Honorable Mention Poster Award

A Boosted Distance Metric: Application to Content Based Image Retrieval and Classification of Digitized Histopathology

Jay Naik[1], Scott Doyle[1], Ajay Basavanally[1], Shridar Ganesan[2], Michael D. Feldman[3], John E. Tomaszewski[3], Anant Madabhushi[1]

[1]Rutgers University, 599 Taylor Road, Piscataway, NJ
[2]Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ
[3]University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA

ABSTRACT

Distance metrics are often used as a way to compare the similarity of two objects, each represented by a set of features in high-dimensional space. The Euclidean metric is a popular distance metric, employed for a variety of applications. Non-Euclidean distance metrics have also been proposed, and the choice of distance metric for any specific application or domain is a non-trivial task. Furthermore, most distance metrics treat each dimension or object feature as having the same relative importance in determining object similarity. In many applications, such as in Content-Based Image Retrieval (CBIR), where images are quantified and then compared according to their image content, it may be beneficial to utilize a similarity metric where features are weighted according to their ability to distinguish between object classes. In the CBIR paradigm, every image is represented as a vector of quantitative feature values derived from the image content, and a similarity measure is applied to determine which of the database images is most similar to the query. In this work, we present a boosted distance metric (BDM), where individual features are weighted according to their discriminatory power, and compare the performance of this metric to 9 other traditional distance metrics in a CBIR system for digital histopathology. We apply our system to three different breast tissue histology cohorts – (1) 54 breast histology studies corresponding to benign and cancerous images, (2) 36 breast cancer studies corresponding to low and high Bloom-Richardson (BR) grades, and (3) 41 breast cancer studies with high and low levels of lymphocytic infiltration. Over all 3 data cohorts, the BDM performs better compared to 9 traditional metrics, with a greater area under the precision-recall curve. In addition, we performed SVM classification using the BDM along with the traditional metrics, and found that the boosted metric achieves a higher classification accuracy (over 96%) in distinguishing between the tissue classes in each of 3 data cohorts considered. The 10 different similarity metrics were also used to generate similarity matrices between all samples in each of the 3 cohorts. For each cohort, each of the 10 similarity matrices were subjected to normalized cuts, resulting in a reduced dimensional representation of the data samples. The BDM resulted in the best discrimination between tissue classes in the reduced embedding space.

Keywords: breast cancer, CBIR, distance metric, similarity, Boosted Distance Metric (BDM), Graph Embedding, SVM, Precision-Recall, histopathology, lymphocytic infiltration

1. INTRODUCTION

1.1 Content-Based Image Retrieval

Digital image databases have become commonplace thanks to advances in computational storage and processing capabilities. In many applications, it is desirable to compare one image, a "query" to a database of images so the database images can be ranked in order of decreasing similarity. Traditionally, the similarity between two images is measured by matching user-defined keywords that describe both the query and database; however, the manual labeling of these images is time-consuming and subject to variability. Advances in image processing and computational power over the last two decades have led to the development of content-based image retrieval (CBIR) systems. A CBIR system is composed of two main components: *Quantitative Image Representation*, or

Contact Author: Anant Madabhushi (E-mail: anantm@rci.rutgers.edu)

Medical Imaging 2009: Computer-Aided Diagnosis, edited by Nico Karssemeijer, Maryellen L. Giger
Proc. of SPIE Vol. 7260, 72603F · © 2009 SPIE · CCC code: 1605-7422/09/$18 · doi: 10.1117/12.813931

the features that are calculated from the image content, and a *Similarity Metric*, or the method of comparing a query image to the images in the database. CBIR takes advantage of quantitative, objective features extracted from the images to determine the similarity between an image pair, and can rank database images according to this similarity metric. Images can be quantified with global features calculated over the entire scene or structure-based features which attempt to quantify higher-level objects and patterns appearing in the image. In most cases, choosing the method of image representation is application-specific. However, choosing the appropriate similarity metric to compare the feature vectors from the query and database images is a non-trivial task.

1.2 Metrics in CBIR and Classification

Let us consider a repository of N images, $\mathcal{C}_j^r$, $j \in \{1, \cdots, N\}$, where each image belongs to one of 2 particular classes (ω_1, ω_2). Let us denote as $\mathcal{L}(\mathcal{C}_j^r) \in \{+1, -1\}$, the labels used to denote the class of $\mathcal{C}_j^r$. The main objective within CBIR is as follows: given a query scene $\mathcal{C}^q$ and repository image $\mathcal{C}_j^r$, find an appropriate distance function $\mathbb{D}_\psi$ using metric ψ such that $\mathbb{D}_\psi(\mathcal{C}^q, \mathcal{C}_j^r)$ yields a small value for $\mathcal{C}_j^r$, $j \in \{1, \cdots, N\}$ for which $\mathcal{L}(\mathcal{C}^q) = \mathcal{L}(\mathcal{C}_j^r)$ and correspondingly large for $\mathcal{L}(\mathcal{C}^q) \neq \mathcal{L}(\mathcal{C}_j^r)$. Note that in this context, "distance" and "similarity" are inversely proportional: if images are far apart in high-dimensional space, they have a low similarity and are likely to be from different classes. Querying the database thus consists of calculating $\mathbb{D}(\mathcal{C}^q, \mathcal{C}_j^r)$ for all repository images and ordering the results in terms of decreasing similarity. The performance of the CBIR system (and specifically the distance metric ψ) can be determined by examining whether the database images that are most similar to the query have the same class label.

Many applications[1] using distance metrics choose the L^2 or Euclidean norm as a way to measure the similarity between two feature vectors. For instance, several supervised (Support Vector Machines,[2] Nearest Neighbor[3]) and unsupervised clustering and classification schemes employ the L^2 norm, either directly or via a kernel, for measuring object similarity. However, the Euclidean norm may not always be the optimal similarity metric. There are many other distance metrics – Bray Curtis, Canberra, Chebychev, Chi-Squared, Manhattan, Minkowski, Squared Chi-Squared, Squared Chord – that have been proposed, some of which may be more appropriate based on the specific domain or application. A drawback of many of these traditional distance measures, however, is the assumption that all dimensions in the image representation space have an equal contribution to measuring object similarity; that is, each feature is equally weighted in the final similarity calculation.

Supervised learning metrics – i.e. metrics that preferentially weight certain dimensions or features – have been suggested for use in similar applications.[4,5] Athitsos, et al.[6] introduced a method for combining multiple "weak" embeddings of image data into a single "strong" embedding. Each embedding's weight is related to how well a classifier performs on each of the embeddings: a better classification accuracy leading to a higher weight in the final classifier. Yang, et al.[7] proposed using a boosted distance metric for assessing the similarity between mammogram images in an interactive search-assisted diagnosis system. They found that the application of boosting to weight individual features in a distance metric increased both retrieval and classification accuracy over the Euclidean metric.

1.3 Histopathology and CBIR

The medical community has been cited[1] as a major beneficiary of CBIR development. The increase in digital medical image acquisition in routine clinical practice coupled with increases in the quantification and analysis of medical images creates an environment where a CBIR system could provide significant benefits. Müller et al.,[1] provide an overview of CBIR applications in medicine, stating that such applications can include teaching, research, diagnostics, and annotation or classification of medical images. Histopathology, in particular, stands to benefit greatly from CBIR, thanks to the advent of high-resolution digital whole-slide scanners. Quantitative image analysis of histology can help to reduce inter- and intra-observer variability, as well as provide a standard for tissue analysis based solely on qualitative image-derived features. Such systems have found application, among others, in breast[8] and prostate.[9] A study by Caicedo, et al.[10] investigated the development of a CBIR system for basal-cell carcinoma images of histopathology, testing five distance metrics and five different feature types, reporting an average precision rate of 67%. However this study indicated that the best-performing distance metric depends on the feature space. Doyle, et al.[11] investigated the use of different feature types for a CBIR system for breast tissue, and found that the choice of feature space had a large impact on the ability of the system to match query images to database images from the same class.

1.4 Contributions

In this paper we introduce a Boosted Distance Metric (BDM) based on the AdaBoost[12] algorithm, which empirically identifies the most significant image features which contribute the most to the discrimination between different tissue types and weights them accordingly. This similarity metric will be important for applications in both object classification and CBIR. Further, we compare the performance of the BDM against 9 traditional distance metrics: Bray Curtis, Canberra, Chebychev, Chi-Squared, Euclidean, Manhattan, Minkowski, Squared Chi-Squared, and Squared Chord distances. The different metrics are evaluated on three different datasets and 3 corresponding experiments, (1) distinguishing between breast cancer samples with and without lymphocytic infiltration, (2) distinguishing cancerous breast tissue samples from benign samples, and (3) distinguishing high Bloom-Richarcson grade tissue from low grade tissue. Additionally, we employ a support vector machine (SVM) classifier using each of the 9 similarity metrics within the radial basis function (RBF) kernel to classify the query images. Finally, we use the similarity metrics to construct low-dimensional embeddings of the data, allowing us to visualize how well the metrics can distinguish between images from different tissue classes in an alternative data space representation.

In this work, we employ texture features similar to those used in,[11,13] as histopathological images tend to have different texture characteristics due to varying degrees of nuclear proliferation: benign tissue regions contain fewer nuclei when compared to cancerous regions. Further, the arrangement of nuclei in a tissue image plays an important role in describing physiological changes, such as the presence and degree of lymphocytic infiltrate (LI).[14] Thus, we have employed the use of architectural features, which use nuclear centroids to generate graph-based statistics to quantify the content of the images. These features have been shown[11,15] to discriminate well between different tissue types.

The remainder of the paper is organized as follows. Section 2 describes the experimental setup and the specific data sets used in this study. Section 3 discusses the feature extraction process, followed by a detailed description of the BDM algorithm in Section 4. Section 5 explains the evaluation criteria, followed by the experimental results in Section 6 and concluding remarks in Section 7.

2. EXPERIMENTAL DATA: BREAST CANCER HISTOPATHOLOGY

In this work, we focus our analysis on tissue samples taken from the breast. The American Cancer Society predicts over 184,000 new cases of breast cancer (BC) in 2008, leading to 41,000 fatalaties. Proper screening and diagnostic techniques can drastically increase the survival rate of a patient with BC, typically involving a biopsy of a suspicious lesion identified on mammography.[16] Tissue samples are then manually examined under a microscope by a pathologist, and a Bloom Richardson (BR) grade[17] is assigned to the cancer. The BR grading scheme is a systematic way of classifying the degree of BC by analyzing the degree of tumor tubule formation, mitotic activity, and nuclear pleomorphism. BR grade is often critical in deciding treatment options. Unfortunately, grading tends to be qualitative and subject to a high degree of inter-, and even intra-observer variability,[18] sometimes leading to suboptimal treatment decisions. Thus, it is desirable to develop quantitative methods for detecting and analyzing these tissue samples leading in turn to quantitative, reproducible image analysis of the tissue.

For the experiments considered in this study, 3 data cohorts were considered. All 3 cohorts comprised Hematoxylin & Eosin (H&E) stained breast biopsy tissues, scanned into the computer on a whole-slide digital scanner. The first two datasets (Cancer Detection, Cancer Grading) were collected and digitized at the University of Pennsylvania Department of Surgical Pathology, while the third dataset (Lymphocytic Infiltrate (LI)) was obtained and digitized at the Cancer Institute of New Jersey (CINJ). The datasets are described below and are summarized in Table 1.

Cancer Detection (D_{CD}): This dataset consists of 54 digital images of breast biopsy tissue scanned into the computer at 40x optical magnification. Each image represents a region of interest (ROI), and has been manually identified by an expert pathologist as containing either cancerous or benign tissue. Of the 54 images, 18 were classified as benign and 36 were identified as cancerous by 2 expert pathologists.

Cancer Grading Dataset (D_{CG}): This is a subset of the D_{CD} dataset comprised of the 36 cancerous images. However, in this dataset the pathologist identified the BC grade of the tissue, classifying the regions as high grade (12 images) or low grade (24 images).

Lymphocytic Infiltration (D_{LI}): This dataset comprises 41 H&E images of breast tissue obtained at CINJ. The ROI's on these images were chosen according to degree of LI present in the image. A pathologist classified each image either as having a high (22 images) or a low degree (19 images) of LI.

Dataset	Notation	Classes (ω_1 / ω_2)	Class Distribution (ω_1/ω_2)
Cancer Detection	D_{CD}	Cancer/Benign	18/36
Cancer Grading	D_{CG}	High Grade/Low Grade	12/24
Lymphocytic Infiltration	D_{LI}	Infiltrated/Not Infiltrated	22/19

Table 1. Datasets used to evaluate the distance metrics considered in this study.

3. FEATURE EXTRACTION

In order to evaluate and compare the performance of similarity metrics, the images need to first be quantitatively represented through the process of feature extraction. In this way, a single image is converted to a point in a high-dimensional feature space, where it can be compared to other points (images) within the same space. In the following sections, we denote a generic image scene as $\mathcal{C} = (C, f)$, where C is a 2D grid of pixels c and f is a function that assigns an intensity to each pixel $c \in C$. From each $\mathcal{C}$ in each of D_{CD}, D_{CG}, and D_{LI}, we extract image features corresponding to textural characteristics at every pixel $c \in C$. In addition, we extract architectural features for the scenes in D_{LI} in order to quantify the arrangement of the nuclei in the images.[11,15] These features are detailed below and are summarized in Table 2. Our goal with feature extraction is to create a set of K feature operators, Φ_i, for $i \in \{1, \cdots, K\}$, where $\Phi_i(\mathcal{C})$ represents the value of feature i from image scene $\mathcal{C}$. In the following sections, we describe how we obtain the feature operators.

3.1 Texture Features

Texture features are useful in characterizing tissue samples in H&E stained slides, since changes in the proliferation of nuclei result in different staining patterns and different resulting textures.[11] This type of textural data has been quantified using Laws texture features, described below.

Laws Features: These features[19] involve the use of a set of 1-dimensional filters that provide various impulse responses in order to quantify specific patterns in the images. These filters are abbreviated as L (Level), E (Edge), S (Spot), R (Ripple), and W (Wave) owing to the shape of the filters. By multiplying combinations of these filters, we generate 15 unique two-dimensional filter masks Γ_l, for $l \in \{1, \cdots, 15\}$, where $\Gamma_l \in \{\mathrm{LE, LS, LR, LW, ES, ER, EW, SR, SW, RW, LL, EE, SS, RR, WW}\}$. Each of these filters is then convolved with the image scene $\mathcal{C}$ to generate feature scenes $\mathcal{F}_l = \mathcal{C} * \Gamma_l = (C, g_l)$, for $l \in \{1, \cdots, 15\}$, where $g_l(c)$ is a function that assigns a value from feature l to pixel $c \in C$. Examples of these feature scenes are shown in Figure 1 for a breast tissue histology scene in D_{CG}. We calculate the following statistics for each l:

$$\Phi_1(\mathcal{C}) = \frac{1}{|C|} \sum_{c \in C} g_l(c), \tag{1}$$

$$\Phi_2(\mathcal{C}) = \sqrt{\frac{1}{|C|} \sum_{c \in C} (g_l(c) - \Phi_1(\mathcal{C}))^2}, \tag{2}$$

$$\Phi_3(\mathcal{C}) = \frac{\frac{1}{|C|} \sum_{c \in C} (g_l(c) - \Phi_1(\mathcal{C})^4}{\left(\frac{1}{|C|} \sum_{c \in C} (g_l(c) - \Phi_1(\mathcal{C})^2\right)^2} - 3, \tag{3}$$

$$\Phi_4(\mathcal{C}) = \mathrm{MEDIAN}_{c \in C} \left[g_l(c) \right], \tag{4}$$

$$\Phi_5(\mathcal{C}) = \frac{\frac{1}{|C|}\sum_{c\in C}(g_l(c) - \Phi_1(\mathcal{C})^3}{\left(\frac{1}{|C|}\sum_{c\in C}(g_l(c) - \Phi_1(\mathcal{C})^2\right)^{3/2}}, \tag{5}$$

where $\Phi_1(\mathcal{C})$ through $\Phi_5(\mathcal{C})$ represents the average, standard deviation, kurtosis, median, and skewness of feature values, respectively. These are calculated from each $\mathcal{F}_l$, $l \in \{1, \cdots, 15\}$, yielding a total of 75 Laws feature values.

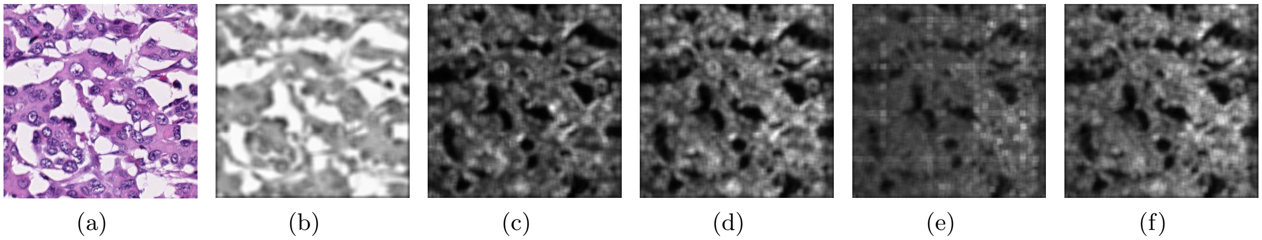


Figure 1. Examples of the feature scenes $\mathcal{F}_l$ used to generate the feature values described in Section 3. Shown are (a) an example of a tissue image and the texture images generated from the following filters: (b) LL, (c) EE, (d) SS, (e) RR, and (f) WW.

3.2 Architectural Features

To extract quantifiable image attributes from the graph-based representations of the image, we first manually label the nuclear centroids within each of the images. A pixel at the centroid of a nucleus is denoted $\hat{c}$, where $\hat{c}^k$ refers to the kth nucleus in the image, for $k \in \{1, \cdots, m\}$. The following features were extracted exclusively for the D_{LI} dataset.

Voronoi Diagram: The Voronoi Diagram is defined by a set of polygons $\mathbf{P} = \{P_1, P_2, \ldots, P_m\}$, where each polygon is constructed around a nuclear centroid by adding surrounding pixels to the nearest centroid. Thus, P_a is constructed around $\hat{c}^a$ by adding pixels for which $\mathbb{D}_{\mathrm{EU}}(c, \hat{c}^a) = \min_k ||c - \hat{c}^k||$ where $a, k \in \{1, 2, \ldots, m\}$; that is, each non-centroid pixel is added to the polygon of the nearest centroid pixel. The metric $\mathbb{D}_{\mathrm{EU}}(c, \hat{c}^a)$ is defined as the Euclidean distance between pixels $c, \hat{c}^a \in C$. Area, perimeter length, and chord length are calculated for all polygons in the image, and the average, standard deviation, min/max ratio, and measurement of disorder are calculated across the entire graph to yield 12 Voronoi features for every image.

Delaunay Triangulation: The Delaunay graph is constructed such that if two unique polygons P_a and P_b, where $a, b \in \{1, \cdots, m\}$ from the Voronoi graph share a side, their nuclear centroids $\hat{c}^a$ and $\hat{c}^b$ are connected by and edge, denoted $E^{a,b}$. The collection of all edges constitutes the Delaunay graph, which is a triangulation connecting each nuclear centroid. The edge lengths and triangle areas are calculated for all triangles in the image, and the average, standard deviation, min/max ratio, and measurement of disorder is calculated across the graph, yielding 10 Voronoi features for every image.

Minimum Spanning Tree: A spanning tree graph $\mathcal{G}$ is a connected, undirected graph connecting all vertices (nuclei) in an image. For any set of vertices, there may be many $\mathcal{G}$. In each $\mathcal{G}$, weights denoted $w_{\mathcal{G}}^E$ are assigned to each edge E based on the length of E and $\mathcal{G}$. The sum of all weights in $\mathcal{G}$ determines the characteristic weight, $\hat{w}_{\mathcal{G}}$ assigned to each $\mathcal{G}$. The minimum spanning tree, denoted as $\mathcal{G}'$, has a weight $\hat{w}_{\mathcal{G}'} \leq \hat{w}_{\mathcal{G}}$ for every other spanning tree $\mathcal{G}$. From $\mathcal{G}'$, we calculate the average, standard deviation, min/max ratio, and disorder of the branch lengths, yielding 4 minimum spanning tree features for every image.

Nuclear Features: Finally, we calculate several non-graph based features from the arrangement of the nuclei in the image. Nuclear density is computed as $\frac{m}{|C|}$, where $|C|$ is the cardinality of C. For each nuclear centroid $\hat{c}^a$, $\mathcal{N}(\rho, \hat{c}^a)$ is the set of pixels $c \in C$ contained within a circle with its center at $\hat{c}^a$ and radius ρ. The number of nuclear centroids $\hat{c}^k$, for $k \neq a$ and $k, a \in \{1, \cdots, m\}$, contained within $\mathcal{N}(\rho, \hat{c}^a)$ are computed for $\rho \in \{10, 20, \cdots, 50\}$. Additionally, the radius ρ required for $\mathcal{N}(\rho, \hat{c}^a)$ to contain 3, 5, and 7 additional nuclei are also computed. The mean, standard deviation, and disorder of these values for all $\hat{c}^a$ in C are calculated to provide 25 features for each $\mathcal{C}$.

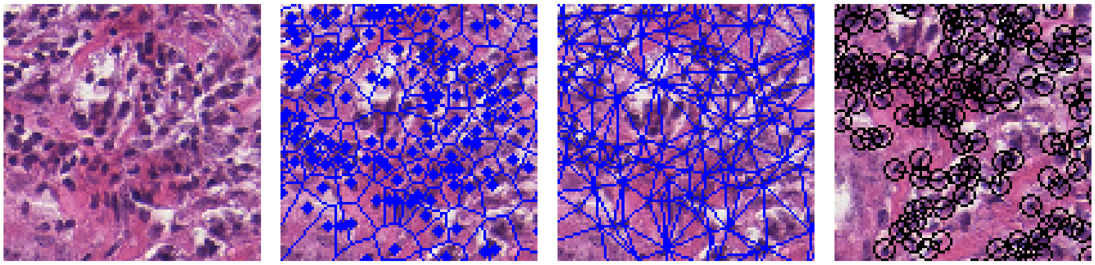


Figure 2. Illustration of the graphs used to generate the feature values described in Section 3 for the D_{LI} dataset. Shown are: (a) the original image, (b) the Voronoi Diagram, (c) the Delaunay Triangulation, and (d) the Minimum Spanning Tree.

Feature Type (Dataset)	**Feature Name**	**Number of Features**
Texture (D_{CD}, D_{CG}, D_{LI})	Laws Filters	75
Architecture (D_{LI})	Voronoi Diagram	12
	Delaunay Triangulation	10
	Minimum Spanning Tree	4
	Nuclear Features	25

Table 2. Features used to quantify the tissue images.

4. BOOSTED DISTANCE METRIC

The purpose of the Boosted Distance Metric (BDM) is to construct a weighted metric where each feature's contribution to the distance between two points is related to its ability to distinguish between object classes. The construction of the BDM is a three-step process: (1) For each feature extracted above, we construct a weak classifier h_i, $i \in \{1, \cdots, K\}$, such that $h_i(\mathcal{C}) \in \{+1, -1\}$ gives a classification label for $\mathcal{C}$. (2) We identify the T most accurate classifiers h_t, $t \in \{1, \cdots, T\}$, and learn the weights α_t associated with each classifier via AdaBoost.[12] (3) Finally, we build the BDM using the weight / classifier pairs (α_t, h_t). Note that independent BDM's are separately learned for each of D_{CD}, D_{CG}, D_{LI}. The details of each step are given below.

Constructing Weak Classifiers: The weak classifiers h_i are generated in the following manner.

1. The feature extraction process involves the use of a feature operator $\Phi_i, i \in \{1, \cdots, K\}$, where for any image scene $\mathcal{C}$, $\Phi_i(\mathcal{C})$ yields a single feature value. Each image has a class label $\mathcal{L}(\mathcal{C}) \in \{+1, -1\}$.
2. We obtain our training set, $\mathcal{S}^{\mathrm{tr}} = \{\mathcal{C}^r_j | j \in \{1, \cdots, N\}\}$, as two-thirds of the total dataset. We obtain two class distributions $\mathcal{B}^+_i, \mathcal{B}^-_i$, for $i \in \{1, \cdots, K\}$.
3. Class distribution means of $\mathcal{B}^+_i$ and $\mathcal{B}^-_i$, $i \in \{1, \cdots, K\}$, are estimated as $\mu(\mathcal{B}^+_i)$ and $\mu(\mathcal{B}^-_i)$, respectively.
4. We define the separating plane between $\mu(\mathcal{B}^+_i)$ and $\mu(\mathcal{B}^-_i), i \in \{1, \cdots, K\}$, as
$$\beta_i = \min[\mu(\mathcal{B}^+_i), \mu(\mathcal{B}^-_i)] + \frac{|\mu(\mathcal{B}^+_i) - \mu(\mathcal{B}^-_i)|}{2}. \tag{6}$$
5. For each $i \in \{1, \cdots, K\}$, define a weak classifier h_i such that for any query scene $\mathcal{C}^q$:
$$h_i(\mathcal{C}^q) = \begin{cases} +1, & \text{if } \Phi_i(\mathcal{C}^q) \geq \beta_i, \\ -1, & \text{otherwise.} \end{cases} \tag{7}$$

Learning Feature Weights via AdaBoost: We use the AdaBoost algorithm[12] to select class discriminatory weak classifiers and learn the associated weights. The AdaBoost algorithm operates in an iterative fashion, choosing the best-performing weak classifiers and assigning weights according to the classification accuracy of that feature. The algorithm maintains an error-weighting distribution, Π, to ensure that subsequent features focus on difficult to classify samples. The output of the algorithm is a set of selected weak classifiers, h_t, and associated weights, α_t, for $t \in \{1, \cdots, T\}$, where $1 \leq T \leq K$. The algorithm is given below.

```
Algorithm BoostMetricWeights()
Input:  $\mathcal{S}^{\mathrm{tr}}$, $\mathcal{L}(\mathcal{C}_j^r)$ for $j \in \{1, \cdots, N\}$, iterations $T$, weak classifiers $h_i$ for $i \in \{1, \cdots, K\}$.
Output: Selected classifiers $h_t$, associated weights $\alpha_t$.
begin
    0. Initialize distribution $\Pi_1(j) = \frac{1}{N}, j \in \{1, ..., N\}$;
    1. for $t = 1$ to $T$ do
    2.     Find $h_t = \arg\min_{h_i} \epsilon_i$, where $\epsilon_i = \sum_{j=1}^{N} \Pi_t(j)[\mathcal{L}(\mathcal{C}_j^r) \neq h_i(\mathcal{C}_j^r)]$;
    3.     if $\epsilon_t \geq 0.5$ then stop;
    4.     $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$;
    5.     Update Distribution, $\Pi_{t+1}(j) = \frac{1}{Z_t} \Pi_t(j) e^{-\alpha_t \mathcal{L}(\mathcal{C}_j^r) h_t(\mathcal{C}_j^r)}$ for all $j$, where $Z_t$ is a normalization term;
    6.     Output $\alpha_t$;
    7. endfor;
end
```

Note that the same feature may be chosen twice in this algorithm, and some features may not be chosen at all if their classification error ϵ_i is consistently above 0.5. In this work we chose $T = 10$.

Constructing the BDM: Once the weights and features have been chosen, the BDM is constructed. To find the distance between query image $\mathcal{C}^q$ and repository image $\mathcal{C}_j^r$, we calculate

$$\mathbb{D}_{\mathrm{BDM}}(\mathcal{C}^q, \mathcal{C}_j^r) = \left[\frac{1}{T} \sum_{t=1}^{T} \alpha_t \| \Phi_t(\mathcal{C}^q) - \Phi_t(\mathcal{C}_j^r) \|_2 \right]^{\frac{1}{2}}. \tag{8}$$

5. EVALUATION OF DISTANCE METRICS

5.1 Precision Recall Curves

In a CBIR system, performance is typically judged based on how many retrieved images for a given query image are "relevant" to the query, and where in the retrieval list order they appear. For our purposes, a repository image $\mathcal{C}_j^r$ is considered relevant if $\mathcal{L}(\mathcal{C}^q) = \mathcal{L}(\mathcal{C}_j^r)$. The ability to recall relevant images is evaluated with respect to two statistics: precision, or the ability to retrieve *only* relevant images, and recall, or the ability to retrieve *all available* relevant images. The training set $\mathcal{S}^{\mathrm{tr}}$ acts as our database, and is the source of our repository images $\mathcal{C}_j^r \in \mathcal{S}^{\mathrm{tr}}, j \in \{1, \cdots, N\}$. The testing set $\mathcal{S}^{\mathrm{te}}$ is the source for the query images, $\mathcal{C}^q \in \mathcal{S}^{\mathrm{te}}$.

First, a query image is selected from the testing set and features are extracted via the feature operator, $\Phi_i(\mathcal{C}^q)$, $i \in \{1, \cdots, K\}$, as described in Section 3. For each repository image $\mathcal{C}_j^r$, $j \in \{1, \cdots, N\}$, the distance is calculated: $\mathbb{D}_\psi(\mathcal{C}^q, \mathcal{C}_j^r)$, where ψ is one of the metrics listed in Table 3. The repository images are arranged in a retrieval list in order of increasing distance (i.e. decreasing similarity). Denoting by R, the number of retrieved images, R' as the number of relevant retrieved images, and N' as the number of relevant images in the database, the precision is calculated as $\frac{R'}{R}$, and recall is calculated as $\frac{R'}{N'}$, for $R \in \{1, \cdots, N\}$. By calculating precision and recall for all values of R, we can build a precision-recall curve (PRC), which illustrates the ability of the system to retrieve relevant images from the database in the appropriate order of similarity. Interpretation of the PRC is similar to a receiver operating characteristic (ROC) curve, where the area under the curve is a measure of how well different metrics perform on the same retrieval task. The average PRC over all $\mathcal{C}^q \in \mathcal{S}^{\mathrm{te}}$ is calculated to represent the performance of each metric ψ listed in Table 3. Three-fold cross-validation is performed to ensure that all images are used as both query and repository images and thus prevent overfitting the BDM.

5.2 Support Vector Machine (SVM) Classifier

In order to additionally evaluate the discriminative power of the metrics, we perform SVM classification to classify the query image using the repository as the training data.[2] An SVM classifier uses a kernel function denoted as $\mathcal{K}(\cdot, \cdot)$ to project training data to a higher-dimensional space, where a hyperplane is established that separates out the different classes. Testing data is then projected into this same space, and is classified according to where the test objects fall with respect to the hyperplane. A modified version of the common radial basis function (RBF) was used in this study:

$$\mathcal{K}(\mathcal{C}^q, \mathcal{C}^r_j) = e^{-\gamma \mathbb{D}_\psi(\mathcal{C}^q, \mathcal{C}^r_j)}, \tag{9}$$

where γ is a parameter for normalizing the inputs, $\mathcal{C}^q$ is a query image, and $\mathcal{C}^r_j$ is a repository image scene, for $j \in \{1, \cdots, N\}$. In our formulation, the RBF kernel determines the high-dimensional projection of the inputs using the distance function $\mathbb{D}_\psi$ utilizing metric ψ. The intuition is that the separation of the objects in the projected high dimensional space will reflect the performance of the distance metric. The general form of the SVM classifier is:

$$\Theta = \sum_{\kappa=1}^{\widehat{N}} \xi_\kappa \mathcal{L}(\mathcal{C}^r_\kappa) \mathcal{K}(\mathcal{C}^q, \mathcal{C}^r_\kappa) + \mathbf{b}, \tag{10}$$

where $\mathcal{C}^r_\kappa$ denotes the subset of the overall training data acting as the $\widehat{N}$ support vectors, $\kappa \in \{1, \cdots, \widehat{N}\}$, $\mathbf{b}$ is a bias obtained from the training set to maximize the distance between the support vectors, and ξ is a model parameter obtained via maximization of an objective function.[2] The output of Equation 10 represents the distance between the query image $\mathcal{C}^q$ and the decision hyperplane midway between the support vectors, while the sign of the distance indicates class membership. We can construct a strong classifier, $\mathbf{h}_{\text{SVM}}$, where the classification label of the query image $\mathcal{C}^q$ is given as:

$$\mathbf{h}_{\text{SVM}}(\mathcal{C}^q) = \begin{cases} +1, & \text{if } \Theta \geq 0, \\ -1, & \text{otherwise.} \end{cases} \tag{11}$$

The parameters γ and $\mathbf{b}$ are found through randomized three-fold cross validation. We run over 50 trials per experiment. The percentage of correctly classified images is recorded for each trial.

5.3 Low-Dimensional Embedding

The distance metrics were also evaluated in terms of their ability to preserve object-class relationships while projecting the data from a high to reduced dimensional space, the reduced dimensional representation being obtained via a non-linear dimensionality reduction scheme called Graph Embedding. In previous work,[20] we have demonstrated the utility of non-linear DR schemes over linear schemes (PCA) for representing biomedical data. For each distance metric $\psi \in \{\text{Bray Curtis}, \cdots, \text{BDM}\}$ given in Table 3, we perform the following steps.

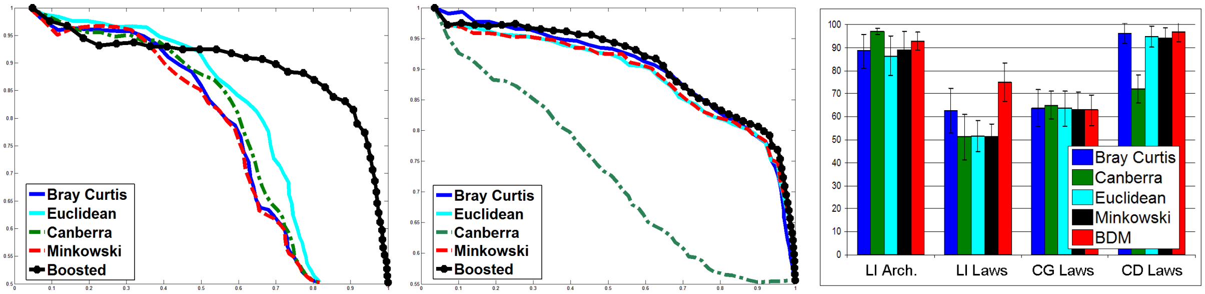
(a) Construct a similarity matrix $W_\psi(u, v) = e^{-\mathbb{D}_\psi(\mathcal{C}^r_u, \mathcal{C}^r_v)}$, for $u, v \in \{1, \cdots, N\}$.

(b) Find the embedding vector $\mathcal{X}$ by maximizing the function:

$$\mathcal{E}_{W_\psi}(\mathcal{X}) = 2\eta \frac{\mathcal{X}^{\mathbf{T}}(Y - W_\psi)\mathcal{X}}{\mathcal{X}^{\mathbf{T}} Y \mathcal{X}}, \tag{12}$$

where $Y(u, u) = \sum_v W_\psi(u, v)$ and $\eta = N - 1$.

(c) The d-dimensional embedding space is defined by the eigenvectors corresponding to the smallest d eigenvalues of $(Y - W_\psi)\mathcal{X} = \lambda Y \mathcal{X}$. For any image $\mathcal{C}$, the embedding $\mathcal{X}(\mathcal{C})$ contains the coordinates of the image in the embedding space and is given as $\mathcal{X}(\mathcal{C}) = [w_z(\mathcal{C}) | z \in \{1, \cdots, d\}]$, where $w_z(\mathcal{C})$ are the z eigenvalues associated with $\mathcal{X}(\mathcal{C})$.



(a) (b) (c)

Figure 3. Precision Recall Curves generated by retrieval of repository images in the high dimensional space using a subset of the metrics (Bray Curtis, Euclidean, Canberra, Minkowski, and BDM) in this study. Shown are the results on: (a) D_{LI} using architectural features and (b) D_{CD} using Laws texture features. (c) SVM classification accuracy in the high-dimensional space for the subset of metrics listed above. Each bar represents a distance metric, and each group of bars represents a dataset. From left to right are: $D_{\rm LI}$ using architectural features, $D_{\rm LI}$ using Laws features, $D_{\rm CG}$, and $D_{\rm CD}$.

6. EXPERIMENTAL RESULTS AND DISCUSSION

The distance metrics described in Table 3 are first evaluated in the original high-dimensional feature space. Each metric operates on two high-dimensional feature vectors, whose elements are defined by the feature operator Φ_i, for $i \in \{1, \cdots, K\}$. For two images $\mathcal{C}^q$ and $\mathcal{C}^r_j$, we define the distance between the images as a single scalar value given by $\mathbb{D}_\psi(\mathcal{C}^q, \mathcal{C}^r_j)$. Note the Minkowski distance, which is parameterized by θ; when $\theta = 2$, this is equivalent to the Euclidean distance, and as $\theta \to \infty$, it becomes the Chebychev distance. For our experiments, $\theta = 3$.

For each experiment given below, the dataset is randomly separated into thirds: two-thirds of the dataset constitute the training set $\mathcal{S}^{\rm tr}$, while the remaining one-third is the testing set $\mathcal{S}^{\rm te}$.

6.1 Distance Metrics in High Dimensional Space

Distance Metric ψ	**Formula** $\mathbb{D}_\psi$
Bray Curtis	$\frac{\sum_{i=1}^K \lvert\Phi_i(\mathcal{C}^q) - \Phi_i(\mathcal{C}^r_j)\rvert}{\sum_{i=1}^K \Phi_i(\mathcal{C}^q) + \Phi_i(\mathcal{C}^r_j)}$
Canberra	$\frac{\sum_{i=1}^K \lvert\Phi_i(\mathcal{C}^q) - \Phi_i(\mathcal{C}^r_j)\rvert}{\sum_{i=1}^K \lvert\Phi_i(\mathcal{C}^q) + \Phi_i(\mathcal{C}^r_j)\rvert}$
Chebychev	$\lim_{\tau \to \infty} \left[\sum_{i=1}^K (\lvert\Phi_i(\mathcal{C}^q) - \Phi_i(\mathcal{C}^r_j)\rvert^\tau)\right]^{\frac{1}{\tau}}$
Chi-Squared	$\sum_{i=1}^K \frac{(\Phi_i(\mathcal{C}^q) - \Phi_i(\mathcal{C}^r_j))^2}{\Phi_i(\mathcal{C}^r_j)}$
Euclidean	$\sqrt{\sum_{i=1}^K (\Phi_i(\mathcal{C}^q) - \Phi_i(\mathcal{C}^r_j))^2}$
Manhattan	$\sum_{i=1}^K \lvert\Phi_i(\mathcal{C}^q) - \Phi_i(\mathcal{C}^r_j)\rvert$
Minkowski	$\left[\sum_{i=1}^K \lvert\Phi_i(\mathcal{C}^q) - \Phi_i(\mathcal{C}^r_j)^\theta\right]^{\frac{1}{\theta}}$
Squared Chi-Squared	$\sum_{i=1}^K \frac{(\Phi_i(\mathcal{C}^q) - \Phi_i(\mathcal{C}^r_j))^2}{\lvert\Phi_i(\mathcal{C}^q) + \Phi_i(\mathcal{C}^r_j)\rvert}$
Squared Chord	$\sum_{i=1}^K (\sqrt{\Phi_i(\mathcal{C}^q)} - \sqrt{\Phi_i(\mathcal{C}^r_j)})^2$
BDM	$\left[\frac{1}{T}\sum_{t=1}^T \alpha_t \lVert\Phi_t(\mathcal{C}^q) - \Phi_t(\mathcal{C}^r_j)\rVert_2\right]^{\frac{1}{2}}$

Table 3. Distance metrics used in this study, operating on a query image scene $\mathcal{C}^q$ and a repository scene $\mathcal{C}^r_j$ for $j \in \{1, \cdots, N\}$.

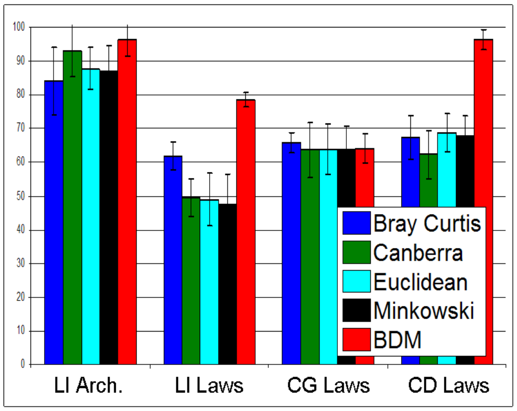


Figure 4. SVM classification accuracy in the low-dimensional space for the subset of metrics (Bray Curtis, Euclidean, Canberra, Minkowski, and BDM) used in this study. Each bar represents a distance metric, and each group of bars represents a dataset and a specific feature type. From left to right are: D_{LI} using architectural features, D_{LI} using Laws features, D_{CG}, and D_{CD}.

Precision Recall Curves: Average PR curves are generated for all distance metrics ψ listed in Table 3. The PR curves resulting from the CBIR query and retrieval tasks for datasets D_{LI} and D_{CD} and for $\psi \in \{\mathrm{Bray\ Curtis}, \mathrm{Euclidean}, \mathrm{Canberra}, \mathrm{Minkowski}, \mathrm{BDM}\}$, are illustrated in Figures 3(a), 3(b), respectively. In both Figures 3(a), (b), the black dotted curve represents the BDM. The area under the PRC for the BDM is higher in both Fig. 3(a) and (b) compared to the Bray Curtis, Euclidean, Canberra, and Minkowski distances.

Support Vector Machines: Figure 3 (c) illustrates the SVM classification accuracy for each of D_{LI}, D_{CG}, D_{CD} and for metrics $\psi \in \{\mathrm{Bray\ Curtis}, \mathrm{Euclidean}, \mathrm{Canberra}, \mathrm{Minkowski}, \mathrm{BDM}\}$ over 50 trials using three-fold cross-validation. The accuracies are broken up according to classification task, with the D_{LI} dataset evaluated using both the Laws texture features and the architectural features. In order from left to right: the blue bar represents the Bray Curtis metric, green represents Canberra, cyan represents Euclidean, red represents Minkowski, and black represents the BDM. Note that for all feature types, the BDM performs comparably to the other metrics.

6.2 Distance Metrics in Low Dimensional Space

Support Vector Machines: Figure 4 plots the SVM classification accuracies for each of D_{LI}, D_{CG}, D_{CD} and for metrics $\psi \in \{\mathrm{Bray\ Curtis}, \mathrm{Euclidean}, \mathrm{Canberra}, \mathrm{Minkowski}, \mathrm{BDM}\}$ in the low-dimensional classification experiments. In most classification tasks the BDM out-performs the other metrics in this study. Table 4 shows the classification accuracies across all metrics for each of the three datasets. These results not only reflect the BDM's efficacy in the original feature space, but also its ability to preserve object-class relationships in the low dimensional data transformation, where the SVM is able to distinguish the different classes.

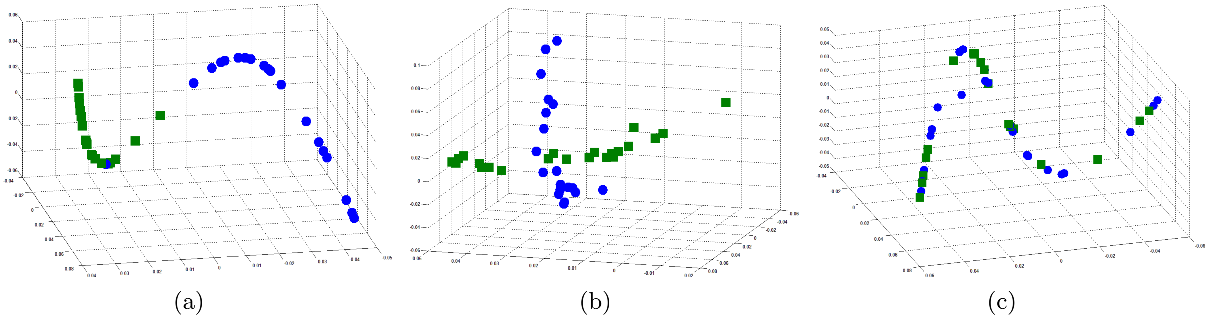
Low-Dimensional Embedding: Figure 5 is an example of the low-dimensional embedding produced for the D_{LI} dataset using three different metrics. Figure 5(a) is a three-dimensional embedding obtained via the BDM, while Figure 5(b) was obtained using the Canberra distance metric and Figure 5(c) using the Minkowski distance metric. In these plots, green squares indicate the LI class, while blue circles indicate the non-infiltrated class. The separation between the two classes is much more apparent in the embedding space obtained via the BDM compared to the embeddings obtained via the non-boosted metrics.

7. CONCLUDING REMARKS

In this paper, we have introduced a Boosted Distance Metric that was shown to be capable of effectively calculating the similarity between high-dimensional feature vectors in a medical CBIR application. We have demonstrated that by weighting individual features, the BDM out-performs many traditional metrics, including the commonly-used Euclidean distance metric. We evaluated the performance of BDM with respect to 9 other similarity metrics on 3 different datasets and using a large number of textural and graph-based image features. The

ψ	Mean Classification Accuracy (Standard Deviation)		
	D_{CD}	D_{CG}	D_{LI}
Bray Curtis	67.33 (6.41)	**65.92 (3.00)**	84.12 (4.28)
Canberra	62.39 (7.17)	63.67 (8.08)	93.14 (5.63)
Chebychev	67.00 (8.83)	65.58 (3.84)	83.29 (10.91)
Chi-Squared	68.89 (6.66)	64.50 (6.53)	87.71 (7.54)
Euclidean	68.72 (5.66)	63.92 (7.46)	87.86 (7.70)
Manhattan	67.50 (7.64)	63.00 (6.87)	89.93 (6.91)
Minkowski	67.78 (6.07)	63.75 (7.16)	87.07 (8.83)
Squared Chi-Squared	67.28 (6.43)	65.08 (4.03)	87.71 (5.50)
Squared Chord	67.61 (5.46)	63.42 (5.85)	87.14 (7.11)
Boosted Weight Metric	**96.30 (2.90)**	64.11 (4.23)	**96.20 (2.27)**

Table 4. SVM Classification accuracy percentages in the low dimensional space, for all distance metrics and all three datasets (D_{LI}, D_{CG}, D_{CD}). Standard deviation is shown in parentheses. Highest accuracy in each dataset shown in boldface.



(a) (b) (c)

Figure 5. Low-dimensional embeddings of the high-dimensional data obtained by using (a) the BDM, (b) the Canberra, and (c) the Minkowski distance metrics on the D_{LI} study using architectural features. Green squares indicate the LI class, while blue circles indicate the non-infiltrated class.

similarity metrics were evaluated via (a) precision-recall curves, (b) SVM classification accuracy, and (c) in terms of their ability to preserve object-class relationships from a high to a reduced dimensional space, via a non-linear dimensionality reduction scheme. For all 3 evaluation criteria, the BDM was superior or comparable to the 9 other distance metrics. Our initial results suggest that for focused biomedical applications, such as CBIR for histopathology, a supervised learning metric may be a better choice compared to traditional measures that do not consider feature weighting. In future work, we intend to evaluate the BDM on a much larger cohort of data.

8. ACKNOWLEDGMENTS

This work made possible via grants from Coulter Foundation (WHCF 4-29368), New Jersey Commission on Cancer Research, National Cancer Institute (R21CA127186-01, R03CA128081-01), the Society for Imaging Informatics in Medicine (SIIM), the Life Science Commercialization Award, the Aresty Undergraduate Award, and the US Department of Defense (427327).

REFERENCES

[1] Muller, H. et al., "A review of content-based image retrieval systems in medical applications - clinical benefits and future directions," *Int. J. of Med. Inf.* **73**(1), 1–23 (2003).

[2] Cortes, C. and Vapnik, V., "Support-vector networks," *Machine Learning* **20**, 273–297 (1995).

[3] Richard O. Duda, Peter E. Hart, D. G. S., [*Pattern Classification*], Wiley (2001).

[4] Zhou, G. et al., "Content-based cell pathology image retrieval by combining different features," in [*SPIE Conf. Series*], **5371**, 326–333 (2004).

[5] Yu, J. et al., "Toward robust distance metric analysis for similarity estimation," in [*CVPR'06*], **1**, 316–322 (2006).

[6] Athitsos, V., Alon, J., Sclaroff, S., and Kollios, G., "Boostmap: An embedding method for efficient nearest neighbor retrieval," *IEEE Trans. on Patt. Recog. and Mach. Learn.* **30**, 89–104 (Jan. 2008).

[7] Yang, L. et al., "Learning distance metrics for interactive search-assisted diagnosis of mammograms," *Proc. of SPIE* (2007).

[8] Choi, H. et al., "Design of the breast carcinoma cell bank system," in [*HEALTHCOM 2004*], 88–91 (2004).

[9] Wetzel, A. W. et al., "Evaluation of prostate tumor grades by content-based image retrieval," *27th AIPR Workshop: Adv. in Comp. Assist. Recog.* **3584**(1), 244–252, SPIE (1999).

[10] Caicedo, J. C. et al., "Design of a medical image database with content-based retrieval capabilities," in [*PSIVT*], *LNCS* **4872**, 919–931 (2007).

[11] Doyle, S. et al., "Using manifold learning for content-based image retrieval of prostate histopathology," in [*Workshop on CBIR for Biomedical Image Archives, (MICCAI)*], (2007).

[12] Schapire, R., "The boosting approach to machine learning: An overview," *Nonlin. Est. and Class.* (2003).

[13] Doyle, S. et al., "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in [*5th IEEE Int. Symp. ISBI 2008*], 496–499 (2008).

[14] Basavanhally, A. et al., "Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade breast cancer histology," in [*MIAAB (in conjunction with MICCAI)*], (2008).

[15] Sudb, J., Marcelpoil, R., and Reith, A., "New algorithms based on the voronoi diagram applied in a pilot study on normal mucosa and carcinomas.," *Anal Cell Pathol* **21**(2), 71–86 (2000).

[16] Saslow, D. et al., "American cancer society guidelines for breast screening with mri as an adjunct to mammography," *CA Cancer J Clin* **57**, 75–89 (2007).

[17] Bloom H.J., R. W., "Histological grading and prognosis in breast cancer," *Br. J. Cancer* **11**, 359–377 (1957).

[18] Dalton, L. W. et al., "Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement.," *Mod Pathol* **13**(7), 730–735 (2000).

[19] Laws, K. I., "Rapid texture identification," in [*SPIE Conf. Series*], **238**, 376–380 (1980).

[20] Lee, G., Rodriguez, C., and Madabhushi, A., "Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies," *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **5**(3), 368–384 (2008).