sectioned transversely and each section is captured on a glass slide, such that slides display a cross section of each core in a grid-like fashion. More than 100 slides can be generated from each TMA block for analysis with a separate probe. This procedure standardizes the hybridization process of the probe across hundreds of tissue samples. The use of TMAs in cancer biology has increased dramatically in recent years [Camp, Neumeister and Rimm (2008), Giltnane and Rimm (2004), Hassan et al. (2008), Voduc, Kenney and Nielsen (2008)] for the rapid evaluation of DNA, RNA and protein expressions on large numbers of clinical tissue samples; they remain the most efficient method for validating proteomics data and tissue biomarkers. We limit our discussion to high-density immunohistochemistry (IHC)

staining, a method used for the measurement of protein expression, as the most common method for subcellular localization.

The evaluation of protein expression requires the quantification, or scoring, of a TMA image. The scores can be used for the validation of biomarkers, assessment of therapeutic targets, analysis of clinical outcome, etc. [Hassan et al. (2008)]. The bottom panel of Figure 1 gives an example of several TMA images with scores assigned at a 4-point scale (see Section 4 for details).

Although the construction of TMAs has been automated for large-scale inter-rogation of markers in tissue samples, several factors limit the use of the TMA as a high-throughput assay. These include the variability, subjectivity and time-intensive effort inherent in the visual scoring of staining patterns [Camp, Neumeis-ter and Rimm (2008), Vrolijk et al. (2003)]. Indeed, a pathologist's score relies on subjective judgments about colors, textures, intensities, densities and spatial rela-tionships. As noted in Giltnane and Rimm (2004), however, the human eye cannot provide an objective quantification that can be normalized to a reference. In gen-eral, problems stemming from the subjective and inconsistent scoring by patholo-gists are well known and have been highlighted by several studies [Bentzen, Buffa and Wilson (2008), Berger et al. (2005), DiVito and Camp (2005), Kirkegaard et al. (2006), Thomson et al. (2001), Walker (2006)]. Thus, as study size grows, the value of TMAs in a rigorous statistical analysis may actually decrease without a consistent and objective scoring process.

These concerns have motivated the recent development of a variety of tools for automated scoring, ranging from sophisticated image enhancement tools, tissue segmentation to computer-assisted pathologist-based scoring. Many are focused on a particular cellular pattern, with HER2 (exhibiting nuclear staining) being the most commonly targeted marker; see, for example, Hall et al. (2008), Joshi et al. (2007), Masmoudi et al. (2009), Skaland et al. (2008), Tawfik et al. (2005). For a survey of commercial systems, we refer to Mulrane et al. (2008) or Rojo, Bueno and Slodkowska (2009), and also the review by Cregger, Berger and Rimm (2006) which acknowledges that, given the rapid changes in this field, this information may become outdated as devices are abandoned, improved or newly developed. A property of most automated TMA scoring algorithms is that they rely on various forms of background subtraction, feature segmentation and thresholds for pixel in-tensity. Tuning of these algorithms can be difficult and may result in models sensi-tive to several variables, including staining quality, background antibody binding, counterstain intensity, and the color and hue of chromogenic reaction products used to detect antibody binding. Moreover, such algorithms typically require tun-ing from the vendors with parameters specific to the markers' staining pattern (e.g., nuclear versus cytoplasmic), or even require a dedicated person for such a system.

To address the further need for scoring of TMAs in large biomarker studies, we propose a framework—called Tissue Array Co-Occurrence Matrix Analysis (TACOMA)—that is trainable to any staining pattern or tissue type. By seeking texture-based patterns invariant in the images, TACOMA does not rely on inten-

sity thresholds, color filters, image segmentation or shape recognition. It recognizes specific staining patterns based on expert input via a preliminary set of image patches. In addition to providing a score or categorization, TACOMA allows to see which pixels in an image contribute to its score. This clearly enhances interpretability and confidence in the results.

It should be noted that TACOMA is not designed for clinical diagnosis but rather a tool for use in large clinical studies that involve a range of potential biomarkers. Since many thousands of samples may be required, the cost and time required for pathologist-based scoring may be prohibitive and so an efficient automated alternative to human scoring can be essential. TACOMA is a framework for such a purpose.

An important concern in biomedical studies is that of the limited training sample size.[2] The size of the training set may necessarily be small due to the cost, time or human efforts required to obtain them. We adopt co-training [Yarowsky (1995), Blum and Mitchell (1998)] in the context of TACOMA to substantially reduce the training sample size. We explore the thinning of the feature set for co-training when a "natural" split is not readily available but the features are fairly redundant, and this is supported by our theory that a thinned slice carries about the same classification power as the whole feature set under some conditions.

The organization of the remainder of this paper is as follows. We describe the TACOMA algorithm in Section 2, this is followed by a discussion on co-training to reduce the training sample size with some theoretical insights on the thinning scheme in Section 3. Then in Section 4, we present our experimental results. We conclude with a discussion in Section 5.

## 2. The TACOMA algorithm.

The primary challenge TACOMA addresses is the lack of easily-quantified criteria for scoring: features of interest are not localized in position or size. Moreover, within any region of relevance—one containing primarily cancer cells—there is no well-defined (quantifiable) shape that characterizes a pattern of staining (see, e.g., the bottom panel of Figure 1 for an illustration). The key insight that underlies TACOMA is that, despite the heterogeneity of TMA images, they exhibit strong statistical regularity in the form of visually observable textures or staining pattern [see, e.g., Figure 2(b)]. And, with the guidance of pathologists, TACOMA can be trained for this pattern regardless of the cancer cell type (breast, prostate, etc.) or marker type (e.g., nucleus, cytoplasmic, etc.).

TACOMA captures the texture patterns exhibited by TMA images through a matrix of counting statistics, the Gray Level Co-occurrence Matrix (GLCM). Through a small number of representative image patches, TACOMA constructs a feature mask so that the algorithm will focus on those biologically relevant features (i.e., a subset of GLCM entries). Besides scoring, TACOMA also reports salient

---

[2]The additional issue of label noise was studied elsewhere [Yan et al. (2011)].
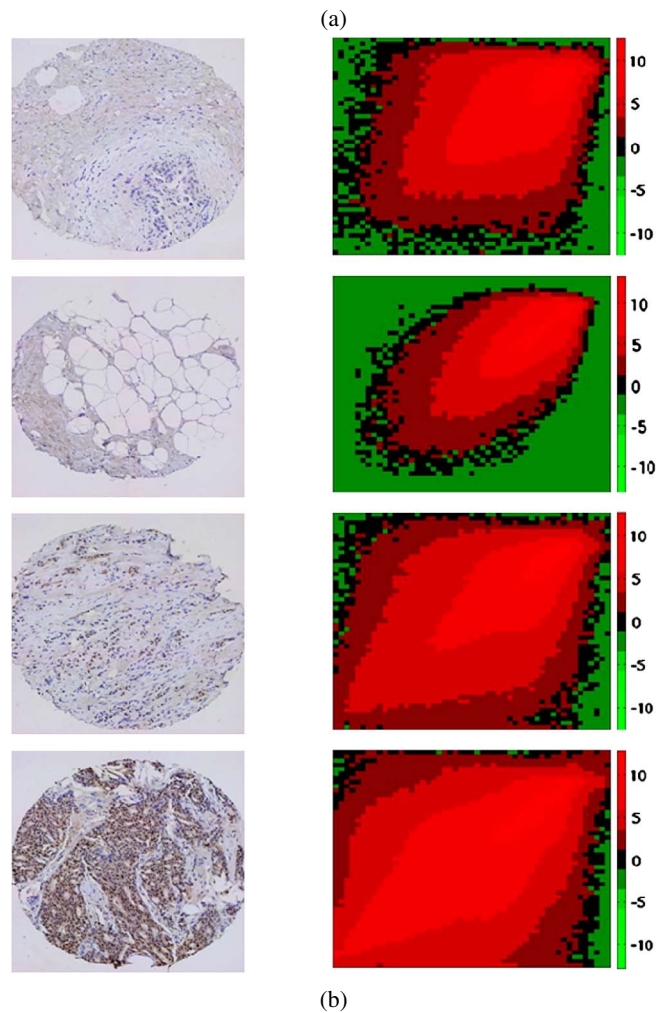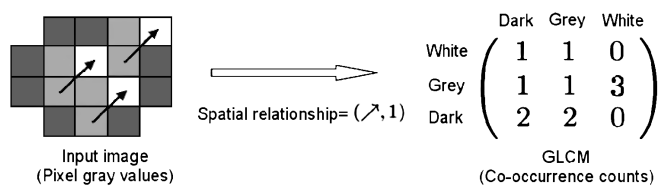
(a)



(b)

FIG. 2. *Example images and their GLCMs.* (a) *Generating the GLCM from an image. This toy "image" (left) has 3 gray levels, {Dark, Grey, White}. Here, under the spatial relationship ($\nearrow$, 1), the transition from Grey to White (indicated by $\nearrow$) occurs three times; accordingly, the entry of the GLCM corresponding to the Grey row and White column has a value of 3.* (b) *Example TMA images. Images of a tissue sample (left panel) and the Heatmap (right panel) of their GLCM (in log scale). The GLCM matrices are all $51 \times 51$; each GLCM cell in the heatmap indicates the frequency of the corresponding transition. The color scale is illustrated by the color bar on the right.*

image pixels (i.e., those contribute to the scoring of an image) which will be useful for the purpose of training, comparison of multiple TMA images, estimation of staining intensity, etc. For the rest of this section, we will briefly discuss these individual building blocks of TACOMA followed by an algorithmic description of TACOMA.

2.1. *The gray level co-occurrence matrix.* The GLCM was originally proposed by Haralick (1979) and has proven successful in a variety of remote-sensing applications [Yan, Bickel and Gong (2006)]. The GLCM, of an image, is a matrix whose entries count the frequency of transitions between pixel intensities across neighboring pixels with a particular spatial relationship; see Figure 2. The description here is essentially adopted from Yan, Bickel and Gong (2006). We start by defining the spatial relationship between a pair of pixels in an image.

DEFINITION. A spatial relationship has two elements, the direction and the distance of interaction. The set of all possible spatial relationships is defined as

$$\Re = D \otimes L$$
$$= \{\nearrow, \searrow, \nwarrow, \swarrow, \downarrow, \uparrow, \rightarrow, \leftarrow\} \otimes \{1, \ldots, d\},$$

where $D$ is the set of potential directions and $L$ is the distance of interaction between the pair of pixels involved in a spatial relationship. The distance of interaction is the minimal number of steps required to move from one pixel to the other along a given direction. The particular spatial relationships used in our application are $(\nearrow, 3)$, $(\searrow, 1)$ and $(\nearrow, 1)$. Details about the choice of these spatial relationships can be seen in Section 4.

Although the definition of spatial relationships can be extended to involve more pixels [Yan, Bickel and Gong (2006)], we have focused on pairwise relationships which appear to be sufficient. Next we define the GLCM.

DEFINITION. Let $N_g$ be the number of gray levels in an image.[3] For a given image (or a patch) and a fixed spatial relationship $\sim\, \in \Re$, the GLCM is defined as

A $N_g \times N_g$ matrix such that its $(a, b)$-entry counts the number of pairs of pixels, with gray values $a$ and $b$, respectively, having a spatial relationship $\sim$, for $a, b \in \{1, 2, \ldots, N_g\}$.

This definition is illustrated in Figure 2(a) with more realistic examples in Figure 2(b). Figure 2(b) gives a clear indication as to how the GLCM distinguishes between TMA images having different staining patterns.

---

[3]Often before the computing of GLCM, the gray level of each pixel in an image is scaled linearly from $[1, N_o]$ to $[1, N_g]$ for $N_o$ the predefined number, typically 256, of gray levels in the image.
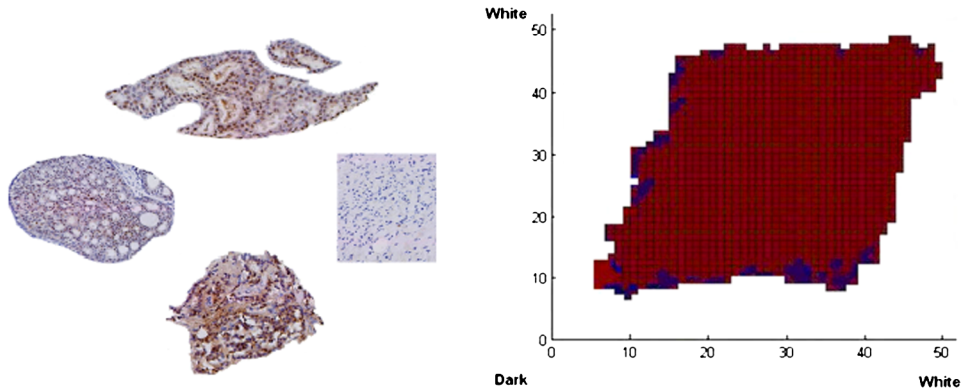
FIG. 3. *Representative image patches and the induced feature mask. Four pathologist-chosen patches* (*left panel*) *and the feature mask as determined by all patches* (*right panel*, *see algorithmic description of TACOMA*). *Nonwhite entries in this matrix indicate the corresponding GLCM entries to be used in scoring. Note that one and only one feature mask is required throughout.*

Our use of the GLCM is nonstandard in that we do not use any of the common scalar-valued summaries of a GLCM [see Haralick (1979) and Conners and Harlow (1980)], but instead employ the entire matrix (with masking) in a classification algorithm [see also Yan, Bickel and Gong (2006)]. A GLCM may have a large number of entries, typically thousands, however, the exceptional capability of Random Forests [Breiman (2001)] in feature selection allows us to directly use all (or a masked subset of) GLCM entries to determine a final score or classification.

2.2. *Image patches for domain knowledge.* In order to incorporate prior knowledge about the staining pattern, we mask the GLCM matrix so that the scoring will focus on biologically pertinent features. The masking is realized by first choosing a set of image patches representing regions that consist predominantly of cancer cells and are chosen to represent the staining patterns; see Figure 3. The collection of GLCMs from these patches are then used to define a template of "significant entries" (cf. TACOMA algorithm in Section 2.5) for all future GLCMs: when the GLCM of a new image is formed, only the entries that correspond to this template are retained. This masking step enforces the idea that features used in a classifier should not be based on stromal, arterial or other nonpertinent tissue which may exhibit nonspecific or background staining. Note that only one small set of image patches is required; these image patches are used to produce a common feature mask which is applied to all images in both training and scoring.

In this fashion, feature selection is initiated by expert biological knowledge. This manner of feature selection involves little human effort but leads to substantial gain in both interpretability and accuracy. The underlying philosophy is that no machine learning algorithms surpass domain knowledge. Since by using image patches we do not indicate which features to select but instead specify their effect,

*Performance comparison of RF, SVM and Boosting of a naive Bayes classifier. The result for RF is adopted from Section* 4.1. *For SVM, we vary the choice of the kernel from* {*Gaussian, polynomial, sigmoid*} *with the best tuning parameters for* $N_g \in \{7, 9, 13, 26, 37, 51, 64, 85\}$ *and the best result is reported. For boosting, the best result is reported by varying* $N_g \in \{7, 9, 13, 26, 37, 51, 64, 85\}$ *and the number of boosting iterations from* {1, 2, 3, 5, 10, 50, 100}

| Classifier | Accuracy |
|------------|----------|
| RF | 78.57% |
| SVM | 65.24% |
| Boosting | 61.28% |

we achieve the benefits of a manual-based feature selection but avoid its difficulty. This is a novel form of nonparametric, or implicit, feature selection which is applicable to settings beyond TMAs.

2.3. *Random forests.* TACOMA uses Random Forests (RF) [Breiman (2001)] as the underlying classifier. RF was proposed by Breiman and is considered one of the best classifiers in a high-dimensional setting [Caruana, Karampatziakis and Yessenalina (2008)]. In our experience, RF achieves significantly better performance than SVM and Boosting on the TMA images we use (see Table 1). Additionally Holmes, Kapelner and Lee (2009) argue that RF is superior to others in dealing with tissue images. The fundamental building block of RF is a tree-based classifier which can be nonstable and sensitive to noise. RF takes advantage of such instability and creates a strong ensemble by bagging a large number of trees [Breiman (2001)]. Each individual tree is grown on a bootstrap sample from the training set. For the splitting of tree nodes, RF randomly selects a number of candidate features or linear combinations of features and splits the tree node with the one that achieves the most reduction in the node impurity as defined by the Gini index [or other measures such as the out of bag (oob) estimates of generalization error] defined as follows:

$$(1) \qquad \phi(\mathbf{p}) = \sum_{i=1}^{C} p_i(1 - p_i),$$

where $\mathbf{p} = (p_1, \ldots, p_C)$ denotes the proportion of examples from different classes and $C$ is the number of different classes. RF grows each tree to the maximum and no pruning is required. For an illustration of RF, see Figure 4.

To test a future example $X$, let $X$ fall from each tree for which $X$ receives a vote for the class of the terminal node it reaches. The final class membership of $X$ is obtained by a majority vote over the number of votes it receives for each class. The features are ranked by their respective reduction of node impurity as measured by the Gini index. Alternatives include the permutation-based measure,
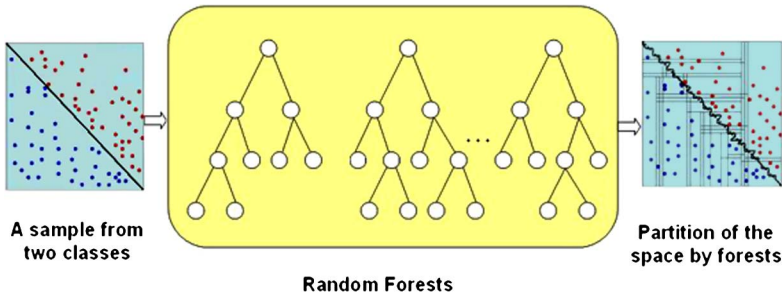
FIG. 4. *Random Forests classification. In this illustration the data points reside in a unit square* (*left panel*). *The two classes are indicated by red and blue dots. The true decision boundary is the diagonal line shown. RF* (*center panel*) *grows many trees. Each tree corresponds to a recursive partition of the data space. These partitions are represented in the right panel by a sequence of horizontal and vertical lines*; *the data space shown here is partitioned by many instances. The RF classifier eventually leads to a decision boundary* (*solid black curve*) *for this two-class classification problem.*

that is, permute variables one at a time and then rank according to the respective amount of decrease in accuracy (as estimated on oob observations).

2.4. *Salient pixels detection.* A valuable property of TACOMA is its ability to report salient pixels in an image that determine its score (see Figure 8). This property is based on a correspondence between the position of pixels in an image and entries in its GLCM, and made possible by the remarkable variable-ranking capability of RF. Here we use the importance measure (Gini index-based) provided by RF to rank the variables (i.e., entries of the GLCM) and then collect relevant image pixels associated with the important entries.

Since each entry of a GLCM is a counting statistic involving pairs of pixels, we can associate the $(a, b)$-entry of a GLCM with those pixels that make up this GLCM entry. The set of image pixels that are associated with the $(a, b)$-entry of a GLCM is formally represented as

$$\mathcal{G}_{a,b} = \{x, y : x \sim y, I(x) = a, I(y) = b\}.$$

In the above representation, $x$ and $y$ represent the position of image pixels and we treat an image $I$ as a map from the position of an image pixel to its gray value. Note that not all pairs of pixels with $x \sim y$ such that $I(x) = a, I(y) = b$ correspond to salient spots in a TMA image. However, if the $(a, b)$-feature is "important" (e.g., as determined by RF), then typically most pixels in the set $\mathcal{G}_{a,b}$ are relevant.

Whereas RF appears to be a black box—taking a large number of GLCM features and producing a score—salient pixels provide a quick peek into its internals. Effectively, RF works in roughly the same manner as a pathologist, that is, they both use salient pixels to score the TMA images; the seemingly mysterious image features are merely a form of representation for use by a computer algorithm.

---

**Algorithm 1** The training in TACOMA

---

1: **for** $i = 1$ **to** $L$ **do**
2:     compute the GLCM for the $i$th image patch and denote by $Z_i$;
3:     $M_i \leftarrow$ the index set of $Z_i$ that survive thresholding at level $\tau_i$;
4: **end for**
5: $M \leftarrow \bigcup_{j=1}^{L} M_j$;
6: **for** $k = 1$ **to** $n$ **do**
7:     compute the GLCM of image $I_k$ and keep only entries in index set $M$;
8:     denote the resulting matrix by $X_k$;
9: **end for**
10: Feed $\bigcup_{l=1}^{n} \{(X_l, Y_l)\}$ into the RF classifier and obtain a classification rule.

---

2.5. *An algorithmic description of TACOMA.* Denote the training sample by $(I_1, Y_1), \ldots, (I_n, Y_n)$ where $I_i$'s are images and $Y_i$'s are scores. Additionally, assume there is a set of $L$ "representative" image patches. The training of TACOMA is described as Algorithm 1.

In the above description, $\tau_i$ is chosen as the median of entries of matrix $Z_i$ for $i = 1, \ldots, L$. Then, for a new image, TACOMA will: (i) derive the GLCM matrix; (ii) select the entries with indices in $M$; (iii) apply the trained classifier on the selected entries and output the score. The training and scoring with TACOMA are illustrated in Figure 5.

**3. Co-training with RF.** The sample size is an important issue in the scoring of TMA images, mainly because of the high cost and human efforts involved in obtaining a large sample of high quality labels. For instance, it may take several
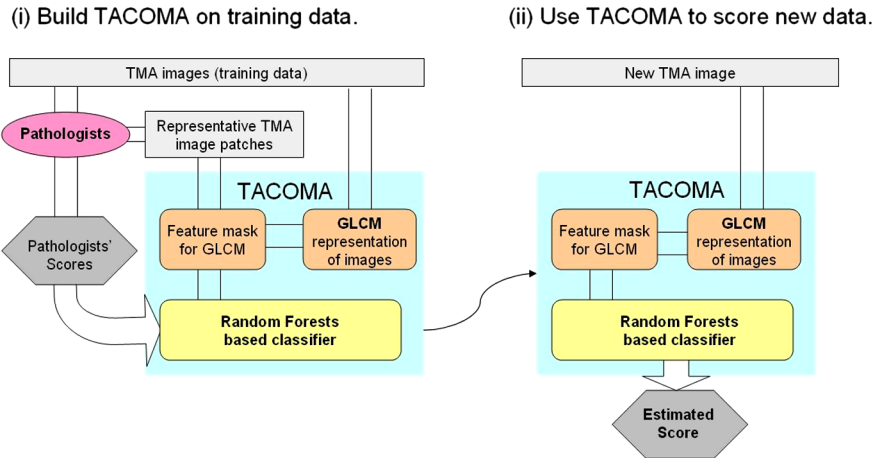


FIG. 5. *TACOMA illustrated. The left and right panels illustrate, respectively, model training and the use of the model on future data.*
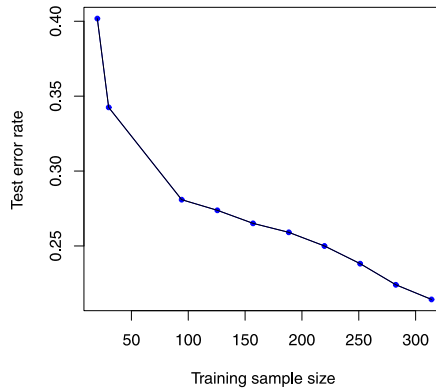
FIG. 6.   *Error rate of TACOMA as the training sample size varies. There are* 328 *TMA images in the test sample.*

hours for a well-trained pathologist to score 100 TMA images. Unfortunately, it is often the case that the classification performance drops sharply when the training sample size is reduced. For example, Figure 6 shows the error rate of TACOMA when the sample size varies. Our aim is to achieve reasonable accuracy for small sample size and co-training is adopted for this purpose.

Co-training was proposed in the landmark papers by Yarowsky (1995) and Blum and Mitchell (1998). It is an effective way in training a model with an extremely small labeled sample and has been successfully applied in many applications [Blum and Mitchell (1998), Nigam and Ghani (2000)]. The idea is to train two separate classifiers (called coupling classifiers) each on a different set of features using a small number of labeled examples. Then the two classifiers iteratively transfer those confidently classified examples, along with the assigned label, to the labeled set. This process is repeated until all unlabeled examples have been labeled. For an illustration of the idea of co-training, see Figure 7. Co-training is relevant here due to the natural redundancy that exists among features that are based on GLCMs corresponding to different spatial relationships.
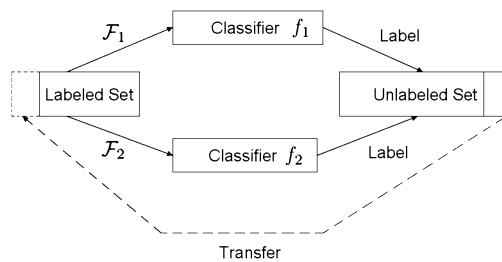


FIG. 7.   *An illustration of co-training.*

---

**Algorithm 2** The co-training algorithm

---

1: **while** the set $\mathcal{U}$ is not empty **do**
2:  **for** $k = 1, 2$ **do**
3:   Train RF classifier $f_k$ on labeled examples from $\mathcal{L}$ using feature set $\mathcal{F}_k$;
4:   Classify examples in the set $\mathcal{U}$ with $f_k$;
5:   Under $f_k$, calculate the margin for each observation in $\mathcal{U}$;
6:   pick $m_k$ observations, $x_1^{(k)}, \ldots, x_{m_k}^{(k)}$, with the largest margins;
7:  **end for**
8:  $\mathcal{L} \leftarrow \mathcal{L} \cup \{x_1^{(1)}, \ldots, x_{m_1}^{(1)}, x_1^{(2)}, \ldots, x_{m_2}^{(2)}\}$;
9:  $\mathcal{U} \leftarrow \mathcal{U} \setminus \{x_1^{(1)}, \ldots, x_{m_1}^{(1)}, x_1^{(2)}, \ldots, x_{m_2}^{(2)}\}$;
10: **end while**

---

A learning mode that is closely related to co-training is self-learning [Nigam and Ghani (2000)], where a single classifier is used in the "*label → transfer → label*" loop (cf. Figure 7). However, empirical studies have shown that co-training is often superior [Nigam and Ghani (2000)]; the intuition is that co-training allows the two coupling classifiers to progressively expand the "knowledge boundary" of each other which is absent in self-learning.

Previous works in co-training use almost exclusively Expectation Maximization or Naive Bayes based classifiers where the posterior probability serves as the "confidence" required by co-training. Here we use RF [Breiman (2001)] where the margin (to be defined shortly) provided by RF is used as a "natural" proxy for the "confidence." The margin is defined through the votes received by an observation. For an observation $x$ in the test set, let the number of votes it receives for the $i$th class be denoted by $N_i(x)$, $i = 1, \ldots, C$, where $C$ is the number of classes. The *margin* of $x$ is defined as

$$\max_{i \in \{1, \ldots, C\}} N_i(x) - \operatorname*{second}_{j \in \{1, \ldots, C\}} N_j(x),$$

where *second* in the above indicates the second-largest element in a list.

To give an algorithmic description of co-training, let the two subsets of features be denoted by $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Let the set of labeled and unlabeled examples be denoted by $\mathcal{L}$ and $\mathcal{U}$, respectively. The co-training process proceeds as Algorithm 2. The final error rate and class membership are determined by a fixed coupling classifier, say, $f_1$. We set $m_1 = m_2 = 2$ in our experiments according to Blum and Mitchell (1998).

3.1. *Feature split for co-training.* Co-training requires two subsets of features (or a feature split). However, co-training algorithms rarely provide a recipe for obtaining these feature splits. There are several possibilities one can explore.

The first is called a "natural" split, often resulting from an understanding of the problem structure. A rule of thumb as to what constitutes a natural split is that each

of the two feature subsets alone allows one to construct an acceptable classifier and that the two subsets somehow complement each other (e.g., conditional independence given the labels). Fortunately, TMA images represented in GLCM's naturally have such properties. For a given problem, often there exist several *spatial relationships* [e.g., $(\nearrow, 3)$ and $(\searrow, 1)$ for TMA images studied in this work], with each inducing a GLCM sufficient to construct a classifier while the "dependence" among the induced GLCM's is usually low. Thus, it is ideal to apply co-training on TMA images using such natural splits.

When there is no natural split readily available, one has to find two proper subsets of features. One way is via random splitting. Co-training via a random split of features was initially considered by Nigam and Ghani (2000) but has since been largely overlooked in the machine learning literature. Here we extend the idea of random splits to "thinning," which is more flexible and potentially may lead to a better co-training performance. Specifically, rather than randomly splitting the original feature set $\mathcal{F} = \{1, \ldots, p\}$ into two halves, we select two disjoint subsets of $\mathcal{F}$ with size not necessarily equal but nonvanishing compared to $p$. This way of feature splitting leads to feature subsets smaller than $\mathcal{F}$, hence the name "thinning." One concrete implementation of this is to divide $\mathcal{F}$ into a number of, say, $J$, equal-sized partitions (each partition is also called a thinned slice of $\mathcal{F}$). In the following discussion, unless otherwise stated, thinning always refers to this concrete implementation. It is clear that this includes random splits as a special case. Thinning allows one to construct a self-learning classifier (the features are taken from one of the $J$ partitions), co-training (randomly pick 2 out of $J$ partitions) and so on. For a given problem, one can explore various alternatives associated with thinning but here we shall focus on co-training.

The extension of random split to thinning may lead to improved co-training performance, as thinning may make features from different partitions less dependent and meanwhile well preserves the classification power in a high-dimensional setting when there is sufficient redundancy among features (see Section 3.2). The optimal number of partitions can be selected by heuristics such as the kernel independence test [Bach and Jordan (2003), Gretton et al. (2007)], which we leave for future work.

3.2. *Some theoretical insights on thinning.* According to Blum and Mitchell (1998), one essential ingredient of co-training is the "high" confidence of the two coupling classifiers in labeling the unlabeled examples. This is closely related to the strength of the two coupling classifiers which is in turn determined by the feature subsets involved. In this section, we study how much a thinned slice of the feature set $\mathcal{F}$ preserves its classification power. Our result provides insight into the nature of thinning and is interesting at its own right due to its close connection to several lines of interesting work [Ho (1998), Dasgupta and Gupta (2002)] in machine learning (see Section 3.2.2). We present our theoretical analysis in Section 3.2.1 and list related work in Section 3.2.2. In Supplement A [Yan et al.

(2012)], we provide additional simulation results related to our theoretical analysis.

3.2.1. *Thinning "preserves" the ratio of separation.* Our theoretical model is the Gaussian mixture specified as

$$\Pi \mathcal{N}(\boldsymbol{\mu}_1, \Sigma) + (1 - \Pi)\mathcal{N}(\boldsymbol{\mu}_2, \Sigma), \tag{2}$$

where $\Pi \in \{0, 1\}$ indicates the label of an observation such that $\mathbb{P}(\Pi = 1) = \pi$, and $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ stands for Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\Sigma$. For simplicity, we consider $\pi = \frac{1}{2}$ and the 0–1 loss. We will define the ratio of separation as a measure of the fraction of "information" carried by the subset of features due to thinning with respect to that of the original feature set and show that this quantity is "preserved" upon thinning. For simplicity, we take $J = 2$ (i.e., random splits of $\mathcal{F}$) and similar discussion applies to $J > 2$.

Let the feature set $\mathcal{F}$ be decomposed as

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \qquad \text{such that } \mathcal{F}_1 \cap \mathcal{F}_2 = \varnothing \text{ and } |\mathcal{F}_1| = \frac{p}{2} \triangleq m. \tag{3}$$

We will show that each of the two subsets of features, $\mathcal{F}_1$ and $\mathcal{F}_2$, carries a substantial fraction of the "information" contained in the original data when $p$ is large, assuming the data is generated from Gaussian mixture (2).

A quantity that is crucial in our inquiry is

$$S_{\mathcal{F}} = \mathbf{u}_{\mathcal{F}}^T \Sigma_{\mathcal{F}}^{-1} \mathbf{u}_{\mathcal{F}}, \tag{4}$$

where $\mathbf{u}_{\mathcal{F}} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)_{\mathcal{F}} \triangleq (U_1, U_2, \ldots, U_p)_{\mathcal{F}}$ and here $\mathcal{F}$, as a subscript, indicates that the associated quantity corresponds to the feature set $\mathcal{F}$. We call $S_{\mathcal{F}}$ the separation of the Gaussian mixture induced by the feature set $\mathcal{F}$. The separation is closely related to the Bayes error rate for classification through a well-known result in multivariate statistics.

LEMMA 1 [Anderson (1958)]. *For Gaussian mixture* (2) *and* 0–1 *loss, the Bayes error rate is given by* $\Phi(-\frac{1}{2}(\mathbf{u}_{\mathcal{F}}^T \Sigma_{\mathcal{F}}^{-1} \mathbf{u}_{\mathcal{F}})^{1/2})$ *where* $\Phi(\cdot)$ *is defined as* $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$.

Let the covariance matrix $\Sigma$ be written as

$$\Sigma = \begin{bmatrix} A & B^T \\ B & D \end{bmatrix},$$

where we assume block $A$ corresponds to features in $\mathcal{F}_1$ after a permutation of rows and columns. Accordingly, write $\mathbf{u}$ as $\mathbf{u}_{\mathcal{F}} = (\mathbf{u}_{\mathcal{F}_1}, \mathbf{u}_{\mathcal{F}_2})$ and define $S_{\mathcal{F}_1}$ (called the separation induced by $\mathcal{F}_1$) similarly as (4). Now we can define the *ratio of separation* for the feature subset $\mathcal{F}_1$ as

$$\gamma = \frac{S_{\mathcal{F}_1}}{S_{\mathcal{F}}}. \tag{5}$$

To see why definition (5) is useful, we give here a numerical example. Assume there is a Gaussian mixture defined by (2) such that $\Sigma_{100 \times 100}$ is a tri-diagonal matrix with diagonals being all 1 and off-diagonals being 0.6, $\mathbf{u}_{\mathcal{F}} = (1, \dots, 1)^T$. Suppose one picks the first 50 variables and form a new Gaussian mixture with covariance matrix $A$ and mixture center distance $\mathbf{u}_{\mathcal{F}_1}$. We wish to see how much is affected in terms of the Bayes error rate. We have

$$S_{\mathcal{F}} = 45.87, \qquad \Phi\left(-\tfrac{1}{2}(\mathbf{u}_{\mathcal{F}}^T \Sigma_{\mathcal{F}}^{-1} \mathbf{u}_{\mathcal{F}})^{1/2}\right) = 3.54 \times 10^{-4},$$

$$S_{\mathcal{F}_1} = 23.32, \qquad \Phi\left(-\tfrac{1}{2}(\mathbf{u}_{\mathcal{F}_1}{}^T A^{-1} \mathbf{u}_{\mathcal{F}_1})^{1/2}\right) = 7.87 \times 10^{-3}$$

and $\gamma = 0.5084$. Here the difference between feature set $\mathcal{F}_1$ and $\mathcal{F}$ is very small in terms of their classification power. In general, if the dimension is sufficiently high and $\gamma$ is nonvanishing, then using a subset of features will not incur much loss in classification power. In Theorem 2, we will show that, under certain conditions, $\gamma$ does not vanish (i.e., $\gamma > c$ for some positive constant $c$) so a feature subset is as good as the whole feature set in terms of classification power.

Our main assumption (i.e., in Theorem 2) is actually a technical one related to the "local" dependency among components of $\mathbf{u}$ after some variable transformation. The exact context will become clear later in the proof of Theorem 2. For now, let $\Sigma$ have a Cholesky decomposition $\Sigma = H H^T$ for some lower triangular matrix $H$. A variable transformation in the form of $\mathbf{y} = H^{-1} \mathbf{u}_{\mathcal{F}}$ will be introduced. The idea is that we desire $\Sigma = H H^T$ to possess a structure such that the components of $\mathbf{y} = H^{-1} \mathbf{u}_{\mathcal{F}}$ are "locally" dependent so that some form of law of large numbers may be applied. To avoid technical details in the main text, we shall discuss the assumption in the supplement [Yan et al. (2012)].

Our main result is the following theorem.

THEOREM 2. *Assume the data are generated from Gaussian mixture* (2). *Further assume the smallest eigenvalue of* $\Sigma^{-1}$, *denoted by* $\lambda_{\min}(\Sigma^{-1})$, *is bounded away from* 0 *under permutations of rows and columns of* $\Sigma$. *Then, under assumptions of Lemma* 3 *(c.f. Supplement A* [Yan et al. (2012)]), *the separation induced by the feature set* $\mathcal{F}_1$ *satisfies*

$$\frac{S_{\mathcal{F}_1}}{S_{\mathcal{F}}} \geq \left(\frac{1}{2}\right)^{-}$$

*in probability as* $p \to \infty$ *where* $(a)^-$ *indicates any constant smaller than* $a$. *When the number of partitions* $J > 2$, *the right-hand side is replaced by* $1/J$.

PROOF. See supplement [Yan et al. (2012)] for proof. □

3.2.2. *Related work.* There are mainly two lines of work closely related to ours. One is the Johnson–Lindenstrauss lemma and related [Johnson and Lindenstrauss (1984), Dasgupta and Gupta (2002)]. The Johnson–Lindenstrauss (or J–L) lemma states that, for Gaussian mixtures in high-dimensional space, upon a random projection to a low-dimensional subspace, the separation between the mixture centers in the projected space is "comparable" to that in the original space with high probability. The difference is that the random projection in J–L is carried out via a nontrivial linear transformation and the separation is defined in terms of the Euclidean distance whereas, in our work, random projection is performed coordinate-wise in the original space and we define the separation with the Mahalanobis distance.

The other related work is the random subspace method [Ho (1998)], an early variant of the RF classifier ensemble algorithm, that is, comparable to bagging and Adaboost in terms of empirical performance. The random subspace method grows a tree by randomly selecting half of the features and then constructs a tree-based classifier. However, beyond simulations there has been no formal justification for the random selection of half of the features. Our result provides support on this aspect. In a high dimensional data setting where the features are "redundant," our result shows that a randomly-selected half of the features can lead to a tree comparable, in terms of classification power, to a classifier that uses all the features; meanwhile the random nature of the set of features used in each tree makes the correlation between trees small, so good performance can be expected.

Our theoretical result, when used in co-training, can be viewed as a manifestation of the "blessings of the dimensionality" [Donoho (2000)]. For high-dimensional data analysis, the conventional wisdom is to do dimension reduction or projection pursuit. As a result, the "redundancy" among the features is typically not used and, in many cases, even becomes the nuisance one strives to get rid of. This is clearly a waste. When the "redundancy" among features is complementary, such redundancy actually allows one to construct two coupling learners from which co-training can be carried out. It should be emphasized that the splitting of the feature set works because of redundancy. We believe the exploration of this type of redundancy will have important impact in high-dimensional data analysis.

**4. Applications on TMA images.** To assess the performance of TACOMA, we evaluate a collection of TMA images from the Stanford Tissue Microarray Database, or STMAD [see Marinelli et al. (2007) and http://tma.stanford.edu/]. TMAs corresponding to the potential expression of the estrogen receptor (ER) protein in breast cancer tissue are used since ER is a histologically well-studied marker that is expressed in the cell nucleus. An example of TMA images can be seen in Figure 2. There are 641 TMA images in this set and each image has been assigned a score from $\{0, 1, 2, 3\}$. The scoring criteria are as follows: "0" representing a definite negative (no staining of cancer cells), "3" a definitive positive (a majority of cancer cells show dark nucleus staining) and "2" for positive

(a minority of cancer cells show nucleus staining or a majority show weak nucleus staining). The score "1" indicates ambiguous weak staining in a minority of cancer cells. The class distribution of the scores is (65.90%, 2.90%, 7.00%, 24.20%). Such an unbalanced class distribution makes the scoring task even more challenging. In our experiments, we assess the accuracy by the proportion of images in the test sample that receives the same score by a scoring algorithm as that given by the reference set (e.g., STMAD). We split the images into a training and a test set of sizes 313 and 328, respectively. In the following, we will first describe our choice of various design options and parameters, and then report performance of TACOMA on a large training set (i.e., a set of 313 TMA images) and small training set (i.e., a set of 30 TMA images) with co-training in Section 4.1 and Section 4.2, respectively.

We use three spatial relationships, $(\nearrow, 3)$, $(\searrow, 1)$ and $(\nearrow, 1)$, in our experiments. In particular, $(\nearrow, 3)$ is used in our experiment on TACOMA for a large training set, while $(\searrow, 1)$ is used along with $(\nearrow, 3)$ in our co-training experiment and $(\nearrow, 1)$ in an additional experiment (see Table 3). Often $(\nearrow, 1)$ is the default choice in applications; we use $(\nearrow, 3)$ here to reflect the granularity of the visual pattern seen in the images. Indeed, as the staining patterns in TMA images for ER markers occur only in the nucleus, $(\nearrow, 3)$ leads to a slightly better scoring performance than $(\searrow, 1)$ according to our experiment on TACOMA; moreover, no significant difference is observed for TACOMA when simply concatenating features derived from $(\nearrow, 3)$ and $(\searrow, 1)$. $(\searrow, 1)$ is used in co-training in hoping that it is less correlated to features derived from $(\nearrow, 3)$ than others, as these two are on the orthogonal directions.

For a good balance of computational efficiency, discriminative power, as well as ease of implementation, we take $N_g = 51$ (our experiments are not particularly sensitive to the choice of $N_g$ in the range of 40 to 60) and apply uniform quantization over the 256 gray levels in our application. One can, of course, further explore this, but we would not expect a substantial gain in performance due to the limitation of a computer algorithm (or even human eyes) in distinguishing subtle differences in gray levels given a moderate sample size.

We use the R package "randomForest"[4] in this work. There are two important parameters, the number of trees in the ensemble and the number of features to explore at each node split. These are searched through $\{50, 100, 200, 500\}$ and $\{0.5\sqrt{p}, \sqrt{p}, 2\sqrt{p}\}$ ($\sqrt{p}$ is the default value suggested by the R package for $p$ the number of features fed to RF), respectively, in this work and the best test set error rates are reported. More information on RF can be found in Breiman (2001).

---

[4]Originally written in Fortran by Leo Breiman and Adele Cutler, and later ported to R by Andy Liaw and Matthew Wiener.

4.1. *Performance on large training set.* The full set of 313 TMA images in the training set are used in this case. We run TACOMA on the training set (scores given by STMAD) and apply the trained classifier to the test set images to obtain TACOMA scores. Then, we blind STMAD scores in the test set of 328 images (100 of which are duplicated so totally there are 428 images) and have them reevaluated by two experienced pathologists from two different institutions. The 100 duplicates allow us to evaluate the self-consistency of the pathologists.

Although the scores from STMAD do not necessarily represent ground truth, they serve as a fixed standard with respect to which the topics of accuracy and reproducibility can be examined. On the test set of 328 TMA images, TACOMA achieves a classification accuracy of 78.57% (accuracy defined as the proportion of images receiving the same score as STMAD). We argue this is close to the optimal. The Bayes rate is estimated for this particular data example (represented as GLCMs) with a simulation using a 1-nearest neighbor (1*NN*) classifier. The Bayes rate refers to the theoretically best classification rate given the data distribution. With the same training and test sets as RF classification, the accuracy achieved by 1*NN* is around 60%. According to a celebrated theorem of Cover and Hart (1967), the error rate by 1*NN* is at most twice that of the Bayes rule. This result implies an estimate of the Bayes rate at around 80% subject to small sample variation (the estimated Bayes rates on the original image or its quantized version are all bounded above by this number according to our simulation). Thus, TACOMA is close to optimal.

In the above experiments, we use four image patches. To see if TACOMA is sensitive to the choice of image patches, we conduct experiments over a range of different patch sets and achieve an average accuracy at $78.66 \pm 0.52\%$. Such an accuracy indicates that TACOMA is robust to the choices of image patches.

It is worth reemphasizing that all reports of test errors for TMA images are not based on absolute truth, as all scores given to these images are subjectively provided by a variable human scoring process.

*Salient spots detection.* The ability of TACOMA to detect salient pixels is demonstrated in Figure 8 where image pixels are highlighted in white if they are associated with a significant scoring feature. These highlighted pixels are verified by the pathologists to be indicative. With relatively few exceptions, these locations correspond to areas of stained nuclei in cancer cells. We emphasize that these highlighted pixels indicate features most important for classification as opposed to identifying every property indicative of ER status. The highlighted pixels facilitate interpretation and the comparison of images by pathologists.
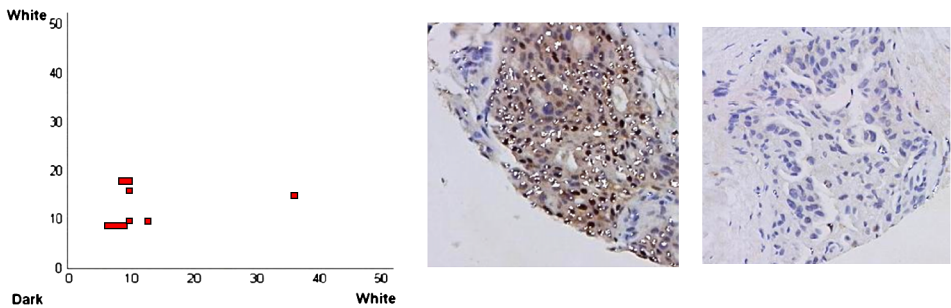
FIG. 8. *The salient pixels* (*highlighted in white*). *The left panel displays top features* (*indices of GLCM entries*) *from the classifier where the x-axis and y-axis indicate the row and column of the GLCM entries. The middle and right panels display images having scores* 3 *and* 0, *respectively*; *the pixels highlighted in white are those that correspond to the GLCM entries shown in the left panel. Note the highlighted pixels in the right panel are notably absent. For visualization, only part of the images are shown* (*see Supplement B* [Yan et al. (2012)]) *for larger images*).

*The experiments with pathologists.* The superior classification performance of TACOMA is also demonstrated by scores provided by the two pathologists. These two copies of scores, along with STMAD, provide three independent pathologist-based scores. Among these, 142 images receive a unanimous score. Consequently, these may be viewed as a reference set of "true" scores against which the accuracy of TACOMA might be evaluated (accuracy being defined as the proportion of images receiving the same score as the reference set). Here, TACOMA achieves an accuracy of 90.14%; see Figure 9.
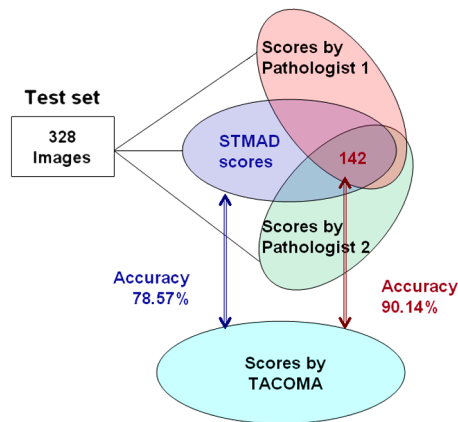


FIG. 9. *Classification performance of TACOMA. On the STMAD test set TACOMA achieves an accuracy of* 78.57%. *On the* 142 *images assigned a unanimous score by two pathologists and STMAD, TACOMA agrees on about* 90%.

Scores provided by the two pathologists are also used to assess their self-consistency. Here self-consistency is defined as the proportion of repeated images receiving an identical score by the same pathologist. Consensus among different pathologists is an issue of concern [Penna et al. (1997), Walker (2006)]. In order to obtain information about the self-consistency of pathologist-based scores, 100 images are selected from the set of 328 images. These 100 images are rotated and/or inverted, and then mixed at random with the 328 images to avoid recognition (so each pathologist actually scored a total of 428 TMA images). The self-consistency rates of the two pathologists are found to be in the range 75–84%. Of course, one desirable feature of any automated algorithm such as TACOMA is its complete (i.e., 100%) self-consistency.

*Performance comparison of RF to SVM and boosting.* The classification algorithm chosen for TACOMA is RF. Some popular alternatives include support vector machines (SVM) [Cortes and Vapnik (1995)], Boosting [Freund and Schapire (1996)] and Bayesian network [Pearl (1985)], etc. Using the same training and test set as that for RF, we conduct experiments with SVM as well as Boosting of a naive Bayes classifier. The input features for both SVM and Boosting are the entries of the GLCM. We use the Libsvm [Chang and Lin (2001)] software for SVM. The naive Bayes classifier is adopted from Yan, Bickel and Gong (2006). The idea is to find the class that maximizes the posterior probability for a new observation, denoted by $(I(x) = a_x, x \in \{1, 2, \ldots, N_g\} \otimes \{1, 2, \ldots, N_g\})$ for a fixed $N_g$. That is, we seek to solve

$$\arg \max_{k \in \{0,1,2,3\}} \text{Prob}\{k | I(1, 1) = a_{1,1}, \ldots, I(N_g, N_g) = a_{N_g, N_g}\}$$

under the assumption that $(I(1, 1), \ldots, I(N_g, N_g))$ follows a multinomial distribution. More details can be found in Yan, Bickel and Gong (2006). The results are shown in Table 1. We can see that RF outperforms SVM and Boosting by a large margin. This is consistent with observations made by Holmes, Kapelner and Lee (2009).

4.2. *Experiments on small training sets.* We conduct experiments on co-training with natural splits and thinning. For natural splits, we use GLCM's corresponding to two spatial relationships, $(\nearrow, 3)$ and $(\searrow, 1)$, as features. For thinning, we combine features corresponding to $(\nearrow, 3)$ and $(\searrow, 1)$ and then split this combined feature set.

The number of labeled examples is fixed at 30 (compared to 313 in experiments with a large training set). This choice is designed to make it easy to get a nonempty class 1 (which carries only about 2.90% of the cases). We suspect this number can be further reduced without suffering much in learning accuracy. The test set is the same as that in Section 4.1. The result is shown in Table 2. One interesting observation is that co-training by thinning achieves an accuracy very close to that by

TABLE 2

*Performance of RF and co-training by thinning on TMA images. The unlabeled set is taken as the test set in Section 4.1 and the labeled set is randomly sampled from the corresponding training set. The subscript for "thinning" indicates the number of partitions. The results are averaged over 100 runs and over the two coupling classifiers for co-training*

| Scheme | Error rate |
|---|---|
| RF on $(\nearrow, 3) \cup (\searrow, 1)$ | 34.36% |
| Thinning$_2$ on $(\nearrow, 3) \cup (\searrow, 1)$ | 34.21% |
| Thinning$_3$ on $(\nearrow, 3) \cup (\searrow, 1)$ | 34.18% |
| Co-training by natural split on $(\nearrow, 3)$ and $(\searrow, 1)$ | 27.49% |
| Co-training by thinning$_2$ on $(\nearrow, 3) \cup (\searrow, 1)$ | 27.89% |
| Co-training by thinning$_3$ on $(\nearrow, 3) \cup (\searrow, 1)$ | 27.62% |

natural splits. Additionally, Table 2 lists error rates given by RF on features corresponding to $(\nearrow, 3) \cup (\searrow, 1)$ and its thinned subsets. Here thinning of the feature set does not cause much loss in RF performance, consistent with our discussion in Section 3.2.

4.3. *Experiments of TACOMA on additional data sets.* This study focuses on the ER marker for which the staining is nuclear. However, the TACOMA algorithm can be applied with equal ease to markers that exhibit cell surface, cytoplasm or other staining patterns. Additional experiments are conducted on the Stanford TMA images corresponding to three additional protein markers: CD117, CD34 and NMB. These three sets of TMA images are selected for their large sample size and relatively few missing scores (excluded from experiment). The results are shown in Table 3. In contrast, the automated scoring of cytoplasmic markers is often viewed as more difficult and refined commercial algorithms for these were reportedly not available in a recent evaluation [Camp, Neumeister and Rimm (2008)] of commercial scoring methods.

TABLE 3

*Accuracy of TACOMA on TMA images corresponding to protein markers CD117, CD34, NMB and ER. Except for ER (which has a fixed training and test set), we use $(\nearrow, 1)$ and 80% of the instances for training and the rest for test; this is repeated for 100 runs and results averaged*

| Marker | Staining | #Instances | Accuracy |
|---|---|---|---|
| ER | Nucleus | 641 | 78.57% |
| CD117 | Cell surface | 1063 | 81.08% |
| NMB | Cytoplasmic | 1036 | 84.17% |
| CD34 | Cytoplasmic and cell surface | 908 | 76.44% |

**5. Discussion.** We have presented a new algorithm that automatically scores TMA images in an objective, efficient and reproducible manner. Our contributions include the following: (1) the use of co-occurrence counting statistics to capture the spatial regularity inherent in a heterogeneous and irregular set of TMA images; (2) the ability to report salient pixels in an image that determine its score; (3) the incorporation of pathologists' input via informative training patches which makes our algorithm adaptable to various markers and cell types; (4) a very small training sample is achievable with co-training and we have provided some theoretical insights into co-training via thinning of the feature set. Our experiments show that TACOMA can achieve performance comparable to well-trained pathologists. It uses the similar set of pixels for scoring as that would be used by a pathologist and is not adversely sensitive to the choice of image patches. The theory we have developed on the thinning scheme in co-training gives insights on why thinning may rival the performance of a natural split in co-training; a thinned slice may be as good as the whole feature set in terms of classification power, hence, thinning can lead to two strong coupling classifiers that will be used in co-training and this is what a natural split may achieve.

The utility of TACOMA lies in large population-based studies that seek to evaluate potential markers using IHC in large cohorts. Such a study may be compromised by a scoring process, that is, protracted, prohibitively expensive or poorly reproducible. Indeed, a manual scoring for such a study could require hundreds of hours of pathologists' time without achieving a reproducibly consistent set of scores. Experiments with several IHC markers demonstrate that our approach has the potential to be as accurate as manual scoring while providing a fast, objective, inexpensive and highly reproducible alternative. Even more generally, TACOMA may be adopted to other types of textured images such as those appearing in remote sensing applications. These properties provide obvious advantages for any subsequent statistical analysis in determining the validity or clinical utility of a potential marker. Regarding reproducibility, we note that the scores provided by two pathologists in our informal pilot study revealed an intra-observer agreement of around 80% and an accuracy only in the range of 70%, as defined by the STMAD reference set (excluding all images deemed unscorable by the pathologists). This low inter-observer agreement may be attributed to a variety of factors, including a lack of a subjective criteria used for scoring or the lack of training against an established standard. This performance could surely be improved upon, but it highlights the inherent subjectivity and variability of human-based scoring.

In summary, TACOMA provides a transparent scoring process that can be evaluated with clarity and confidence. It is also flexible with respect to marker patterns of cellular localization: although the ER marker is characterized by staining of the cell nucleus, TACOMA applies with comparable ease and success to cytoplasmic or other marker staining patterns (see Table 3 in Section 4.3).

A software implementation of TACOMA is available upon request and the associated R package will be made available to the R project.

**Acknowledgments.** The authors would like to thank the Associate Editor and the anonymous reviewers for their constructive comments and suggestions.

## SUPPLEMENTARY MATERIAL

**Supplement A: Assumption $\mathcal{A}_1$, proof of Theorem 2 and simulations on thinning** (DOI: 10.1214/12-AOAS543SUPPA; .pdf). We provide a detailed description of Assumption $\mathcal{A}_1$, a sketch of the proof of Theorem 2 and simulations on the ratio of separation upon thinning under different settings.

**Supplement B: TMA images with salient pixels marked** (DOI: 10.1214/12-AOAS543SUPPB; .pdf). This supplement contains a close view of some TMA images where the salient pixels are highlighted.

## REFERENCES

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York. MR0091588

BACH, F. R. and JORDAN, M. I. (2003). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051

BENTZEN, S. M., BUFFA, F. M. and WILSON, G. D. (2008). Multiple biomarker tissue microarrays: Bioinformatics and practical approaches. *Cancer Metastasis Rev.* **27** 481–494.

BERGER, A. J., DAVIS, D. W., TELLEZ, C., PRIETO, V. G., GERSHENWALD, J. E., JOHNSON, M. M., RIMM, D. L. and BAR-ELI, M. (2005). Automated quantitative analysis of activator protein-2alpha subcellular expression in melanoma tissue microarrays correlates with survival prediction. *Cancer Res.* **65** 11185–11192.

BLUM, A. and MITCHELL, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (*Madison, WI*, 1998) 92–100. ACM, New York. MR1811574

BREIMAN, L. (2001). Random Forests. *Mach. Learn.* **45** 5–32.

CAMP, R. L., NEUMEISTER, V. and RIMM, D. L. (2008). A decade of tissue microarrays: Progress in the discovery and validation of cancer biomarkers. *J. Clin. Oncol.* **26** 5630–5637.

CARUANA, R., KARAMPATZIAKIS, N. and YESSENALINA, A. (2008). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning* (*ICML*), *Helsinki, Finland* 96–103. ACM, New York.

CHANG, C. C. and LIN, C. J. (2001). LIBSVM: A library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

CONNERS, R. and HARLOW, C. (1980). A theoretical comparison of texture algorithms. *IEEE Transactions on Pattern Analyses and Machine Intelligence* **2** 204–222.

CORTES, C. and VAPNIK, V. (1995). Support vector networks. *Machine Learning* **20** 273–297.

COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13** 21–27.

CREGGER, M., BERGER, A. and RIMM, D. (2006). Immunohistochemistry and quantitative analysis of protein expression. *Archives of Pathology & Laboratory Medicine* **130** 1026–1030.

DASGUPTA, S. and GUPTA, A. (2002). An elementary proof of the Johnson–Lindenstrauss lemma. *Random Structures Algorithms* **22** 60–65.

DIVITO, K. and CAMP, R. (2005). Tissue microarrays—automated analysis and future directions. *Breast Cancer Online* **8**.

DONOHO, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century*. Available at http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/AMS2000.html.

FREUND, Y. and SCHAPIRE, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning* (*ICML*), *Bari*, *Italy* 148–156. Morgan Kaufman, San Francisco, CA.

GILTNANE, J. M. and RIMM, D. L. (2004). Technology insight: Identification of biomarkers with tissue microarray technology. *Nat*. *Clin*. *Pract*. *Oncol*. **1** 104–111.

GRETTON, A., FUKUMIZU, K., TEO, C. H., SONG, L., SCHÖLKOPF, B. and SMOLA, A. J. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* (*NIPS*), *Vancouver*, *BC*, *Canada* 585–592. MIT Press, Cambridge, MA.

HALL, B. H., IANOSI-IRIMIE, M., JAVIDIAN, P., CHEN, W., GANESAN, S. and FORAN, D. J. (2008). Computer-assisted assessment of the human epidermal growth factor receptor 2 immunohistochemical assay in imaged histologic sections using a membrane isolation algorithm and quantitative analysis of positive controls. *BMC Med*. *Imaging* **8** 11.

HARALICK, R. (1979). Statistical and structural approaches to texture. *Proceedings of IEEE* **67** 786–803.

HASSAN, S., FERRARIO, C., MAMO, A. and BASIK, M. (2008). Tissue microarrays: Emerging standard for biomarker validation. *Curr*. *Opin*. *Biotechnol*. **19** 19–25.

HO, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** 832–844.

HOLMES, S., KAPELNER, A. and LEE, P. (2009). An interactive JAVA statistical image segmentation system: GemIdent. *Journal of Statistical Software* **30** 1–20.

JOHNSON, W. B. and LINDENSTRAUSS, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability* (*New Haven*, *Conn*., 1982). *Contemp*. *Math*. **26** 189–206. Amer. Math. Soc., Providence, RI. MR0737400

JOSHI, A. S., SHARANGPANI, G. M., PORTER, K., KEYHANI, S., MORRISON, C., BASU, A. S., GHOLAP, G. A., GHOLAP, A. S. and BARSKY, S. H. (2007). Semi-automated imaging system to quantitate Her-2/neu membrane receptor immunoreactivity in human breast cancer. *Cytometry Part A* **71** 273–285.

KIRKEGAARD, T., EDWARDS, J., TOVEY, S., MCGLYNN, L. M., KRISHNA, S. N., MUKHERJEE, R., TAM, L., MUNRO, A. F., DUNNE, B. and BARTLETT, J. M. S. (2006). Observer variation in immunohistochemical analysis of protein expression, time for a change? *Histopathology* **48** 787–794.

KONONEN, J., BUBENDORF, L., KALLIONIMENI, A., BÄRLUND, M., SCHRAML, P., LEIGHTON, S., TORHORST, J., MIHATSCH, M., SAUTER, G. and KALLIONIMENI, O. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine* **4** 844–847.

MARINELLI, R., MONTGOMERY, K., LIU, C., SHAH, N., PRAPONG, W., NITZBERG, M., ZACHARIAH, Z., SHERLOCK, G., NATKUNAM, Y., WEST, R. et al. (2007). The Stanford tissue microarray database. *Nucleic Acids Res*. **36** D871–D877.

MASMOUDI, H., HEWITT, S. M., PETRICK, N., MYERS, K. J. and GAVRIELIDES, M. A. (2009). Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer. *IEEE Trans. Med. Imaging* **28** 916–925.

MULRANE, L., REXHEPAJ, E., PENNEY, S., CALLANAN, J. J. and GALLAGHER, W. M. (2008). Automated image analysis in histopathology: A valuable tool in medical diagnostics. *Expert Rev. Mol. Diagn.* **8** 707–725.

NIGAM, K. and GHANI, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management, McLean, VA* 86–93. ACM, New York.

PEARL, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the Seventh Conference of the Cognitive Science Society, Irvine, CA* 329–334. Cognitive Science Soc., Austin, TX.

PENNA, A., GRILLI, R., FILARDO, G., MAININI, F., ZOLA, P., MANTOVANI, L. and LIBERATI, A. (1997). Do different physicians' panels reach similar conclusions? A case study on practice guidelines for limited surgery in breast cancer. *European Journal of Public Health* **7** 436–440.

ROJO, M. G., BUENO, G. and SLODKOWSKA, J. (2009). Review of imaging solutions for integrated quantitative immunohistochemistry in the pathology daily practice. *Folia Histochem. Cytobiol.* **47** 349–354.

SKALAND, I., ØVESTAD, I., JANSSEN, E. A. M., KLOS, J., KJELLEVOLD, K. H., HELLIESEN, T. and BAAK, J. (2008). Digital image analysis improves the quality of subjective HER-2 expression scoring in breast cancer. *Applied Immunohistochemistry & Molecular Morphology* **16** 185–190.

TAWFIK, O. W., KIMLER, B. F., DAVIS, M., DONAHUE, J. K., PERSONS, D. L., FAN, F., HAGEMEISTER, S., THOMAS, P., CONNOR, C., JEWELL, W. et al. (2005). Comparison of immunohistochemistry by automated cellular imaging system (ACIS) versus fluorescence in-situ hybridization in the evaluation of HER-2/neu expression in primary breast carcinoma. *Histopathology* **48** 258–267.

THOMSON, T. A., HAYES, M. M., SPINELLI, J. J., HILLAND, E., SAWRENKO, C., PHILLIPS, D., DUPUIS, B. and PARKER, R. L. (2001). HER-2/neu in breast cancer: Interobserver variability and performance of immunohistochemistry with 4 antibodies compared with fluorescent in situ hybridization. *Mod. Pathol.* **14** 1079–1086.

VODUC, D., KENNEY, C. and NIELSEN, T. O. (2008). Tissue microarrays in clinical oncology. *Semin. Radiat. Oncol.* **18** 89–97.

VROLIJK, H., SLOOS, W., MESKER, W., FRANKEN, P., FODDE, R., MORREAU, H. and TANKE, H. (2003). Automated acquisition of stained tissue microarrays for high-throughput evaluation of molecular targets. *J. Mol. Diagn.* **5** 160–167.

WALKER, R. A. (2006). Quantification of immunohistochemistry—issues concerning methods, utility and semiquantitative assessment I. *Histopathology* **49** 406–410.

WAN, W. H., FORTUNA, M. B. and FURMANSKI, P. (1987). A rapid and efficient method for testing immunohistochemical reactivity of monoclonal antibodies against multiple tissue samples simultaneously. *J. Immunol. Methods* **103** 121–129.

YAN, D., BICKEL, P. J. and GONG, P. (2006). A discrete log density expansion based approach to Ikonos image classification. In *American Society for Photogrammetry and Remote Sensing Fall Speciality Conference, San Antonio, TX*. Am. Soc. Photogrammetry and Remote Sensing, Bethesda, MD.

YAN, D., GONG, P., CHEN, A. and ZHONG, L. (2011). Classification under data contamination with application to image mis-registration in remote sensing. Available at arXiv:1101.3594.

YAN, D., WANG, P., LINDEN, M., KNUDSEN, B. and RANDOLPH, T. (2012). Supplement to "Statistical methods for tissue array images—algorithmic scoring and co-training." DOI:10.1214/12-AOAS543SUPPA, DOI:10.1214/12-AOAS543SUPPB.

YAROWSKY, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, *Cambridge*, *MA* 189–196. Morgan Kaufmann, San Francisco, CA.

D. YAN
P. WANG
T. RANDOLPH
BIOSTATISTICS AND BIOMATHEMATICS PROGRAM
FRED HUTCHINSON CANCER RESEARCH CENTER
SEATTLE, WASHINGTON 98109
USA
E-MAIL: dyan@fhcrc.org
        pwang@fhcrc.org
        trandolp@fhcrc.org

M. LINDEN
DEPARTMENT OF LABORATORY MEDICINE
  AND PATHOLOGY
UNIVERSITY OF MINNESOTA MEDICAL SCHOOL
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: linde013@umn.edu

B. KNUDSEN
DEPARTMENT OF PATHOLOGY AND
  LABORATORY MEDICINE
DEPARTMENT OF BIOMEDICAL SCIENCES
CEDARS-SINAI MEDICAL CENTER
LOS ANGELES, CALIFORNIA 90048
USA
E-MAIL: Beatrice.Knudsen@cshs.org