

## Review

# Integrated diagnostics: a conceptual framework with examples

Anant Madabhushi<sup>1</sup>, Scott Doyle<sup>1</sup>, George Lee<sup>1</sup>, Ajay Basavanahally<sup>1</sup>, James Monaco<sup>1</sup>, Steve Masters<sup>2</sup>, John Tomaszewski<sup>2</sup> and Michael Feldman<sup>2,\*</sup>

<sup>1</sup> Laboratory for Computational Imaging and Bioinformatics, Department of Biomedical Engineering, Rutgers University, NJ, USA

<sup>2</sup> Department of Pathology and Laboratory Medicine, University of Pennsylvania Health System, Philadelphia, PA, USA

## Abstract

With the advent of digital pathology, imaging scientists have begun to develop computerized image analysis algorithms for making diagnostic (disease presence), prognostic (outcome prediction), and theragnostic (choice of therapy) predictions from high resolution images of digitized histopathology. One of the caveats to developing image analysis algorithms for digitized histopathology is the ability to deal with highly dense, information rich datasets; datasets that would overwhelm most computer vision and image processing algorithms. Over the last decade, manifold learning and non-linear dimensionality reduction schemes have emerged as popular and powerful machine learning tools for pattern recognition problems. However, these techniques have thus far been applied primarily to classification and analysis of computer vision problems (e.g., face detection). In this paper, we discuss recent work by a few groups in the application of manifold learning methods to problems in computer aided diagnosis, prognosis, and theragnosis of digitized histopathology. In addition, we discuss some exciting recent developments in the application of these methods for multi-modal data fusion and classification; specifically the building of meta-classifiers by fusion of histological image and proteomic signatures for prostate cancer outcome prediction. Clin Chem Lab Med 2010;48:989–98.

**Keywords:** diagnostic pathology; digitalized histopathology; image analysis algorithms; pattern recognition; proteomic signature; theragnostic.

\*Corresponding author: Michael Feldman, MD, PhD, Department of Pathology and Laboratory Medicine, University of Pennsylvania Health System, Founders Bldg, Rm 6.058, 3400 Spruce St Philadelphia, PA 19104, USA  
Phone: +215-662-6503, Fax: +215-349-5910,  
E-mail: feldmanm@mail.med.upenn.edu  
Previously published online May 23, 2010

## Introduction

Pathology laboratories generate vast amounts of data on a regular basis. These data take many forms and have been used for diagnostic analysis, prognostic analysis and theragnostic (response to therapies) analysis. Some of these data elements are discreet and easily integrated into computational models to classify disease into diagnostic categories, prognostic categories or theragnostic categories. Data generated in pathology laboratories exists in many forms that vary in size, dimension and scale.

Clinical laboratory data is frequently composed of discreet data elements that are often numerical in value (sodium, potassium, PSA, etc.). Other clinical laboratory data may exist as more complex data, such as a molecular diagnostic data which may reside as identifying a point mutation (discreet data element, such as the presence or absence of a point mutation) vs. more complex molecular data, such as a multiplexed in vitro diagnostic assay with 10 or more data elements for a single assay, or multicolor flow cytometric analysis of leukemia or lymphoma (routinely done with six or more simultaneous fluorochromes). Historically, these more complex laboratory data sets are reduced to single discreet values (a risk score in the case of a multiplexed assay or a series of scatterplots in the case of flow cytometry). In both of these cases, the data has been reduced as a stand-alone assay, and often involves data reduction with human intervention (flow cytometry scatterplots), a step which introduces human bias into the generation of the interpreted data. Some groups have been developing more sophisticated methods of analyzing some of these complex data sets. Finn et al. (1, 2) have developed a statistical manifold to explore multiplexed flow cytometry data which is based on non-parametric embedding of high dimensional flow data into a lower dimensional space meant to preserve object adjacencies that were present in the high dimensional space. Rogers et al. have used a different pattern discovery algorithm termed cytometric fingerprinting to analyze flow cytometry data (3). These two approaches to data analysis illustrate a new paradigm in data exploration within laboratory medicine, an approach that involves the use of computer computation to learn or discern complex patterns within high dimensional data sets. This analysis can occur with human intervention (supervised analysis), or in a completely unsupervised approach.

In anatomic pathology, diagnostic information has historically been generated in the form of textual data. As a result, anatomic pathology results have been difficult to integrate into data analysis that combines complex data types. Typi-

cally, data from anatomic pathology have only been represented as a diagnostic type (prostate adenocarcinoma, breast carcinoma, etc.) or more recently, with the increasing utilization of synoptic tumor reporting, the addition of tumor subtypes, tumor grade, tumor stage and other diagnostic or prognostic variables have been abstracted from glass slide review and placed into anatomic pathology reports. This data can be used for clinical reporting as well as used in more complex data analysis. However, all of these approaches in anatomic pathology rely on expert analysis and data extraction, which introduces operator bias into the data extraction. Numerous studies over time have shown the difficulty in using this type of data due to poor inter-observer reliability of the expert data extraction step as measured by kappa statistics (4–7).

The advent of more powerful computer and sophisticated computational approaches now allows us to rethink our approaches to data analysis in pathology and laboratory medicine. Advances in image analysis, whole slide scanners, fluorescent and multispectral scanners now allow feature extraction of pathology images so that the resulting slide is not reduced to a single or a series of discrete textual results, but instead resides in a complex series of image features which may be several thousand features in number and include both primary and secondary image features and graph based features. This review paper will discuss some of the computer vision and non-linear dimensionality reduction (NLDR) approaches which are changing how we think about and use pathology imagery. These pattern recognition and machine learning tools are ideally suited to dealing with high dimensional data, such as digitized pathology, and help provide a rich feature set of data derived from the original images. The image analysis feature set derived from these feature extraction algorithms may also be integrated with non-imaging based features including genomic, epigenomic and proteomic analyses of the same tissue used for image based feature extraction. The resulting data set from these two approaches (image and omics) form the inputs of a high dimensional data set that is characterized by multimode registered data. The data sets are registered at the spatial level since they are derived from the same areas of a pathology slide and are

multimode since the individual data elements are derived from multiple modalities. This registered multimode data set can then be used as inputs into a second set of analytic approaches called machine learning, designed to discern classes of information. There are numerous machine learning approaches, the one we will describe in this manuscript is called manifold learning. This approach is based on data fusion and representation of high dimensional data into a lower dimensional data space that preserves the relative distances of objects from each other from the high dimensional space as it is projected into the lower dimensional space in which the class differences being modeled can be visualized along a manifold surface (Figure 1).

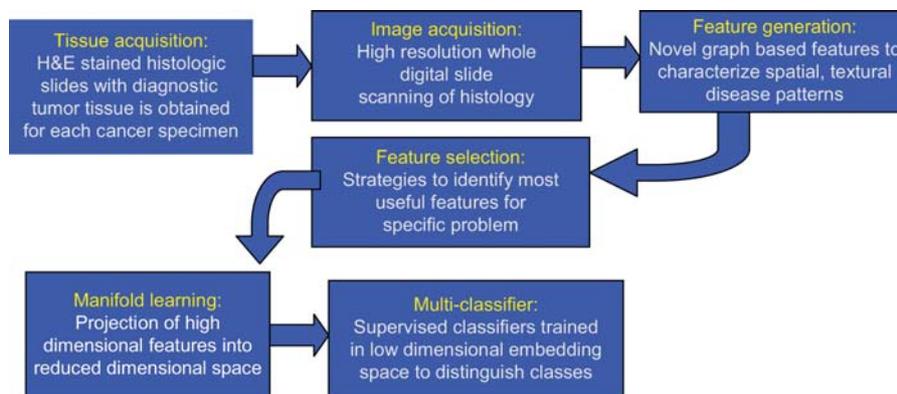
## Feature extraction

Research on useful features for cancer classification and diagnosis has often been inspired by grading features defined by clinicians as particularly important for the 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 (8), there is relatively little such work for histopathology imagery.

Human observers (pathologists) concept of the world is inherently object-based, as opposed to the largely pixel-based representation of computer vision. As such, pathology experts describe and understand images in terms of such objects. For pathologists, diagnosis criteria are inevitably described using terms, such as ‘nucleus’ and ‘cell’ and relationship of larger objects to each other and to benign adjacent tissue. It is thus important to develop computer vision methods capable of such object-level analysis (Figure 2).

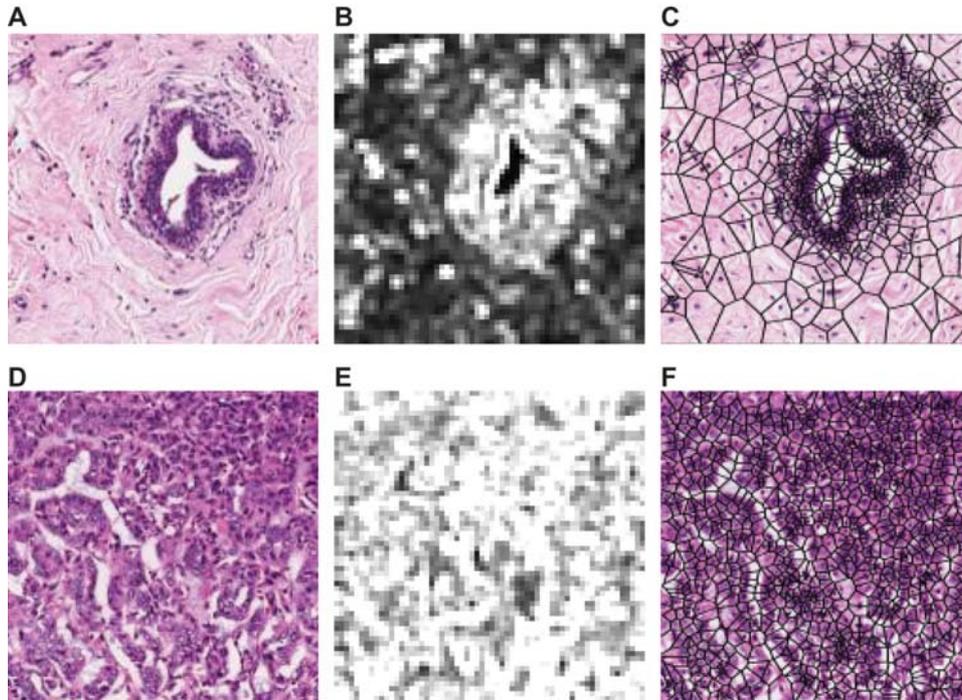
## Spatially related features

Graphs are effective techniques to represent *structural information* by defining a large set of topological features. Appli-



**Figure 1** Data acquisition paradigm.

Illustrates the method of data acquisition, feature extraction, feature selection, dimensionality reduction and classification.



**Figure 2** Supervised extraction of histological features to describe tissue appearance of (A) benign epithelium, and (B) ductal carcinoma in situ (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.

cation of graph theory to other problem domains is impressive. Real-world graphs of varying types and scales have been extensively investigated in technological, social (9) and biological systems (10). 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 (Table 1).

The use of spatial-relation features for quantifying cellular arrangement was proposed in the early 1990s (11, 12), but did not find applications in clinical imagery until recently. Graphs have now been constructed for modeling different tissue states and for distinguishing 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. 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

**Table 1** Summary of spatial-arrangement features used in histopathology image analysis.

Graph structure	Features
Voronoi tessellation	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

(e.g., spectral radius, eigenexponent, number of connected components, sum of the eigenvalues in the spectrum, etc.).

The end result of these feature extraction algorithms is a set of features that can be used for image classification. Prior to classification, a pruning of the most relevant features must be performed in a step termed feature selection.

## Feature selection

While humans have innate abilities to process and understand imagery, they tend not 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 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) (14). 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 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 (14). 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 (15–17). While these methods still cannot guarantee optimality of the select-

ed feature subset, they have been shown to perform very well compared to other feature selection methods (18), and furthermore, are much more computationally efficient. SFFS is one of the most commonly encountered feature selection methods in the pathology image analysis literature.

More recent feature selection research has focused on such methods as genetic algorithms, simulated annealing, boosting (19) and grafting (20). A taxonomy of feature selection algorithms is presented by Jain and Zonker (18). Genetic algorithms and simulated annealing are applications of traditional optimization techniques to feature selection. Boosting basically acts as a greedy feature selection process. Grafting (from “gradient feature testing”) (20) 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 by Qureshi et al. (21). These authors compute the discriminatory power of each of the wavelet packet subbands (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.

## Dimensionality reduction and manifold learning

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).

PCA (22) attempts 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. Thus, by retaining only the first few dimensions of the PCA transform, the sources of the largest amount of variation in the data are maintained. PCA is based on an eigenanalysis of the data, and thus can be computationally expensive. Ordering the eigenvectors according to their associated eigenvalues provides the principal components.

ICA (23–26), in contrast, looks to find some mixing matrix  $W$  such that the components of  $Y = WX$  are statistically independent. This provides a stronger constraint on the resulting components than PCA, which only requires that the components be uncorrelated. 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 (23–26)], which are generally very computationally intensive.

Recently, NLDR methods have become popular in learning applications. These methods overcome a major limitation of linear dimensionality reduction methods, such as PCA and LDA, which assume that geometrical structure of the high-dimensional feature space is linearized. In reality, high-dimensional feature spaces comprised of highly non-linear structures and locality preserving dimensionality reduction methods are highly sought after. Many manifold learning algorithms have been constructed over the years to deal with different types of data (27–33). Graph embedding is one such algorithm that aims to non-linearly project high dimensional data into a reduced dimensional space, while simultaneously preserving object adjacencies (34–37). The high dimensional feature space  $F$  will be significantly reduced in terms of number of dimensions to a lower dimensional eigen feature vector  $X$  (where  $|F| \gg |X|$  and  $|F|$  is the cardinality of set  $F$ ).  $X$  now represents the low dimensional embedding coordinates (or principal eigenvalues) of the original object (e.g., a histology image), in turn representing the variance in the data in  $F$ . Hence, if two objects (e.g., pathology images)  $L_1$  and  $L_2$  are embedded close to each other in the lower dimensional sub-space, then it suggests that they are similar to each other [potentially in terms of survival, or similarity in spatial arrangement of breast cancer (BC) nuclei]. *NLDR schemes differ from linear DR schemes, such as PCA in that they attempt to preserve geodesic distances between objects in the high dimensional space.*

## Classification

For histopathology imagery, unlike some other applications of image analysis, one of the primary considerations in the choice of a classifier is its ability to deal with large, highly dense datasets. Also due to multiple image scales at which relevant information may be extracted from histological imagery, use of an ensemble of classifiers as opposed to a single classifier has been proposed. Following feature extraction, selection and dimensionality reduction, different schema for classification may be applied to the histopathologic images.

### Multi-classifier ensemble schemes

Both theoretical and empirical results have established that, in terms of accuracy, ensembles of classifiers generally outperform monolithic solutions. Learning ensembles or multiple classifier systems are methods for improving classification accuracy through aggregation of several similar classifiers' predictions, and thereby reducing either the bias or variance of the individual classifiers (38–40).

**Support vector machines (SVM)** SVMs project a set of training data,  $E$ , representing two different classes into a high-dimensional space by means of a kernel function  $\mathbf{K}$ . The algorithm then generates a discriminating hyper-plane to

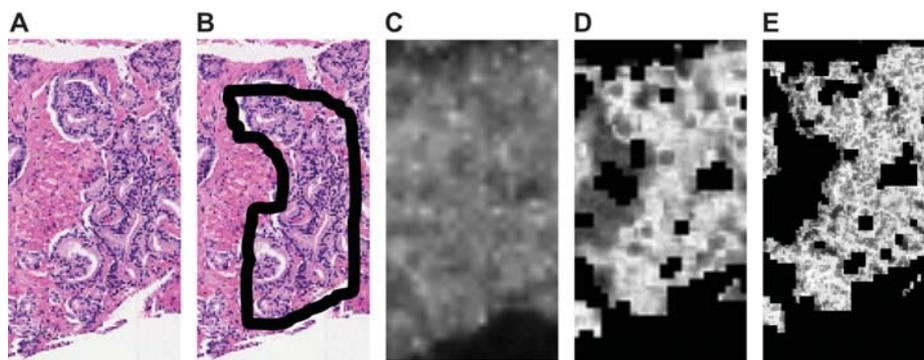
separate out the classes in such a way to maximize a cost function. Testing data is then projected into the high-dimensional space via  $\mathbf{K}$ , and the test data is classified based on where it falls with respect to the hyper-plane. The kernel function  $\mathbf{K}$  defines the method in which data is projected into the high-dimensional space. A commonly used kernel known as the radial basis function has been employed to distinguish between three different classes of prostate tissue (41). Radial basis functions with a grid search for their parameters have also been used to differentiate colon adenocarcinoma histopathology images from benign histopathology images (42), and to classify four different subtypes of meningioma from their histopathology images (43).

**Adaboost** The AdaBoost algorithm is used to combine a number of weak learners to generate a strong classifier. In the study by Doyle et al. (44), a hierarchical boosted cascade scheme for detecting suspicious areas on digitized prostate histopathology, inspired by the work of Viola and Jones (45) on face detection was presented. Efficient and accurate analysis is performed by first detecting those areas only found to be suspicious at lower scales. Analysis at subsequent higher magnifications is limited to those regions deemed to be suspicious at lower 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. The process is repeated using an increasingly larger number of image features and an increasing classification threshold at each iteration. Qualitative results with a hierarchical boosted classifier at three different image resolutions are shown in Figure 3.

## Applications

### Prostate cancer grading

The classification of histopathology imagery is often the ultimate goal in image analysis, particularly in cancer applications. Features derived from segmented nuclei and glands from histopathology are usually a pre-requisite to extracting higher level information regarding the state of the disease. For instance, the grading of prostate cancer by Jafari-Khouzani and Soltanian-Zadeh (46) yielded 97% accuracy for Hematoxylin & Eosin (H&E) stained imagery based on features derived from nuclear structures in histopathology. Weyn et al. (47) reported 87.1%–96.8% accuracy in the correct diagnosis (3 diagnoses) of Feulgen-stained lung cancer specimens, 79.5%–92.3% accuracy in typing (3 types) malignant mesothelioma, and 74.3%–82.9% accuracy in the prognosis (3 classes of survival time) of malignant mesothelioma cases. Analysis of Feulgen-stained breast tissue sections by van de Wouwer et al. found 67.1% accuracy in classifying nuclei as benign or malignant, but 100% classification on a patient level. Tabesh et al. (48) found 96.7% accuracy in discriminating between prostate tissue slides with cancer and no cancer, and 81% accuracy in the discrimination between low



**Figure 3** Hierarchical boosting to identify prostate carcinoma in whole digitized pathology slides. From left to right, (A) a digitized histopathology image, (B) cancer extent delineated in black by an expert pathologist, and cancer probability images generated by an Adaboost classifier at (C) low-, (D) intermediate, and (E) high-image resolutions.

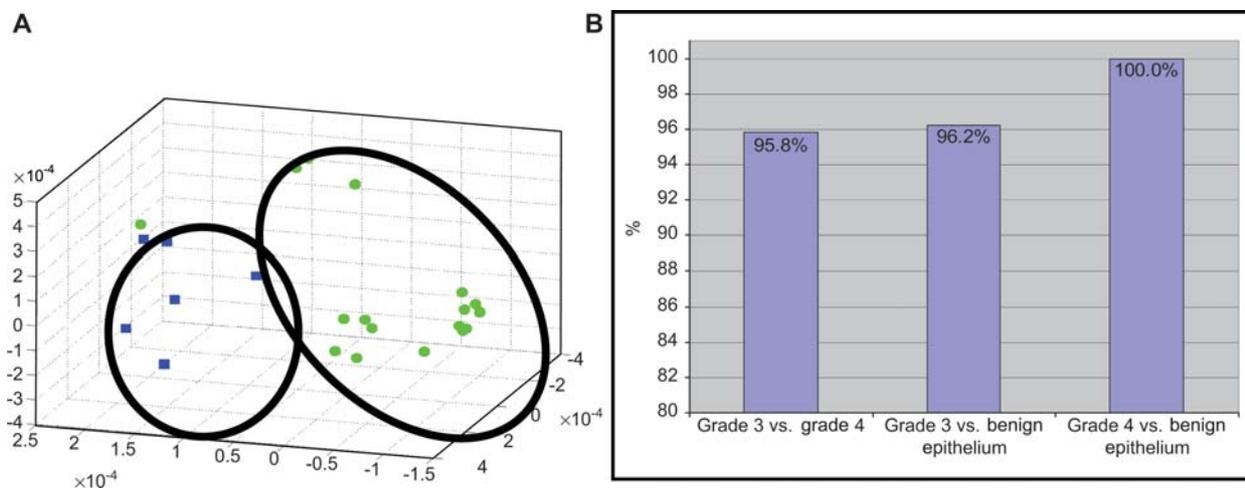
and high Gleason grades in the same imagery. Immunohistochemically stained colon mucosa allowed for an accuracy of 92% for classification of benign and malignant images by Esgiar et al. (49, 50).

Figure 4A shows the low dimensional embedding of the high dimensional attribute space via locally linear embedding of 20 images representing prostate cancer grades 3 (green circles) and 4 (blue squares). Each image is displayed as a point in 3D eigenspace. The clustering clearly shows very good discrimination between these two classes, which clinically is the most challenging problem in terms of Gleason grading. Figure 4B shows bar plots reflecting the classification accuracy obtained via a supervised classifier in distinguishing between pairs of tissue classes – grade 3/4, grade 3 vs. benign epithelium, and grade 4 vs. benign epithelium via a SVM classifier. Note that in every case the classification accuracy is over 90%.

### Quantifying extent of lymphocytic infiltration on Her2+ breast cancer imagery

Molecular changes in BC are sometimes accompanied by corresponding changes in phenotype. Therefore, the identification of these phenotypic changes in BC histopathology is of significant prognostic and theragnostic value. One such phenotype is the presence of lymphocytic infiltration (LI), a form of immune response often seen in high grade BC. The presence of LI in BC histopathology has been shown to correlate with tumor recurrence and nodal metastasis. To characterize LI computationally, we presented a computer assisted diagnosis (CAD) methodology to automatically detect and quantitatively grade the extent of LI in high-grade, HER2+ BC histopathology images.

The lymphocyte CAD algorithm comprises three main stages. First, we employ an automated lymphocyte detection



**Figure 4** Non-linear dimensionality reduction used to classify prostate adenocarcinoma into Gleason grade 3 and grade 4 patterns. (A) Low dimensional embedding of the high dimensional attribute space via locally linear embedding of 20 images representing prostate cancer grades 3 (circles) and 4 (squares). Each image is displayed as a point in 3D eigenspace. The clustering clearly shows very good discrimination between these two classes, which clinically is the most challenging problem in terms of Gleason grading. (B) Bar plots reflecting the classification accuracy obtained via a supervised classifier in distinguishing between pairs of tissue classes – grade 3/4, grade 3 vs. benign epithelium, and grade 4 vs. benign epithelium via a SVM classifier. Note that in every case the classification accuracy is over 90%.

scheme to identify lymphocyte nuclei from surrounding stroma and cancer cell nuclei. A region-growing algorithm initially segments all candidate nuclei. Since many cancer cell nuclei are also found, a Bayesian classifier is used in conjunction with a Markov random field (MRF) to refine the segmentation. The Bayesian classifier calculates features for each candidate nucleus and uses the size and luminance information to distinguish lymphocyte nuclei from cancer cell nuclei. The MRF is an iterative algorithm that models the infiltration phenomenon by incorporating the spatial proximity of lymphocytes in a histopathology image.

Using the segmented lymphocytes, a total of 50 architectural (i.e., graph-based and nuclear) features are extracted for each image. We construct three graphs (Voronoi Diagram, Delaunay Triangulation, and Minimum Spanning Tree) and calculate 25 graph-based features for each image. The remaining 25 features represent nuclear statistics, such as lymphocyte density and nearest neighbor statistics.

For a dataset containing 41 images, we distinguish images with high and low LI by reducing the architectural feature set with NLDR and applying a SVM classifier. We compare our method to two textural feature sets: the popular Varma-Zisserman (VZ) texton-based features and global texture features. While the architectural feature set achieves a classification accuracy of 92%, the texton-based and global texture feature sets perform at 60% and 50%, respectively (51).

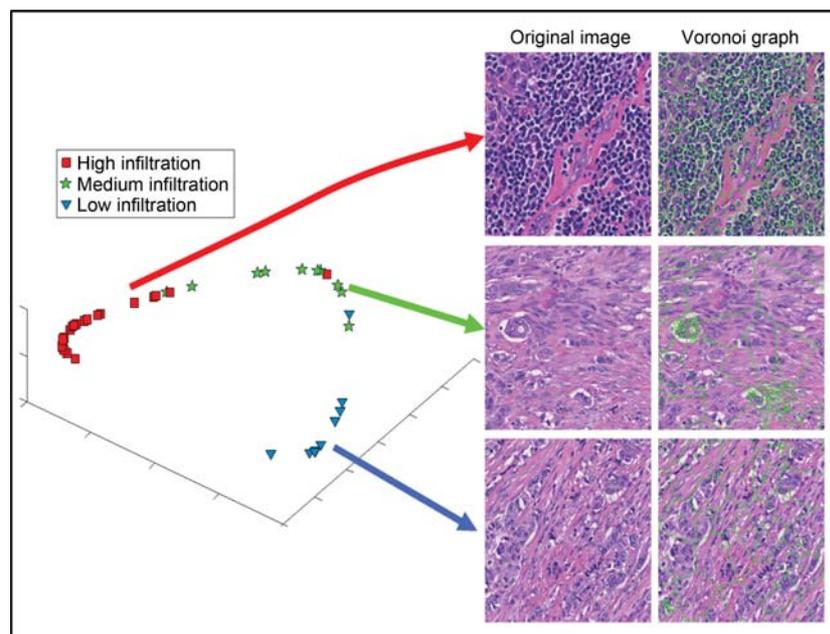
Figure 5 shows that the low dimensional representation, obtained by the application of Graph embedding on a set of HER2+ BC histopathology images from which Voronoi graph based features were derived to quantitatively characterize the extent, pattern, and density of LI (presence of lym-

phocytic nuclei), results in a smooth curvilinear manifold with a continuous transition for low to intermediate to high levels of LI (levels of LI have been clinically correlated to disease outcome; high levels of LI result in better outcome/survival) (54). *By mapping new samples onto this manifold and based on the location of the sample on the manifold, a prediction of disease outcome could be made.* The manifold in the meta-space captures the biological transformation of the disease in its transition from good to poor prognosis cancer (51).

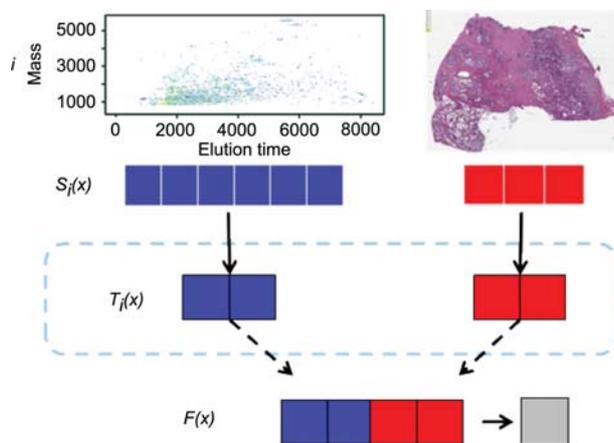
This study demonstrates the ability of a CAD algorithm to automatically detect and quantitatively characterize the nature of LI in breast histopathology. These tasks are of particular importance due to the connection between LI in BC and patient outcome. Furthermore, since the methodology proposed in this study is independent of the underlying biology, it may also be applicable to the histopathology of other diseases.

### Generalized fusion framework

*H&E stained prostate whole mount histology and corresponding proteomic spectra from mass spectrometry can be used to predict prostate cancer recurrence via a generalized fusion framework (GFF).* GFF provides seamless integration of multiple modalities (in this case histology and proteomics) while addressing issues with previous combination of decision (COD) and combination of interpretation (COI) strategies. Owing to the disparity of the dimensionalities that the two different modalities resides in, meta-space, a novel knowledge representation framework is presented which



**Figure 5** Non-linear dimensionality reduction used to classify lymphocytic infiltration in Her2 positive breast carcinoma. Visualization of HER2+ breast cancer tissue samples with low-, to medium-, to high-levels of lymphocytic infiltration (LI) (52, 53) in the meta-space. Graphs constructed on the LI allow for extraction of architectural features, resulting in a smooth manifold (obtained via CEmbed) with clear separation b/w different LI levels.



**Figure 6** A generalized fusion framework (GFF) is illustrated in which data sources (proteomics and histology) are transformed into a common meta-space prior to information fusion. The fused data can then be classified to make a prediction for cancer recurrence.

involves first performing dimensionality reduction on the individual modalities. The histological and proteomic features are transformed into a uniform representation in the meta-space. Data fusion and meta-classifier construction can now take place in this reduced dimensional meta-space (see Figure 6 below).

Data  $S_i(x)$  is transformed into meta-space  $T_i$  where fusion of disparate sources  $i$  is possible. The meta-space preserves the discriminative properties of the original sources in a lower dimensional space, representing both in terms of eigenvector features of equal dimensionality and scale. The fused data representation thus contains the discriminative power of a joint feature space, but in low dimensionality, making it more amenable to classification.

## Concluding remarks

Together this pipeline for processing data provides a framework for image data (whether it be from pathology images, gene chips, proteomics, or other sources not discussed (flow cytometry, radiology images) that allows for pattern discovery in high dimensional data sets that describe complex biologic systems. The exact algorithms for optimum feature selection, feature extraction, data reduction, and classification methodologies or pattern finding approaches will vary depending upon the systems being analyzed and the distribution of data. However, what is clear is that the ability to classify and understand biologic systems in diagnostic medicine will get ever more complex and the need to describe and model them will require the creation of complex high dimensional data sets which will be analyzed with complex algorithms. This analyzed data will then be vetted by domain experts, medical diagnosticians that perhaps do not fall into the classic categories of physicians we train today. They will wear different hats (not radiology or pathology caps) and will learn to be comfortable developing and interacting with these complex data models. The logistics of how this will

evolve, the training that will be required, the rules that will apply to training and credentialing will need to be developed as we move from classic diagnostic into a new realm of complex systems based approaches to understanding diseases, diagnosing and prognosticating disease and developing tools to discover what the best therapeutic options are for patients given their genetic makeup and how their specific disease interacts with their unique constitution.

## Acknowledgements

This work was made possible via grants from the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute Grant Nos. R01CA136535-01, R21CA127186, R03CA128081-01, and R03CA143991-01, The Cancer Institute of New Jersey, Life Science Commercialization Award from Rutgers University.

## Conflict of interest statement

**Authors' conflict of interest disclosure:** The authors stated that there are no conflicts of interest regarding the publication of this article.

**Research funding:** None declared.

**Employment or leadership:** None declared.

**Honorarium:** None declared.

## References

1. Finn WG, Carter KM, Raich R, Stoolman LM, Hero AO. Analysis of clinical flow cytometric immunophenotyping data by clustering on statistical manifolds: treating flow cytometry data as high-dimensional objects. *Cytometry B Clin Cytom* 2008; 76:1–7.
2. Habib LK, Finn WG. Unsupervised immunophenotypic profiling of chronic lymphocytic leukemia. *Cytometry B Clin Cytom* 2006;70:124–35.
3. Rogers WT, Holyst HA. FlowFP: a bioconductor package for fingerprinting flow cytometric data. *Adv Bioinformatics* 2009; 193947.
4. Burchardt M, Engers R, Muller M, Burchardt T, Willers R, Epstein JI, et al. Interobserver reproducibility of Gleason grading: evaluation using prostate cancer tissue microarrays. *J Cancer Res Clin Oncol* 2008;134:1071–8.
5. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L. Interobserver reproducibility of percent Gleason grade 4/5 in total prostatectomy specimens. *J Urol* 2002;168:2006–10.
6. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L. Interobserver reproducibility of percent Gleason grade 4/5 in prostate biopsies. *J Urol* 2004;171:664–7.
7. Veloso SG, Lima MF, Salles PG, Berenstein CK, Scalon JD, Bamber EA. Interobserver agreement of Gleason score and modified Gleason score in needle biopsy and in surgical specimen of prostate cancer. *Int Braz J Urol* 2007;33:639–46.
8. Rodenacker K, Bengtsson E. A feature set for cytometry on digitized microscopic images. *Anal Cell Pathol* 2003;25:1–36.
9. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998;393:440–2.

10. Wuchty S, Stadler PF. Centers of complex networks. *J Theor Biol* 2003;223:45–53.
11. Marceau R. Normalization of the minimum spanning tree. *Anal Cell Pathol* 1993;5:177–86.
12. Albert R, Schindewolf T, Baumann I, Harms H. Three-dimensional image processing for morphometric analysis of epithelium sections. *Cytometry* 1992;13:759–65.
13. Boucheron LE, Harvey NR, Manjunath BS. On using cytoplasm and stroma features in automated cancer diagnosis workshop on bio-image informatics: biological imaging, computer vision and data mining. Santa Barbara, CA, USA, 2008.
14. Somol P, Pudil P, Kittler J. Fast branch & bound algorithms for optimal feature selection. *IEEE Trans Pattern Anal Mach Intell* 2004;26:900–12.
15. Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. *J Biomed Inform* 2009;43:15–23.
16. Sima C, Dougherty ER. What should be expected from feature selection in small-sample settings. *Bioinformatics* 2006;22:2430–6.
17. Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. *Genome Res* 2001;11:1878–87.
18. Jain A, Zonker D. Feature selection: evaluation, application, and small sample performance. *IEEE Trans Pattern Anal Mach Intell* 1997;19:153–8.
19. Freund Y, Schapire RE, editors. Experiments with a new boosting algorithm. Proceedings of National Conference on Machine Learning, 1996.
20. Perkins S, Theiler J. Online feature selection using grafting. *International Conference on Machine Learning* 2003:592–9.
21. Qureshi H, Sertel O, Rajpoot N, Wilson R, Gurcan M. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. *Med Image Comput Assist Interv Int Conf Med Image Comput Assist Interv* 2008;11:196–204.
22. Jolliffe IT. *Principal component analysis*. Springer, 2002.
23. Hyvarinen A, Ramkumar P, Parkkonen L, Hari R. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. *Neuroimage* 2010;49:257–71.
24. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans Neural Netw* 1999;10:626–34.
25. Hyvarinen A, Pajunen P. Nonlinear independent component analysis: existence and uniqueness results. *Neural Netw* 1999;12:429–39.
26. Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. *Neural Netw* 2000;13:411–30.
27. Qiao H, Zhang P, Zhang B, Zheng S. Learning an intrinsic-variable preserving manifold for dynamic visual tracking. *IEEE Trans Syst Man Cybern B Cybern* 2009 Nov 13 (Epub ahead of print).
28. Takahashi S, Fujishiro I, Okada M. Applying manifold learning to plotting approximate contour trees. *IEEE Trans Vis Comput Graph* 2009;15:1185–92.
29. Hou Y, Zhang P, Xu X, Zhang X, Li W. Nonlinear dimensionality reduction by locally linear inlaying. *IEEE Trans Neural Netw* 2009;20:300–15.
30. Lin T, Zha H. Riemannian manifold learning. *IEEE Trans Pattern Anal Mach Intell* 2008;30:796–809.
31. Li X, Lin S, Yan S, Xu D. Discriminant locally linear embedding with high-order tensor data. *IEEE Trans Syst Man Cybern B Cybern* 2008;38:342–52.
32. Keeratipranon N, Maire F, Huang H. Manifold learning for robot navigation. *Int J Neural Syst* 2006;16:383–92.
33. Law MH, Jain AK. Incremental nonlinear dimensionality reduction by manifold learning. *IEEE Trans Pattern Anal Mach Intell* 2006;28:377–91.
34. Madabhushi A, Yang P, Rosen M, Weinstein S. Distinguishing lesions from posterior acoustic shadowing in breast ultrasound via non-linear dimensionality reduction. *Conf Proc IEEE Eng Med Biol Soc* 2006;1:3070–3.
35. Yan S, Xu D, Zhang B, Zhang HJ, Yang Q, Lin S. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans Pattern Anal Mach Intell* 2007;29:40–51.
36. Madabhushi A, Shi J, Rosen M, Tomaszewski JE, Feldman MD. Graph embedding to improve supervised classification and novel class detection: application to prostate cancer. *Med Image Comput Assist Interv* 2005;8:729–37.
37. Higgs BW, Weller J, Solka JL. Spectral embedding finds meaningful (relevant) structure in image and microarray data. *BMC Bioinformatics* 2006;7:74.
38. Kuncheva L. An aggregation of pro and con evidence for medical decision support systems. *Comput Biol Med* 1993;23:417–24.
39. Kuncheva LI. Evaluation of computerized medical diagnostic decisions via fuzzy sets. *Int J Biomed Comput* 1991;28:91–100.
40. Kuncheva LI. Fuzzy multi-level classifier for medical applications. *Comput Biol Med* 1990;20:421–31.
41. Doyle S, Hwang M, Shah K, Madabhushi A, Tomaszewski J, Feldman M. Automated grading of prostate cancer using architectural and textural image features. *International Symposium on Biomedical Imaging*. Washington, DC. 2007;1284–7.
42. Rajpoot K, Rajpoot N. SVM optimization for hyperspectral colon tissue cell classification. *MICCAI* 2004;829–37.
43. Qureshi H, Rajpoot N, Wilson R, Nattkemper T, Hans V, editors. Comparative analysis of discriminant wavelet packet features and raw image features for classification of meningioma subtypes. *Medical Image Understanding and Analysis* 2007; Aberystwyth (UK).
44. Doyle S, Madabhushi A, Feldman M, Tomaszewski J. A boosting cascade for automated detection of prostate cancer from digitized histology. *Med Image Comput Assist Interv* 2006;9:504–11.
45. Viola P, Jones MJ. Robust real-time face detection. *Int J Computer Vision* 2004;57:137–54.
46. Jafari-Khouzani K, Soltanian-Zadeh H. Multiwavelet grading of pathological images of prostate. *IEEE Trans Biomed Eng* 2003;50:697–704.
47. Weyn B, van de Wouwer G, van Daele A, Scheunders P, van Dyck D, van Marck E, et al. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry* 1998;33:32–40.
48. Tabesh A, Teverovskiy M, Pang HY, Kumar VP, Verbel D, Kotsiantzi A, et al. Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans Med Imaging* 2007;26:1366–78.
49. Esgiar AN, Naguib RN, Sharif BS, Bennett MK, Murray A. Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa. *IEEE Trans Inf Technol Biomed* 1998;2:197–203.
50. Esgiar AN, Naguib RN, Bennett MK, Murray A. Automated feature extraction and identification of colon carcinoma. Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology. 1998;20:297–301.
51. Basavanthally AN, Ganesan S, Agner S, Monaco JP, Feldman MD, Tomaszewski JE, et al. Computerized image-based detec-

- tion and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. *IEEE Trans Biomed Eng* 2010;57:642–53.
52. Basavanhally A, Agner A, Alexe G, Ganesan S, Bhanot G, Madabhushi A. Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade breast cancer histology. MMBIA workshop in conjunction with MICCAI 2008; New York, NY 2008.
53. Fatakawala H, Basavanhally A, Xu J, Bhanot G, Ganesan S, Feldman M, et al., editors. Expectation maximization driven geodesic active contour: application to lymphocyte segmentation on digitized breast cancer histopathology. Taiwan: IEEE International Conference on Bioinformatics and Bioengineering (IEEE BIBE), 2009.
54. Alexe G, Dalgin GS, Scanzfeld D, Tamayo P, Mesirov JP, DeLisi C, et al. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. *Cancer Res* 2007;67:10669–76.

Copyright of Clinical Chemistry & Laboratory Medicine is the property of De Gruyter and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.