IRSES Project 247091 MIRACLE

WP1, Deliverable 1: Current Methods of Feature Extraction for Microscopic Images

Note: This report is based on Gurcan, M.N.; Boucheron, L.E.; Can, A.; Madabhushi, A.; Rajpoot, N.M.; Yener, B.; , "Histopathological Image Analysis: A Review," Biomedical Engineering, IEEE Reviews in , vol.2, pp.147-171, 2009. The first and fifth authors are members of OSU and WARWICK, respectively.

Introduction

The widespread use of Computer-assisted diagnosis (CAD) can be traced back to the emergence of digital mammography in the early 1990's [1]. Recently, CAD has become a part of routine clinical detection of breast cancer on mammograms at many screening sites and hospitals [2] in the United States. In fact, CAD has become one of the major research subjects in medical imaging and diagnostic radiology. Given recent advances in high-throughput tissue bank and archiving of digitized histological studies, it is now possible to use histological tissue patterns with computer-aided image analysis to facilitate disease classification. There is also a pressing need for CAD to relieve the workload on pathologists by sieving out obviously benign areas, so that pathologist can focus on the more difficult-to-diagnose suspicious cases. For example, approximately 80% of the 1 million prostate biopsies performed in the US every year are benign; this suggests that prostate pathologists are spending 80% of their time sieving through benign tissue.

Researchers both in the image analysis and pathology fields have recognized the importance of quantitative analysis of pathology images. Since most current pathology diagnosis is based on the subjective (but educated) opinion of pathologists, there is clearly a need for quantitative image-based assessment of digital pathology slides. This quantitative analysis of digital pathology is important not only from a diagnostic perspective, but also in order to understand the underlying reasons for a specific diagnosis being rendered (e.g., specific chromatin texture in the cancerous nuclei which may indicate certain genetic abnormalities). In addition, quantitative characterization of pathology imagery is important not only for clinical applications (e.g., to reduce/eliminate inter- and intra-observer variations in diagnosis) but also for research applications (e.g., to understand the biological mechanisms of the disease process).

A large focus of pathological image analysis has been on the automated analysis of cytology imagery. Since cytology imagery often results from the least invasive biopsies (e.g., the cervical Pap smear), they are some of the most commonly encountered imagery for both disease screening and biopsy purposes. Additionally, the characteristics of cytology imagery, namely the presence of isolated cells and cell clusters in the images and the absence of more complicated

structures such as glands make it easier to analyze these specimens compared to histopathology. For example, the segmentation of individual cells or nuclei is a relatively easier process in such imagery since most of the cells are inherently separated from each other.

Histopathology slides, on the other hand, provide a more comprehensive view of disease and its effect on tissues, since the preparation process preserves the underlying tissue architecture. As such, some disease characteristics, e.g., lymphocytic infiltration of cancer, may be deduced only from a histopathology image. Additionally, the diagnosis from a histopathology image remains the 'gold standard' in diagnosing considerable number of diseases including almost all types of cancer [3]. The additional structure in these images, while providing a wealth of information, also presents a new set of challenges from an automated image analysis perspective. It is expected that the proper leverage of this spatial information will allow for more specific characterizations of the imagery from a diagnostic perspective. The analysis of histopathology imagery has generally followed directly from techniques used to analyze cytology imagery. In particular, certain characteristics of nuclei are hallmarks of cancerous conditions. Thus, quantitative metrics for cancerous nuclei were developed to appropriately encompass the general observations of the experienced pathologist, and were tested on cytology imagery. These same metrics can also be applied to histopathological imagery, provided histological structures such as cell nuclei, glands, and lymphocytes have been adequately segmented (a complication due to the complex structure of histopathological imagery). The analysis of the spatial structure of histopathology imagery can be traced back to the works of Wiend *et al*. [4], Bartels [5] and Hamilton [6] but has largely been overlooked perhaps due to the lack of computational resources and the relatively high cost of digital imaging equipment for pathology. However, spatial analysis of histopathology imagery has recently become the backbone of most automated histopathology image analysis techniques. Despite the progress made in this area thus far, this is still a large area of open research due to the variety of imaging methods and disease-specific characteristics.

Feature extraction

Research on useful features for disease classification has often been inspired by visual attributes defined by clinicians as particularly important for disease grading and diagnosis. The vast majority of these features are nuclear features, and many have been established as useful in analysis of both cytopathology and histopathology imagery. Other features that assume discriminatory importance include the margin and boundary appearance of ductal, stromal, tubular and glandular structures. While there is a compilation of features for cytopathology imagery [10], there is relatively little such work for histopathology imagery.

Humans' concept of the world is inherently object-based, as opposed to the largely pixel-based representation of computer vision. As such, human experts

describe and understand images in terms of such objects. For pathologists, diagnosis criteria are inevitably described using terms such as “nucleus” and “cell.” It is thus important to develop computer vision methods capable of such object-level analysis.

TABLE I
SUMMARY OF OBJECT-LEVEL FEATURES USED IN HISTOPATHOLOGY IMAGE ANALYSIS

Category	Features
Size and Shape	Area
	Elliptical Features: Major and minor axis length, eccentricity, orientation, elliptical deviation
	Convex Hull Features: Convex area, convex deficiency, solidity
	Filled Image Features: Filled area, Euler number
	Bounding Box Features: Extent, aspect ratio
	Boundary Features: Perimeter, radii, perimeter Fourier energies, perimeter curvature, bending energy, perimeter fractal dimension
	Other Shape Features: Equivalent diameter, sphericity, compactness, inertia shape
	Center of Mass
	Reflection Symmetry
Radiometric and Densitometric	Image Bands, Intensity
	Optical density, integrated optical density, and mean optical
	Hue
Texture	Co-occurrence Matrix Features: Inertia, energy, entropy, homogeneity, maximum probability, cluster shade, cluster
	Fractal Dimension
	Run-length Features: Short runs emphasis, long runs emphasis, gray-level non-uniformity, run-length non-uniformity, runs percentage, low gray-level runs emphasis, high gray-level runs
	Wavelet Features: Energies of detail and low resolution images
	Entropy
Chromatin-Specific	Area, integrated optical density, mean optical density, number of regions, compactness, distance, center of mass

5.1 Object Level Features:

Fundamentally, object-level analysis depends greatly on some underlying segmentation mechanism. It is the segmentation methodology that determines what constitutes an object. Commonly, an object is defined as a connected group of pixels satisfying some similarity criterion. The main focus is often on the segmentation of nuclei; there exists little work that explicitly uses features of cytoplasm and stroma, although some researchers have hinted at the need for such features [11,12]. Preliminary work [13] has demonstrated the feasibility of other histologic features for image classification in H&E stained breast cancer. Madabhushi et al. [8] used cytoplasmic and stromal features to automatically segment glands in prostate histopathology. Moreover, it appears that histologic objects may not need to be perfectly segmented to be properly classified when a list of comprehensive features is used in a feature selection framework [13][84]. Classification performance in distinguishing between different grades of prostate cancer was found to be comparable using manual and automated gland and nuclear segmentation [8]. These results suggest that perfect segmentation is not a prerequisite for good classification.

Object-level features can be categorized as belonging to one of four categories: size and shape, radiometric and densitometric, texture, and chromatin-specific. While the radiometric and densitometric, texture, and chromatin-specific

features could be considered low-level features that can be extracted from local neighborhoods, the size and shape metrics are true object-level metrics. A summary of object-level features is listed in Table 5.1; definitions for all listed features can be found in reference [13]. These features were compiled from a comprehensive literature search on cytopathology and histopathology image analysis. In addition, various statistics measures for any of the vector quantities are also commonly calculated. Thus, the mean, median, minimum, maximum, standard deviation, skewness, and kurtosis can be calculated for all vector features. For an RGB image, all relevant features are extracted for each individual color channel; hence the total number of object-level features can easily exceed 1000 for the list of features in Table 1. It should be noted that these features are most commonly extracted from high-resolution imagery (see next section), but are relevant for any resolution.

An approach that semantically describes histopathology images using model based intermediate representation (MBIR) and incorporates low-level color texture analysis was presented in [14]. In this approach, basic cytological components in the image are first identified using an unsupervised clustering in the La*b* color space. The connected components of nuclei and cytoplasm regions were modeled using ellipses. An extensive set of features can be constructed from this intermediate representation to characterize the tissue morphology as well as tissue topology. Using this representation, the relative amount and spatial distribution of these cytological components can be measured. In the application of follicular lymphoma grading, where the spatial distribution of these regions varies considerably between different histological grades, MBIR provides a convenient way to quantify the corresponding observations. Additionally, low-level color texture features are extracted using the co-occurrence statistics of the color information. Due to the staining of the tissue samples, the resulting digitized images have considerably limited dynamic ranges in the color spectrum. Taking this fact into account, a non-linear color quantization using self-organizing maps (SOM) is used to adaptively model the color content of microscopic tissue images. The quantized image is used to construct the co-occurrence matrix from which low-level color texture features are extracted. By combining the statistical features constructed from the MBIR with the low-level color texture features, the classification performance of the system can be improved significantly.

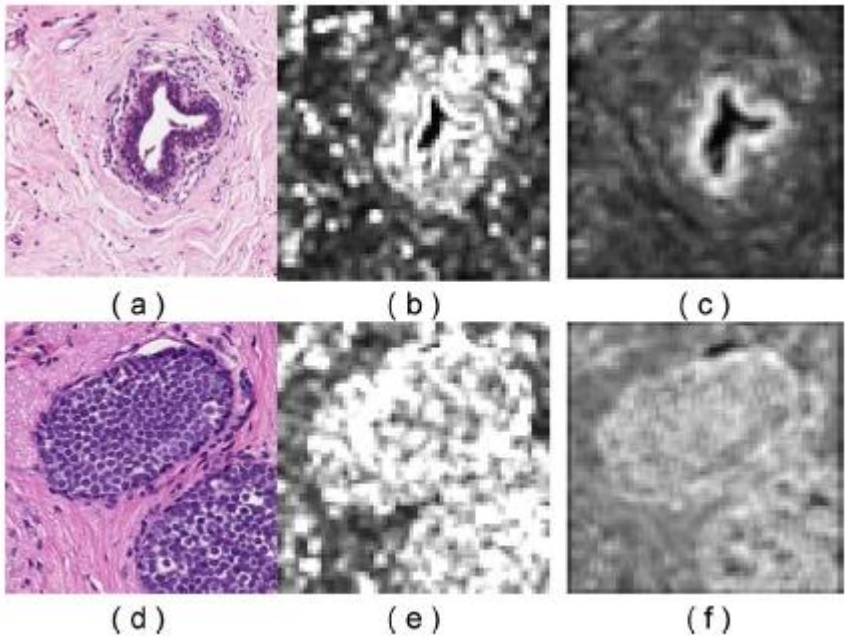

Fig. 1: Supervised extraction of histological features to describe tissue appearance of (a) benign epithelium, and (b) DCIS. Feature images for the two tissue classes (benign epithelium, DCIS) corresponding to Gabor wavelet features (b), (e) and Haralick second order features (c), (f) are shown.

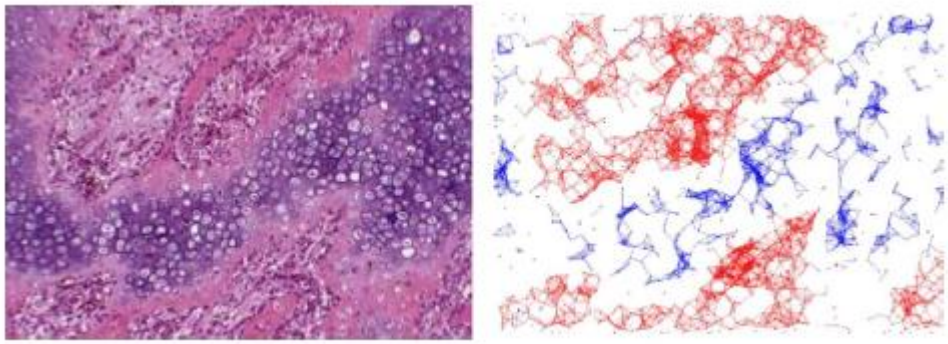

Fig. 2: Bone fracture and its corresponding ECM-aware cell-graph representation. Note the presence of a link between a pair of nodes in an ECM-aware cell-graph indicates not only topological closeness but also it implies the similarity in the surrounding ECM [19].

Figure 1 shows some of the textural image features for discriminating between benign breast epithelial tissue [9] (DCIS, Figure 1(a)) and DCIS (Figure 1(d)). Figures 1(b, e) show the corresponding Gabor filter responses while Figures 1(c, f) show the corresponding Haralick feature images.

5. 2: Spatially Related Features

Graphs are efficient data structures to represent spatial data and an effective way to represent *structural information* by defining a large set of topological features. Formally, a simple graph $G=(V,E)$ is an undirected and un-weighted

graph without self-loops, with V and E being the node and edge set of graph G, respectively.

Application of graph theory to other problem domains is impressive. Real-world graphs of varying types and scales have been extensively investigated in technological [15], social [16] and biological systems [17]. In spite of their different domains, such self-organizing structures unexpectedly exhibit common classes of descriptive spatial (topological) features. These features are quantified by definition of computable metrics.

The use of spatial-relation features for quantifying cellular arrangement was proposed in the early 1990's [18], but didn't find application to clinical imagery until recently. Graphs have now been constructed for modeling different tissue states and to distinguish one state from another by computing metrics on these graphs and classifying their values. Overall, however, the use of spatial arrangement of histological entities (generally at low resolutions) is relatively new, especially in comparison to the wealth of research on nuclear features (at higher resolutions) that has occurred during the same timeframe. A compilation of all the spatial-relation features published in the literature is summarized in Table 2. Definitions for all graph structures and features can be found in reference [13]. The total number of spatial-relation features extracted is approximately 150 for all graph structures.

Graph theoretical metrics that can be defined and computed on a cell-graph induce a rich set of descriptive features that can be used for tissue classification. These features provide structural information to describe the tissue organization such as: (i) the distribution of local information around a single cell cluster (e.g., degree, clustering coefficient, etc), (ii) the distribution of global information around a single cell cluster (e.g., eccentricity, closeness, between-ness, etc.), (iii) the global connectivity information of a graph (e.g., ratio of the giant connected component over the graph size, percentage of the isolated and end data points in the graph, etc), (iv) the properties extracted from the spectral graph theory (e.g., spectral radius, eigen exponent, number of connected components, sum of the eigenvalues in the spectrum, etc). Refer to Table 2 for a list of commonly extracted graph features.

TABLE II
SUMMARY OF SPATIAL-ARRANGEMENT FEATURES USED IN HISTOPATHOLOGY IMAGE ANALYSIS.

Graph Structure	Features
Voronoi Tesselation	Number of nodes, number of edges, cyclomatic number, number of triangles, number of k-walks, spectral radius, eigenexponent, Randic index, area, roundness factor, area disorder, roundness factor homogeneity
Delaunay Triangulation	Number of nodes, edge length, degree, number of edges, cyclomatic number, number of triangles, number of k-walks, spectral radius, eigenexponent, Wiener index, eccentricity, Randic index, fractal dimension
Minimum Spanning Tree	Number of nodes, edge length, degree, number of neighbors, Wiener index, eccentricity, Randic index, Balaban index, fractal dimension
O'Callaghan Neighborhood Graph	Number of nodes, number of edges, cyclomatic number, number of neighbors, number of triangles, number of k-walks, spectral radius, eigenexponent, Randic index, fractal dimension
Connected Graph	Number of nodes, edge length, number of triangles, number of k-walks, spectral radius, eigenexponent, Wiener index, eccentricity, Randic index, fractal dimension
Relative Neighbor Graph	Number of nodes, number of edges, cyclomatic number, number of neighbors, number of triangles, number of k-walks, spectral radius, eigenexponent, Randic index, fractal dimension
k-NN Graph	Number of nodes, edge length, degree, number of triangles, number of k-walks, spectral radius, eigenexponent, Wiener index, eccentricity, Randic index, fractal dimension

5.2.1 2D Cell-graph construction

In cell-graph generation as proposed in [19], there are three steps: (i) color quantization, (ii) node identification, and (iii) edge establishment. In the first step, the pixels belonging to cells from those of the others are distinguished. These steps are explained in the next sub-sections.

i. Node identification:

The class information of the pixels is translated to the node information of a cell-graph. At the end of this step, the spatial information of the cells is translated to their locations in the two-dimensional grid. After computing the probabilities, these are compared against a threshold value.

ii. Edge establishment:

This step aims to model pair-wise relationships between cells by assigning an edge between them. Cells that are in physical contact are considered to be in communication, thus edges can be established between them deterministically. For other node pairs, a probability function is used to establish edges between a pair of nodes randomly. Since structural properties of different tissues (e.g., breast, bone and

brain) are quite different from each other, edge establishment must be guided by biological hypothesis.

5.2.2. 3D Cell-graphs:

The first step in 3D cell-graph construction is to define the distance between a pair of nodes, which is simply the 3D Euclidean distance between a pair of nodes. Based on this distance definition, edges can be established between a pair of nodes. In addition to the simple spatial distance metrics, a multi-dimensional distance measure can be defined using the cell-level *attributes* that can be provided by sophisticated image analysis and segmentation. Cell-level attributes include: x, y, z *physical contact*, *volume* with respect to number of pixels, *peripheral* (i.e., surface area), *shared border* as percentage of shared voxels relative to total, and *polarity*. Then each node of the 3D cell-graph can be represented by a vector of v-dimensions, each dimension corresponding to an attribute. The *Lp* norm can be used to compute the multidimensional distance between them. Once the notion of distance is determined, edge functions of cell-graphs can be applied to construct 3D cell-graphs. The mathematical properties of cell-graphs in 3D can be calculated as the feature set. Although most of the features defined on 2D cell-graphs can be extended to the 3D case, their calculation is not trivial.

5.2.3 Application of Graph based modeling for different histopathology related applications

***A. Graph based Modeling of Extra Cellular Matrix*:**

The Extra Cellular Matrix (ECM) is composed of a complex network of proteins and oligosaccharides that play important roles in cellular activities such as division, motility, adhesion, and differentiation. Recently, a new technique was introduced for constructing ECM-aware cell-graphs that incorporates the ECM information surrounding the cells [20]. ECM-aware cell-graphs aim to preserve the interaction between cells and their surrounding ECM while modeling and classifying the tissues. The ECM-aware cell-graphs successfully distinguish between different types of cells that co-exist in the same tissue sample. For example, in bone tissue samples there are usually several cell types, including blood cells, normal cells, and sometimes fracture cells (e.g., chondrocytes and osteoblasts) and cancerous cells. Since these cells are functionally different from each other, the hypothesis is that they would exhibit different spatial organization and structural relationships in the same tissue. This hypothesis has been validated by showing that ECM-aware cell-graphs yield better classification results for different states of bone tissues than the current state of art. In the construction a color value is assigned to each cell (i.e., vertex) based on the RGB values of its surrounding ECM. This is done by examining the k neighboring pixels in each direction, and computing a dominant color for the ECM surrounding each cell using the RGB values of nearly $4k^2$ neighboring pixels.

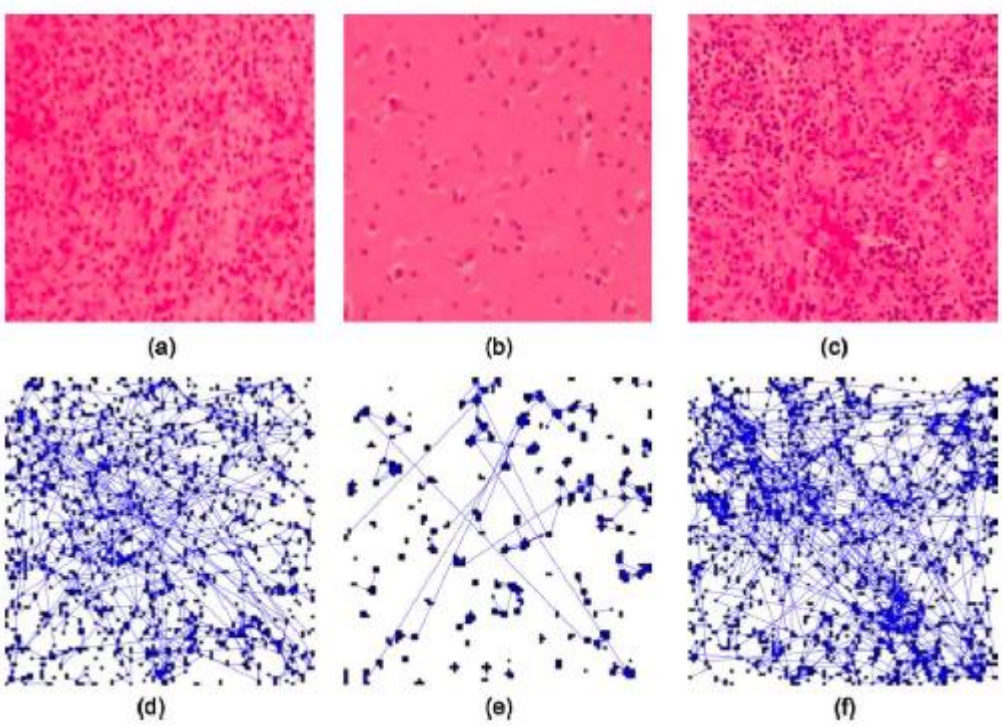

Fig. 3: Illustrating the differences between cell-graphs for cancerous, healthy, and inflamed brain tissues. Panels (a)-(c) show brain tissue samples that are (a) cancerous (gliomas), (b) healthy, and (c) inflamed, but noncancerous. Panels (d)-(f) show the cell-graphs corresponding to each tissue image. While the number of cancerous and inflamed tissue samples appear to have similar numbers and distributions of cells, the structure of their resulting cell-graphs shown in (d) and (f) are dramatically different. (Figure is taken from [20]).

B. Application to Discriminating Different States of Brain Tissue

Figure 3 shows the cell-graphs of brain tissues exhibiting distinctive graph properties that enable discrimination between the different states of brain tissue.

C. Application to Studying Temporal Activity of Adult Human Mesenchymal Stems Cells in a 3D Collagen Matrix

Figure 4 shows relationships between adult human mesenchymal stem cells in a 3D collagen protein matrix over time in culture [20]. The graphs are generated from 3D sections of tissue (900X900X80 μm) imaged using confocal microscopy. The nuclei of stem cells in the constructs were stained and imaged at the time points indicated (0 – 24 hours).

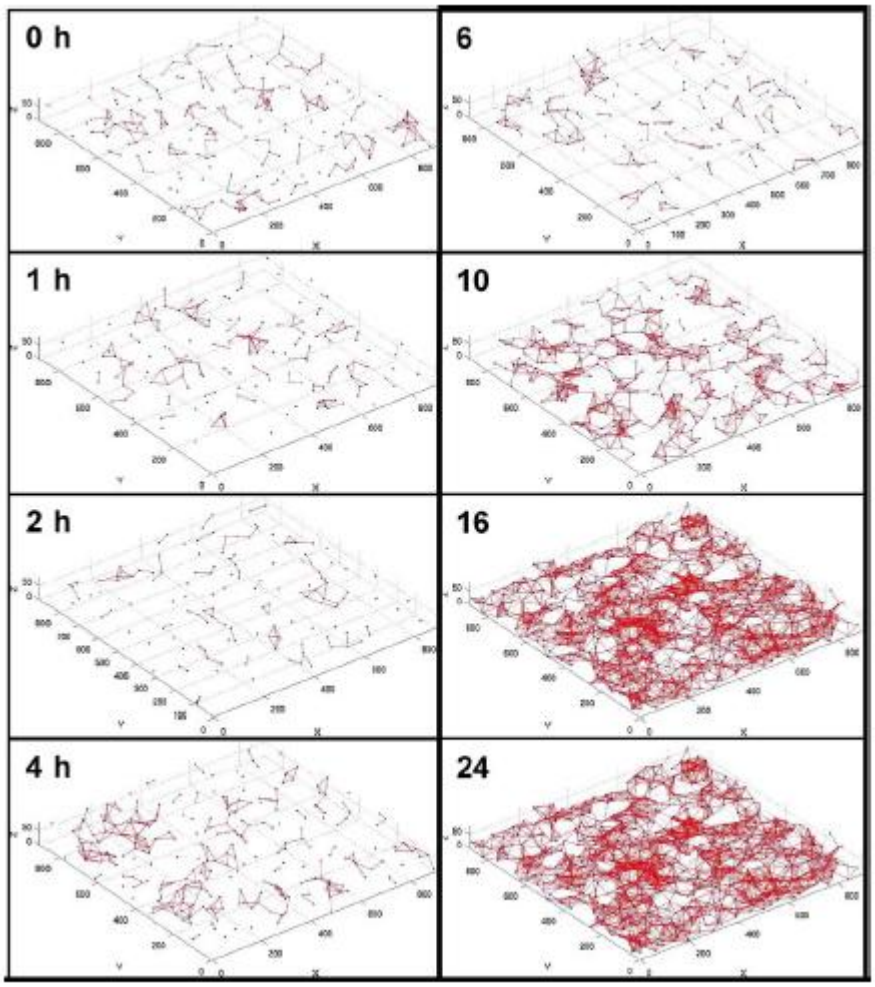

Fig. 4: Cell graphs produced from human MSC embedded in 3-D collagen matrices. Graphs shownuclei and development of edges (relationships) between them over time [19]. There is a phase transition sometime between hour 10 and hour 16 and the graph becomes connected.

D. Application of Graph Theory to Modeling Cancer Grade

In [22], the Voronoi diagram is constructed from a set of seed-like points that denote the centers of each structure of interest (nuclei). From the Voronoi diagram, two more graphs of interest can be constructed: the Delaunay triangulation, which is created by connecting points that share an edge in the Voronoi diagram, and the minimum spanning tree, which is the series of lines that spans the set of points such that the Euclidean sum of the lengths of the lines is smaller than any other spanning tree. From each of these three graphs, a series of features are calculated that captures the size, shape, and arrangement of the structures of the nuclei. The graph based representations of a Gleason grade 4 prostate histopathology image are shown in Figure 5.

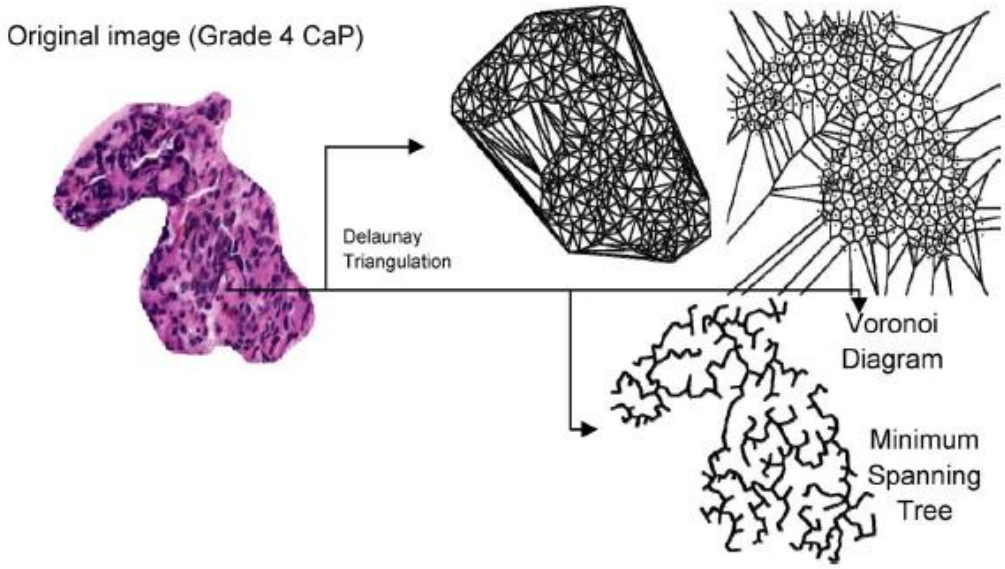

Fig. 5: (a) A digitized histopathology image of Grade 4 CaP and different graph-based representations of tissue architecture via Delaunay Triangulation, Voronoi Diagram, and Minimum Spanning tree.

5.3. Multi-scale feature extraction

Owing to the density of the data and the fact that pathologists tend to employ a multi-resolution approach to analyzing pathology data, feature values are related to the viewing scale or resolution. For instance at low or coarse scales color or texture cues are commonly used and at medium scales architectural arrangement of individual histological structures (glands and nuclei) start to become resolvable. It is only at higher resolutions that morphology of specific histological structures can be discerned.

In [23, 24], a multi-resolution approach has been used for the classification of high-resolution whole-slide histopathology images. The proposed multi-resolution approach mimics the evaluation of a pathologist such that image analysis starts from the lowest resolution, which corresponds to the lower magnification levels in a microscope and uses the higher resolution representations for the regions requiring more detailed information for a classification decision. To achieve this, images were decomposed into multi-resolution representations using the Gaussian pyramid approach [25]. This is followed by color space conversion and feature construction followed by feature extraction and feature selection at each resolution level. Once the classifier is confident enough at a particular resolution level, the system assigns a classification label (e.g., stroma-rich, stroma-poor or undifferentiated, poorly differentiating, differentiating) to the image tile. The resulting classification map from all image tiles forms the final classification map. The classification of a whole-slide image is achieved by dividing into smaller image tiles and processing each image tile independently in parallel on a cluster of computer nodes.

As an example, refer to Figure 6, showing a hierarchical cascaded scheme for detecting suspicious areas on digitized prostate histopathology slides as presented in [26].

Fig. 6: Digitized histological image at successively higher scales (magnifications) yields incrementally more discriminatory information in order to detect suspicious regions.

Figure 7 shows the results of a hierarchical classifier for detection of prostate cancer from digitized histopathology. Figure 7(a) shows the original images with tumor outlined in black. The next 3 columns show the classifier results at increasing analysis scales. Pixels classified as "non-tumor" at a lower magnification (scale) are discarded at the subsequent higher scale, reducing the number of pixels needed for analysis at higher scales. Additionally, the presence of more discriminating information at higher scales allows the classifier to better distinguish between tumor and non-tumor pixels.

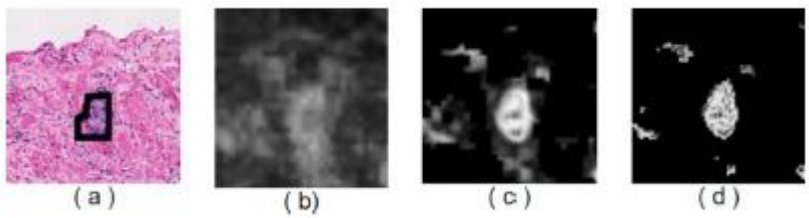

Fig. 7: Results from the hierarchical machine learning classifier. (a) Original image with the tumor region (ground truth) in black contour, (b) results at scale 1, (c) results at scale 2, and (d) results at scale 3. Note that only areas determined as suspicious at lower scales are considered for further analysis at higher scales.

At lower resolutions of histological imagery, textural analysis is commonly used to capture tissue architecture, i.e. the overall pattern of glands, stroma and organ organization. For each digitized histological image several hundred corresponding feature scenes can be generated. Texture feature values are assigned to every pixel in the corresponding image. 3D statistical, gradient, and Gabor filters can be extracted in order to analyze the scale, orientation, and anisotropic information of the region of interest. Filter operators are applied in order to extract features within local neighborhoods centered at every spatial location. At medium resolution, architectural arrangement of nuclei within each cancer grade can be described via several graph-based algorithms. At higher resolutions, nuclei and the margin and boundary appearance of ductal and glandular structures have proved to be of discriminatory importance. Many of these features are summarized in Tables 1 and 2.

5.4. Feature Selection, Dimensionality Reduction, and Manifold Learning

A. Feature Selection

While humans have innate abilities to process and understand imagery, they do not tend to excel at explaining how they reach their decisions. As such, large feature sets are generated in the hopes that some subset of features incorporates the information used by the human expert for analysis. Therefore, many of the generated features could be redundant or irrelevant. Actually, a large set of features may possibly be detrimental to the classification performance, a phenomenon known as "the curse of dimensionality." Feature selection is a means to select the relevant and important features from a large set of features. This is an increasingly important area of research now that automated quantitative image analysis techniques are becoming more mainstream.

Feature selection in histopathological image analysis provides several benefits in addition to improving accuracy. Since images tend to be relatively large, a smaller subset of features needs to be calculated, reducing the computational complexity of classification algorithms. In some applications, it may be preferable to sacrifice the overall performance slightly if this sacrifice greatly reduces the number of selected features. A smaller number of features would also make it easier to explain the underlying model and improve the chances of generalization of the developed system. Additionally, in a multi-resolution framework, a set of features proven useful at a given resolution may not be relevant at another resolution, even within the same image. A feature selection algorithm helps determine which features should be used at a given resolution.

An optimal feature selection method would require an exhaustive search, which is not practical for a large set of features generated from a large dataset. Therefore, several heuristic algorithms have been developed, which use classification accuracy as the optimality criterion. Well-known feature selection methods include the sequential search methods, namely sequential forward selection (SFS) and sequential backward selection (SBS) [27]. SFS works by sequentially adding the feature that most improves the classification performance; similarly, SBS begins with the entire feature set and sequentially removes the feature that most improves the classification performance. Both SFS and SBS suffer from the "nesting effect" whereby features that are selected (SFS) or discarded (SBS) cannot be revisited in a later step and are thus suboptimal [27]. The use of floating search methods, sequential floating forward search (SFFS) and sequential floating backward search (SFBS), in which previously selected or discarded features can be re-evaluated at later steps avoids the nesting problem [27]. While these methods still cannot guarantee optimality of the selected feature subset, they have been shown to perform very well compared to other feature selection methods [28] and are, furthermore, much more computationally efficient [27]. SFFS is one of the most commonly encountered feature selection methods in pathology image analysis literature.

More recent feature selection research has focused on such methods as genetic algorithms, simulated annealing, boosting [29] and grafting [30]. A taxonomy of feature selection algorithms is presented in [28]. Genetic algorithms and simulated annealing are applications of traditional optimization techniques to feature selection. Boosting, which will be explained in Section 6.c, basically acts as a greedy feature selection process. Grafting (from "gradient feature testing") [30] is based on an elegant formulation of the feature selection problem,

whereby the classification of the underlying data and the feature selection process are not separated. Within the grafting framework, a loss function is used that shows preference for classifiers that separate the data with larger margins. Grafting also provides an efficient framework for selection of relevant features. Feature selection based on a measure of discriminatory power was proposed in [31], whereby the authors compute the discriminatory power of each of the wavelet packet sub-bands (features) using a dissimilarity measure between approximated probability density functions for different classes. Derived features are then sorted according to the discriminatory power values associated with the corresponding features.

B. Dimensionality Reduction

While feature selection aims to select features (and reduce the feature dimensionality) that best optimize some criterion related to the class labels of the data (e.g., classification performance), dimensionality reduction techniques aim to reduce dimensionality based on some other criterion. Three well-known and commonly used methods of linear dimensionality reduction are Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA).

Principal Component Analysis (PCA) [32] looks to find a new orthogonal coordinate system whereby the maximum variance of the data is incorporated in the first few dimensions. Projection of the data onto the individual coordinates encompasses varying degrees of variance; the first coordinate encompasses the largest variance in the data, the second coordinate the next largest variance, and so forth.

On the other hand, the LDA is a supervised method; it thus requires class labels for each data sample, mapping the data onto a lower dimensional subspace that best discriminates data. The goal is to find the mapping, where the sum of distances between samples in different classes is maximized; while the sum of distances between samples in same classes is minimized. LDA can also be formulated in terms of eigenanalysis. A comprehensive discussion of PCA and LDA can be found in [33].

Independent Component Analysis [34], looks to find some mixing matrix such that a mixture of the observations (features) are statistically independent. This provides a stronger constraint on the resulting components than PCA, which only requires that the components be uncorrelated. This is why it is particularly well suited for decorrelating independent components from hyperspectral data. Rajpoot & Rajpoot [7] have shown ICA to perform well for extracting three independent components corresponding to three tissue types for segmentation of hyperspectral images of colon histology samples. ICA, however, provides no ranking of the resulting independent components, as does PCA. There are a variety of methods for calculating the independent components (refer to [34]), which are generally very computationally intensive. ICA is a higher order method

that seeks linear projections, not necessarily orthogonal to each other, as in the case of PCA.

C. Manifold Learning

Recently, non-linear dimensionality reduction methods have become popular in learning applications. These methods overcome a major limitation of summarized linear dimensionality reduction methods, which assume that geometrical structure of the high-dimensional feature space is linearized. In reality, high-dimensional feature spaces comprise of highly nonlinear structures and locality preserving dimensionality reduction methods are highly sought after. Manifold learning is a method of reducing a data set from M to N dimensions, where $N < M$ while preserving inter- and intra-class relationships between the data. This is done by projecting the data into a low-dimensional feature space in such a way to preserve high dimensional object adjacency. Many manifold learning algorithms have been constructed over the years to deal with different types of data.

Graph Embedding constructs a confusion matrix Y describing the similarity between any two images C_p and C_q with feature vectors $\boldsymbol{F}_p$ and $\boldsymbol{F}_q$, respectively, where $p, q \in \{1, 2, \ldots., k\}$ and k is the total number of images in the data set

$$Y(p, q) = e^{-||\mathrm{F}p - \mathrm{F}q||} \in \mathbb{R}^{k \times k} \tag{5. 1}$$

The embedding vector X is obtained from the maximization of the function:

$$E_Y(X) = \frac{2\eta \mathrm{X}^{\mathrm{T}}(\mathrm{D} - \mathrm{Y})\mathrm{X}}{\mathrm{X}^{\mathrm{T}}\mathrm{D}\mathrm{X}}, \tag{5. 2}$$

where D is the so-called degree matrix, with non-zero values being along the diagonal $D(p, p) = \sum_q Y(p, q)$ and $\eta = k - 1$. The k dimensional embedding space is defined by the eigenvectors corresponding to the smallest N eigenvalues of $(D - Y)X = \lambda DX$. The value of N is generally optimized by obtaining classification accuracies for $N \in \{1, 2, \cdots, 10\}$ and selecting the N that provided the highest accuracy for each classification task. For image C, the feature vector $\boldsymbol{F}$ given as input to the Graph Embedding algorithm produces an N-dimensional eigenvector $X(C) = [e_j(C) / j \in \{1, 2, \ldots.., N\}]$, where $e_j(C)$ is the principal eigenvalue associated with C.

In [22], a Graph Embedding algorithm employing the normalized cuts algorithm was used to reconstruct the underlying manifold on which different breast cancer grades were distributed. Figure 8 shows the embedding of different grades of breast cancer histopathology (low, intermediate, high) on the reconstructed manifold; low grades (yellow triangles), intermediate grades (green squares and blue circles), and high grades (red triangles). The manifold

captures the biological transformation of the disease in its transition from low to high-grade cancer.

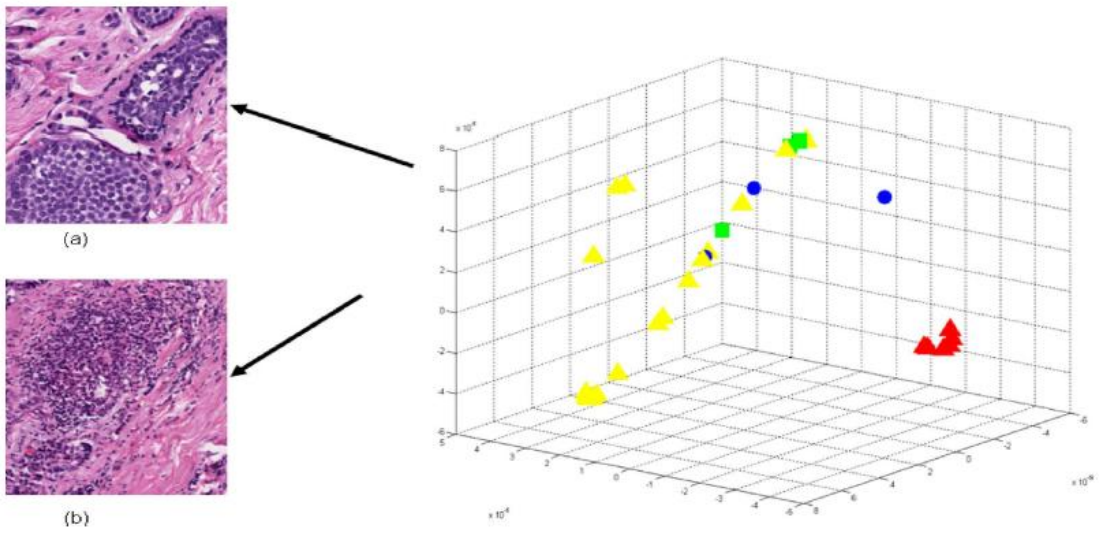

Fig. 8: Low-dimensional embedding reveals innate structure in textural features of invasive breast cancers, with clear separation of high grade tumors from low and intermediate grade tumors as assessed by Nottingham score. Combined Nottingham score 5 (yellow triangle), 6 (green squares), 7 (blue circles), and 8 (red triangles). The score of 8 corresponds to high grade tumors. (a) Low grade (Yellow triangles). (b) High grade (Red triangles).

Manifold learning has also been shown to be useful for shape-based classification of prostate nuclei [35]. Rajpoot *et al.* [35] employ Diffusion Maps [36] in order to reduce the dimensionality of shape descriptors down to two dimensions and a fast classification algorithm is derived based on a simple thresholding of the diffusion coordinates.

[1] A. J. Mendez, P. G. Tahoces, M. J. Lado, M. Souto, and J. J. Vidal, "Computer-aided diagnosis: automatic detection of malignant masses in digitized mammograms," Med Phys, vol. 25, pp. 957-64, Jun 1998.

[2] J. Tang, R. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, "Computer-Aided Detection and Diagnosis of Breast Cancer with Mammography: Recent Advances," IEEE Trans Inf Technol Biomed, Jan 20 2009.

[3] R. Rubin, D. Strayer, E. Rubin, and J. McDonald, Rubin's pathology: clinicopathologic foundations of medicine: Lippincott Williams & Wilkins, 2007.

[4] K. L. Weind, C. F. Maier, B. K. Rutt, and M. Moussa, "Invasive carcinomas and fibroadenomas of the breast: comparison of microvessel distributions--implications for imaging modalities," Radiology, vol. 208, pp. 477-83, Aug 1998.

[5] P. H. Bartels, D. Thompson, M. Bibbo, and J. E. Weber, "Bayesian belief networks in quantitative histopathology," Anal Quant Cytol Histol, vol. 14, pp. 459-73, Dec 1992.

[6] P. W. Hamilton, N. Anderson, P. H. Bartels, and D. Thompson, "Expert system support using Bayesian belief networks in the diagnosis of fine needle aspiration biopsy specimens of the breast," J Clin Pathol, vol. 47, pp. 329-36, Apr 1994.

[7] K. Rajpoot and N. Rajpoot, "Hyperspectral Colon Tissue Cell Classification," in Proceedings, 2004.

[8] S. Naik, Doyle,S., Madabhushi A, Tomaszeweski,J., Feldman,M., "Automated Gland Segmentation and Gleason Grading of Prostate Histology by Integrating Low-, High-level and Domain Specific Information," in Workshop on Microscopic Image Analysis with Applications in Biology Piscataway, NJ, 2007.

[9] S. Naik, Doyle, S, Agner, S, Madabhushi, A, Tomaszeweski, J, Feldman, M, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopatholog," in ISBI Special Workshop on Computational Histopathology (CHIP) Paris, France: IEEE, 2008, pp. 284-287.

[10] K. Rodenacker and E. Bengtsson, "A feature set for cytometry on digitized microscopic images," Analytical Cellular Pathology, vol. 25, pp. 1-36, 2003.

[11] A. J. Sims, M. K. Bennett, and A. Murray, "Image analysis can be used to detect spatial changes in the histopathology of pancreatic tumours," Physics in Medicine and Biology, vol. 48, pp. N183-N191, 2003.

[12] J. Gil, H. Wu, and B. Y. Wang, "Image Analysis and Morphometry in the Diagnosis of Breast Cancer," Microscopy Research and Technique, vol. 59, pp. 109-118, 2002.

[13] L. E. Boucheron, "Object- and Spatial-Level Quantitative Analysis of Multispectral Histopathology Images for Detection and Characterization of Cancer," in PhD Thesis, 2008.

[14] O. Sertel, J. Kong, U. Catalyurek, G. Lozanski, J. Saltz, and M. Gurcan, "Histopathological Image Analysis using Model-based Intermediate Representations and Color Texture: Follicular Lymphoma Grading," Journal of Signal Processing Systems (in print), 2009.
[15] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," 1999, pp. 251-262.
[16] D. Watts and S. Strogatz, "Collective dynamics ofsmall-world'networks," Nature, vol. 393, pp. 440-442, 1998.
[17] S. Wuchty, E. Ravasz, and A. Barabasi, "The architecture of biological networks," Complex Systems in Biomedicine, 2003.
[18] R. Albert, T. Schindewolf, I. Baumann, and H. Harms, "Three-Dimensional Image Processing for Morphometric Analysis of Epithelium Sections," Cytometry, vol. 13, pp. 759-765, 1992.
[19] C. Bilgin, C. Demir, C. Nagi, and B. Yener, "Cell-Graph Mining for Breast Tissue Modeling and Classification," 2007, pp. 5311-5314.
[20] C. Bilgin, P. Bullough, G. Plopper, and B. Yener, "ECM Aware Cell-Graph Mining for Bone Tissue Modeling and Analysis," RPI Computer Science Technical Report 08-07, 2008.
[21] C. Gunduz, B. Yener, and S. Gultekin, "The cell graphs of cancer," Bioinformatics, vol. 20, 2004.
[22] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," in ISBI, 2007, pp. 1284-1287.
[23] J. Kong, O. Sertel, H. Shimada, K. Boyer, J. Saltz, and M. Gurcan, "Computer-aided evaluation of neuroblatoma on whole-slide histology images: Classifying grade of neuroblastic differentiation," Pattern Recognition, vol. 42, pp. 1080-1092, 2009.
[24] O. Sertel, J. Kong, H. Shimada, U. Catalyurek, J. Saltz, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," Pattern Recognition, vol. 42, pp. 1093-1103, 2009.
[25] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," IEEE Transactions on Communications, vol. 31, pp. 532-540, 1983.
[26] S. Doyle, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "A boosting cascade for automated detection of prostate cancer from digitized histology," Lecture Notes in Computer Science, vol. 4191, p. 504, 2006.
[27] P. Pudil, J. Novovi\vcov\'a, and J. Kittler, "Floating search methods in feature selection," Pattern Recognition Letters, vol. 15, pp. 1119-1125, 1994.
[28] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 153-158, 1997.
[29] Y. Freund and R. E. Shapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, pp. 119-139, 1997.
[30] S. Perkins, K. Lacker, and J. Theiler, "Grafting: Fast, Incremental Feature Selection by Gradient Descent in Function Space," Journal of Machine Learning Research, vol. 3, pp. 1333-1356, 2003.
[31] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. Gurcan, "Adaptive Discriminant Wavelet Packet Transform and Local Binary Patterns for Meningioma Subtype Classification," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008, 2008, pp. 196-204.
[32] I. Jolliffe, Principal component analysis, 2002.
[33] A. Martinez and A. Kak, "Pca versus lda," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 228-233, 2001.
[34] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," Neural networks, vol. 13, pp. 411-430, 2000.
[35] N. Rajpoot, M. Arif, and A. Bhalerao, "Unsupervised Learning of Shape Manifolds," in Proceedings, 2007.
[36] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," Proceedings of the National Academy of Sciences, vol. 102, pp. 7426-7431, 2005.