

# Feature Importance in Nonlinear Embeddings (FINE): Applications in Digital Pathology

Shoshana B. Ginsburg, George Lee, Sahirzeeshan Ali, Anant Madabhushi

Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA

**Abstract**—Quantitative histomorphometry (QH) refers to the process of computationally modeling disease appearance on digital pathology images. This procedure typically involves extraction of hundreds of features, which may be used to predict disease presence, aggressiveness, or outcome, from digitized images of tissue slides. Due to the “curse of dimensionality”, constructing a robust and interpretable classifier is very challenging when the dimensionality of the feature space is high. Dimensionality reduction (DR) is one approach for reducing the dimensionality of the feature space to facilitate classifier construction. When DR is performed, however, it can be challenging to quantify the contribution of each of the original features to the final classification or prediction result. In QH it is often important not only to create an accurate classifier of disease presence and aggressiveness, but also to identify the features that contribute most substantially to class separability. This feature transparency is often a prerequisite for adoption of clinical decision support classification tools since physicians are typically resistant to opaque “black box” prediction models. We have previously presented a method for scoring features based on their importance for classification on an embedding derived via principal components analysis (PCA). However, nonlinear DR (NLDR), which is more useful for many biomedical problems, involves the eigen-decomposition of a kernel matrix rather than the data itself, compounding the issue of classifier interpretability. In this paper we extend our PCA-based feature scoring method to kernel PCA (KPCA). We demonstrate that our KPCA approach for evaluating feature importance in nonlinear embeddings (FINE) applies to several popular NLDR algorithms that can be cast as variants of KPCA, such as Isomap and Laplacian eigenmaps. FINE is applied to four digital pathology datasets with 53–2343 features to identify key QH features describing nuclear or glandular arrangements for predicting the risk of recurrence of breast and prostate cancers. Measures of nuclear and glandular architecture and clusteredness were found to play an important role in predicting the likelihood of recurrence of both breast and prostate cancers. Additionally, FINE was able to identify a stable set of features that provide good classification accuracy on four publicly available datasets from the NIPS 2003 Feature Selection Challenge. Compared to the  $t$ -test, Fisher score, and Gini index, FINE was found to yield more stable feature subsets that achieve higher classification accuracy for most of the datasets considered.

**Index Terms**—Quantitative histomorphometry, digital pathology, dimensionality reduction, feature selection.

## I. INTRODUCTION

Quantitative histomorphometry (QH) is the process of computationally modeling the appearance of disease on digital pathology images via image-based features. QH approaches typically involve the extraction of a large number of features

describing the texture, color, and spatial arrangement of nuclei and glands on digitized images of tissue slides [1], [2], [3], [4], [5], [6]. These features have been shown to be useful in determining cancer aggressiveness [4], [5], [6] and in predicting the likelihood of a patient’s cancer recurring following treatment [3], [7]. Nevertheless, since it is typical to extract hundreds or even thousands of features from digital pathology images, the dimensionality of the feature space poses a formidable challenge to the construction of robust classifiers for predicting disease presence and aggressiveness. These challenges have to do with the “curse of dimensionality” [8] and the “curse of data sparsity” [9] since the number of features is significantly larger than the number of training exemplars (typically patient studies). A per-class sample-to-feature ratio of 10:1 is generally recommended for building reliable and generalizable classifiers and predictive models [10].

In order to reduce the dimensionality of the feature space and thereby facilitate classifier construction, dimensionality reduction (DR) is often performed [11], [12], [13], [14]. DR involves transforming a high dimensional dataset into a low-dimensional eigenspace; the resulting eigenvectors replace the original features as inputs to the classifier. The rationale here is that classifiers trained on features in a reduced dimensional space are more robust and reproducible than classifiers constructed in the original high dimensional feature space. Additionally, since the reduced dimensional features capture most of the data variance embedded in the original high dimensional space, the reduced dimensional classifiers are no worse off in terms of class discriminability than classifiers in the original space. However, one major disadvantage of DR is the fact that the new, transformed features tend to be divorced from domain-specific meaning. Consequently, resulting classifiers are opaque, and it is a challenge to identify which features contribute most substantially to class discriminability. In problems involving clinical decision support for medical imaging data, it is often important to not only be able to create an accurate classifier of disease presence and aggressiveness, but also to identify the features that contribute most substantially to class separability. This feature transparency is often a prerequisite for adoption of clinical decision support classification tools since physicians are typically resistant to opaque “black box” prediction models.

However, linear DR methods (e.g. principal components analysis (PCA)) rely on Euclidean distances to estimate similarities between features and do not account for the inherent nonlinear structures underlying most biomedical data. For

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org).

high dimensional biomedical datasets, nonlinear DR (NLDR) methods can capture the underlying nonlinear manifold structure of the data. As a result, classifiers modeled on the NLDR representations have been shown to provide higher class discrimination compared to the corresponding classifiers trained off linear DR representations [15]. Since QH data is inherently nonlinear, NLDR provides a superior embedding of the data when compared to linear DR [6], [16], [17]. However, whereas PCA involves the eigen decomposition of the data itself, NLDR involves the eigen decomposition of a similarity matrix instead. Consequently, compared to PCA, it is even more difficult to reconcile the contributions of the individual features to the low dimensional representations constructed via NLDR schemes.

Apart from DR, feature selection is another approach to reduce the dimensionality of the feature space and thereby enable classifier construction. Feature selection methods include wrappers and embedded methods [18], which typically involve selection of feature subsets that perform well in a particular classifier training framework, and filters [19], [20], which select features based on a variable ranking scheme. Because wrappers and embedded methods utilize a learning algorithm for feature selection, the selected features may not perform well in conjunction with other classifiers. As a result, these methods may not be ideal for relating which computer-extracted features are most strongly associated with disease appearance on digital pathology images. Instead, filter methods are more robust against overfitting, providing more stable sets of selected features [21]. However, one limitation of filters is their inability to account for interactive effects among features. Also, filter methods, as well as wrapper and ensemble methods, may lead to a number of different “optimal” feature subsets that produce equally good classification results [9].

In this paper we present a new classification method that relies on NLDR to overcome the curse of dimensionality but that allows for identifying the relative contributions of the individual features to the reduced dimensional manifold. This is accomplished by developing a filter to rank high dimensional features based on the extent to which they contribute to accurate classification in a low dimensional eigenspace obtained via DR. Since feature ranking is performed in the low-dimensional embedding space, which is more robust to small data perturbations than the original high-dimensional feature space, feature rankings are more stable than those provided by traditional filter methods.

We have previously introduced Variable Importance in Projection (VIP), a variable ranking scheme to quantify the contributions of individual features to classification on an embedding derived via PCA [22]. This variable ranking scheme exploits the mapping between the original feature space and the data in the PCA embedding space to quantify the contributions of individual features. The challenge with directly extending this idea to NLDR schemes is that NLDR involves the eigen-decomposition of a kernelized form of the data. Furthermore, this kernel need not be explicitly defined. Since there is no closed-form expression for the mapping between the original feature space and the embedding space, the reverse mapping is also not defined.

In this paper we present a general approach, which we call feature importance in nonlinear embeddings (FINE), for ranking features based on their contributions to accurate classification in a low dimensional embedding space obtained via NLDR. This is accomplished by approximating the mapping between the data in the original feature space and the low dimensional representation of the data obtained by kernel PCA. Once this mapping has been estimated, the contributions of individual features to classification can be computed in the same manner as VIP. Furthermore, several NLDR schemes, including Isomap [23] and Laplacian eigenmaps [24], have been shown to be analogous to kernel PCA by simply modifying the kernel function [25]. In this work we show that FINE can be applied to assess the importance of individual features in disease characterization in the context of a variety of different NLDR schemes. We apply FINE to four publicly available datasets designed for benchmarking feature selection algorithms. Additionally, we employ FINE to identify QH features that are associated with disease aggressiveness and outcome in the context of four different scenarios involving breast and prostate cancer digitized tissue microarray and whole slide images.

In the context of breast cancer, we focus on two problems that relate to predicting treatment outcome and risk assessment of estrogen receptor positive (ER+) breast cancers. The working hypothesis for the two breast cancer problems we address in this paper is that computer-extracted features describing cancer patterns in ER+ breast cancer tissue slides can predict (a) which cancer patients will have recurrence following treatment with tamoxifen and (b) risk category as determined by a 21 gene expression assay called Oncotype DX. Oncotype DX is a commercial assay that allows for prediction of which patients would not benefit from adjuvant chemotherapy (identified as low risk patients) and which patients would benefit from adjuvant chemotherapy (identified as high risk patients). In the context of prostate cancer, we employ FINE to identify computer-extracted image features of cancer patterns on tissue images that are associated with risk of biochemical recurrence. Toward this end, we look at QH features extracted from tissue microarray images obtained from needle core biopsies and whole mount radical prostatectomy specimens. For both breast and prostate cancers, knowledge about which QH features are most predictive of risk of recurrence could potentially lead to improved disease characterization upon biopsy and better planning of therapy in the adjuvant and neoadjuvant settings. FINE provides the ability to identify key QH features that contribute most substantially to class discriminability.

The rest of this paper is organized as follows. In Section III we show how several DR algorithms can be formulated in terms of eigenvalue problems, and in Section IV we demonstrate how FINE serves as a general framework for ranking features in terms of their contribution to class discrimination on low dimensional embeddings obtained by solving an eigenvalue problem. In Section V FINE is (1) evaluated in conjunction with four DR methods on publicly available datasets from the NIPS 2003 Feature Selection Challenge [26] and (2) applied for feature discovery in four different digital pathology scenarios involving breast and prostate cancer. In

Section VI we present our results, which are further discussed in Section VII. In Section VIII we present our concluding remarks.

## II. RELATED WORK AND NOVEL CONTRIBUTIONS

A number of groups have applied DR to reduce the dimensionality of high dimensional biomedical data and thereby construct more robust classifiers [11], [12], [13], [27], [28], [29], [30]. These groups first extract a large number of features from biomedical images or signals, apply DR to obtain a set of low dimensional features, and use the newly derived features as input to a classifier or predictive model. However, to the best of our knowledge, these approaches have not attempted to identify which of the original, high dimensional features contribute most substantially to class discriminability in the transformed, low-dimensional feature space.

Several groups have combined feature selection and DR in a single optimization routine by achieving a sparse reconstruction of the high dimensional data [31], [32], [33]. Other groups have employed more traditional feature selection routines (e.g. sequential feature selection) in conjunction with NLDR to preserve the structure of the embedding while eliminating redundant features [34], [35], [36]. In both cases the number of high dimensional features contributing to the embedding is significantly decreased. Consequently, this makes the problem of model interpretation in the low dimensional embedding space a little more tractable. However, these approaches do not permit interpretation or quantification of the contributions of the selected features to class discriminability in the reduced dimensional embedding space.

With regard to linear DR variants, methods for quantifying the extent that individual, high dimensional features contribute to classification on embeddings obtained via partial least squares (PLS) and PCA exist. Chong and Jun introduced the concept of variable importance in projections (VIP), a measure of each feature's contribution to classification on a low dimensional embedding obtained via PLS [37]. We extended VIP to PCA, employing VIP to identify computer-extracted features that play an important role in detecting prostate cancer on MRI [22]. However, VIP relies on the mapping between the original feature space and the data in the PCA embedding space to quantify the contributions of individual features. When NLDR is performed, however, this mapping is not explicitly defined, so the contributions of individual features to classification in an NLDR-derived embedding space cannot be determined using VIP.

The novel methodological contribution of this paper is the extension of VIP to NLDR methods. Our new approach, feature ranking in nonlinear embeddings (FINE), extends VIP from PCA to KPCA. Since VIP exploits the mapping between the original feature space and the data in the embedding space, a mapping that is not defined in the context of KPCA, FINE involves approximating this mapping so that feature importance can be measured in the same manner as VIP. Then, by substituting a similarity matrix in place of the kernel matrix in KPCA, FINE is made applicable to NLDR algorithms such as Isomap [23], Laplacian eigenmaps [24],

Symbol	Size	Description
A	$n \times m$	Data matrix
X	$n \times m$	Centered data matrix
$\Lambda$	$n \times m$	Diagonal matrix of singular values
U	$n \times n$	Matrix of eigenvectors
V	$m \times m$	Matrix of right-singular vectors, or loadings
$\Sigma$	$n \times m$	Diagonal matrix of eigenvalues
T	$n \times m$	Principal components matrix
Y	$n \times 1$	Class labels
K	$n \times n$	Kernel matrix
Z	$n \times n$	Eigenvectors of kernel matrix
L	$n \times m$	Graph Laplacian

TABLE I: List of common mathematical notation in this paper.

and locally linear embeddings [38]. An illustration of the FINE method is shown in Figure 1. In general, the FINE method involves performing DR to reduce the dimensionality of high dimensional data, identifying key features, and using the selected features identified by FINE to construct a robust classifier.

In this work we leverage FINE to identify QH features from digitized tissue images for (a) predicting the likelihood of breast cancer recurrence following chemotherapy and (b) predicting risk of prostate cancer biochemical recurrence following radical prostatectomy. For both of these problems, we extract QH features describing nuclear texture, morphology, and architecture from digitized tissue specimens (e.g., biopsy samples, microarrays, surgical specimens). These features are concatenated in a data matrix, and NLDR is performed. Finally, FINE is used to identify the most useful features, and these features are exploited in conjunction with a classifier to predict disease outcome.

## III. DIMENSIONALITY REDUCTION AS AN EIGENVALUE PROBLEM

In this section we review the existing theory showing how several linear and nonlinear DR algorithms can be formulated in terms of eigenvalue problems. A list of common mathematical notation used in the coming section is in Table I.

### A. Eigenvalue Problems for Dimensionality Reduction

Given a matrix  $A \in \mathbb{R}^{n \times m}$ , the eigenvectors associated with A can be found by solving an ordinary eigenvalue problem:

$$AU = \Lambda U; \quad (1)$$

where the columns of U are the eigenvectors associated with A. This eigenvalue problem can be solved by using an iterative technique or the singular value decomposition. According to the singular value decomposition,

$$A = U \Sigma V^T; \quad (2)$$

where the columns of U  $\in \mathbb{R}^{n \times n}$  are the left-singular vectors, or eigenvectors of the covariance matrix  $A^T A$ ; the columns of V  $\in \mathbb{R}^{m \times m}$  are the right-singular vectors, or eigenvectors of the Gram matrix  $AA^T$ ; and  $\Sigma \in \mathbb{R}^{n \times m}$  is a diagonal matrix whose diagonal entries are the singular values of A, or the square roots of the eigenvalues shared by both the covariance and Gram matrices.

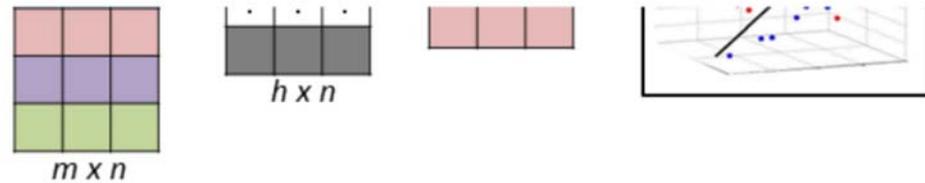


Fig. 1: Illustration of modules comprising FINE to identify key features in a high dimensional dataset: First a large number  $m$  of features (typically far greater than the number of patient samples  $n$ ) are extracted, DR is performed to reduce the dimensionality of the data to  $h$ . FINE is then leveraged to identify key features. The predictive power of these key features is then evaluated via a logistic regression classifier.

Many DR algorithms, including PCA, PLS, Isomap, and their kernelized versions, involve solving an eigenvalue problem to find an optimal subspace to embed the data. Below we review how several popular DR methods can be formulated as eigenvalue problems [39].

1) Principal Components Analysis: Consider a centered data matrix  $X \in \mathbb{R}^{n \times m}$ , whose rows represent  $n$  samples and whose columns represent  $m$  features. PCA involves solving eq. (1) in the case that  $A = X$ . This problem can be solved by the SVD as

$$X = U \Sigma V^T \quad (3)$$

or, alternatively, the eigen decomposition of  $X^T X$ :

$$X^T X = U \Sigma U^T \quad (4)$$

Here  $\Sigma$  is a diagonal matrix whose diagonal entries are the eigenvalues of  $X^T X$ . Letting  $T = U \Sigma$ , the principal components of  $X$  are the columns of  $T$ , the scores matrix containing the projection of the original data into the PCA embedding space.  $V$  is the loadings matrix, whose elements describe the correlation between the scores and the original features. Thus,

$$X = T V^T; \quad (5)$$

and  $T = X V$  since  $V^T V = I$ .

When DR is desired, only a subset of the eigenvectors are used to reconstruct the data. Thus,

$$X \approx T_h V_h^T; \quad (6)$$

where  $h \in \mathbb{N}; h < m$ ,  $T_h \in \mathbb{R}^{n \times h}$ , and  $V_h \in \mathbb{R}^{m \times h}$ .

2) Kernel PCA: Whereas both PCA and PLS only take into account linear relationships among features, kernel PCA (KPCA) involves a nonlinear mapping of the data. Kernel PCA involves the eigen decomposition of a centered kernel matrix  $K$ , as follows:

$$K = Z \Lambda Z^T \quad (7)$$

Thus, the principal components of  $K$  are contained in

$$T = K Z^{\frac{1}{2}} \quad (8)$$

In KPCA the kernel matrix  $K$  does not need to be explicitly defined. Rather,  $K$  can be defined in the dual form:

$$K_{ij} = \langle \phi(x_i); \phi(x_j) \rangle \quad (9)$$

Consequently,  $K$  can represent any kernel, such as a Gaussian or radial basis function kernel, or a distance or similarity matrix.

## B. NLDR Methods as Variants of KPCA

Many NLDR algorithms, including Isomap, Laplacian eigenmaps, and locally linear embeddings (LLE), involve the eigen decomposition of a similarity matrix. Here we review how these three NLDR algorithms can be formulated as variants of KPCA [39].

1) Isomap: Isomap [23] involves creating a neighborhood graph by using  $K$  nearest neighbors or  $\epsilon$ -neighborhoods to determine a set of neighboring points associated with each data-point. Then, Isomap estimates the geodesic distances  $d(i;j)$  between points  $i$  and  $j$  on the graph by computing the shortest

path between points. Thus, we obtain the kernel matrix  $K$  whose elements are

$$K_{ij} = d(i;j) \quad (10)$$

Since all elements of  $K$  are positive,  $K$  must be centered before Isomap can be solved by eq. (7) as a variant of KPCA.

2) Laplacian Eigenmaps: Given a neighborhood graph constructed using  $K$  nearest neighbors or  $\square$ -neighborhoods, Laplacian eigenmaps (LE) [24] determines an optimal manifold representation of the data by computing the graph Laplacian  $L$ :

$$L_{ij} = \begin{cases} \sum_{\theta} P_{i\theta} e^{-\frac{jx_i - x_j}{2\sigma^2}} & \text{if } i = j \\ \frac{1}{2} \sum_{\theta} P_{i\theta} e^{-\frac{jx_i - x_j}{2\sigma^2}} & \text{if } i \neq j \\ 0 & \text{else} \end{cases}$$

Laplacian eigenmaps is equivalent to KPCA such that  $K = L^y$ , where  $y$  denotes the pseudo-inverse.

3) Locally Linear Embeddings: LLE [38] constructs a neighborhood-preserving mapping by computing the Euclidean distances  $D_{ij}$  between points and their nearest neighbors, creating a local covariance matrix  $C$ :

$$C_{ij} = \frac{1}{2} (D_i + D_j - D_{ij})^2 \quad (11)$$

LLE is equivalent to KPCA such that

$$K = \square_{\max} I \square M \quad (12)$$

Here  $I$  is an identity matrix and  $M = (I \square W)(I \square W^T)$ , where the  $w_i = \frac{C_{ij}}{\sum_{ij} C_{ij}}$  are the linear coefficients that optimally reconstruct  $x_i$  from its nearest neighbors.

#### IV. CALCULATING FEATURE IMPORTANCE

In this section we review how the variable importance in projection (VIP) score allows for feature weighting and ranking based on their contributions to classification within a linearly-derived embedding. Then we discuss our extension to VIP that enables the computation of FINE scores when the embedding is derived using an NLDR scheme.

##### A. Variable Importance in Projections (VIP)

The importance of an individual feature to classification on an embedding depends on two factors: how much each eigenvector contributes to the embedding, and how much each feature contributes to each eigenvector. All of the DR algorithms discussed in Section III provide a projection matrix  $T$  and a loadings matrix  $V$ . The features that contribute most to the  $i^{\text{th}}$  dimension of the embedding are those with the largest weights in the  $i^{\text{th}}$  loading vector. Thus, the fraction  $\frac{v_{ji}}{\sum_{jj} v_{jj}}$  reveals how much the  $j^{\text{th}}$  feature contributes to the  $i^{\text{th}}$  principal component in the low-dimensional embedding. The overall importance of the  $j^{\text{th}}$  feature depends also on (a) the regression coefficients  $b_i$ , which relate the transformed data back to the class labels, and (b) the transformed data  $t_i$ . The

variable importance in projections (VIP) score is computed for each feature as follows:

$$\Delta_j = \frac{\sum_{i=1}^m b_i^2 t_i^T t_i \frac{v_{ji}}{\sum_{jj} v_{jj}}}{\sum_{i=1}^m b_i^2 t_i^T t_i} \quad (13)$$

where  $m$  is the number of features in the original, high-dimensional feature space,  $t_i$  is the  $i^{\text{th}}$  principal component vector, and  $v_i$  is the  $i^{\text{th}}$  loading vector. The  $b_i$  are the coefficients that solve the regression equation

$$y = T b^T \quad (14)$$

which correlates the scores with the outcome vector  $y$ .

For DR methods that are intrinsically unsupervised, the exploitation of class labels in computing the VIP score leads to the identification of features that provide good class discrimination in the embedding space. The degree to which a feature contributes to classification in the transformed space is directly proportional to its associated VIP score. Thus, features with VIP scores near 0 have little predictive power, and the features with the highest VIP scores contribute the most to class discrimination on the embedding.

Alternatively, VIP scores can be normalized to the interval  $[0; 1]$  as follows:

$$\hat{\Delta}_j = \frac{\Delta_j^2}{m} \quad (15)$$

When the VIP scores are normalized in this way,  $\sum_{j=1}^m \hat{\Delta}_j = 1$ . Consequently,  $\hat{\Delta}_j$  is the fraction that feature  $j$  contributes to classification in the embedding space compared to the entire feature set. Furthermore, the aggregate VIP score associated with a feature subset  $J = \{1, \dots, m\}$  can be calculated as

$$\hat{\Delta}_J = \frac{1}{m} \sum_{j \in J} \Delta_j^2 \quad (16)$$

##### B. Feature Importance in Nonlinear Embeddings (FINE)

The expression for computing the importance of features according to eq. (13) relies on the loadings matrix  $V$ . For KPCA and its variants (e.g., Isomap, Laplacian eigenmaps, and locally linear embeddings)  $V$  is not defined because the transformed data is not directly related to the original data, but only to the kernel matrix (see eq. (7)). Furthermore, the mapping  $\square : X \rightarrow K$  that relates the kernel matrix to the original data is not necessarily computed, as  $\square$  is only implicitly defined.

Nevertheless, combining equations (5) and (8) yields

$$X = K Z^{\frac{1}{2}} V^T \quad (17)$$

Thus, without explicit knowledge of  $\square$ , we can estimate  $V$  as follows:

$$V^T \square (K Z^{\frac{1}{2}})^y X \quad (18)$$

Approximating  $V$  as  $V^0 = X^T ((K Z^{\frac{1}{2}})^y)^T$  facilitates the computation of the feature importance in nonlinear embeddings (FINE) using eq. (13).

Title	Topic	Features	Feature Type	Training Samples	Testing Samples
Madelon	Random data	500	Synthetic	2000	600
Arcene	Cancer detection	10,000	Protein expression	100	100
Dexter	Text filtering	20,000	Word frequencies	300	300
Gisette	Handwritten digit recognition	5000	Pixel values	6000	1000

TABLE II: Description of four publicly available datasets used in this paper.

Dataset	Objective	Training Samples	Testing Samples	Features	Feature Type
S <sub>1</sub>	Predict recurrence of ER+ breast cancer based on tissue microarrays	36	12	53	Nuclear graph-based, texture
S <sub>2</sub>	Classify high/low OncotypeDX scores on biopsy samples from breast cancer patients	106	34	2343	Nuclear shape, texture
S <sub>3</sub>	Predict post-prostatectomy biochemical recurrence on whole mounts	30	10	242	Glandular graph-based, texture
S <sub>4</sub>	Predict biochemical recurrence of prostate cancer on tissue microarrays during active surveillance	34	10	242	Nuclear graph-based, texture

TABLE III: Description of datasets used in this paper to address four digital pathology problems. ER=estrogen receptor.

## V. EXPERIMENTAL DESIGN

### A. Datasets

1) Publicly Available Datasets: In order to evaluate FINE in terms of its ability to identify a feature subset that (a) is stable and (b) provides good classification accuracy, we chose the NIPS 2003 Feature Selection Challenge datasets because they all suffer from the “curse of dimensionality”. All of these datasets have been made publicly available as benchmarking datasets for feature selection algorithms. The NIPS 2003 Feature Selection Challenge included five datasets, all involving binary classification problems; we used the four datasets that contain non-binary features. More details regarding these four datasets can be found in Table II.

2) Predicting Breast Cancer Recurrence Based on Tissue Microarrays: Tissue microarrays were obtained from a cohort of 48 patients with ER+ breast cancer who underwent chemotherapy, some of whom experienced recurrence and some of whom remained recurrence-free (see Table III). A total of 53 QH features were extracted from tissue microarrays stained with hematoxylin and eosin. These features included quantitative descriptors of Voronoi, Delaunay, and minimum spanning tree graphs connecting nuclei within each of the individual tissue cylinders (see Figure 2). FINE is leveraged to identify which of the QH features are most useful for predicting recurrence risk.

3) Predicting OncotypeDX Risk Categories on Whole Sides: QH features were extracted from whole slide tissue biopsy images obtained from a cohort of 140 subjects (see Table III) for whom OncotypeDX recurrence risk scores were known (OncotypeDX risk scores of 0–18 indicate low risk, 19–30 indicate intermediate risk, and 30–99 indicate high risk of recurrence). Biopsy samples were stained with hematoxylin and eosin and digitized. 2343 QH features describing the texture and shape of nuclei were computed on digital pathology (see Figure 2). FINE is leveraged to identify which of the QH features are most useful for classifying low versus high OncotypeDX scores, a surrogate predictor of recurrence risk.

4) Predicting Biochemical Recurrence of Prostate Cancer Based on Tissue Whole Mounts: Tissue whole mounts were acquired from 40 subjects with biopsy-confirmed prostate cancer who underwent radical prostatectomy and were followed for at least five years afterwards. Tissue whole mounts were stained with hematoxylin and eosin, and a total of 56 cancer regions were delineated on digitized whole mounts. Finally, 242 graph-based features describing nuclear arrangements on pathology that were previously shown to be useful for prostate cancer characterization [3] were extracted from each cancerous region. These features included quantitative descriptors of Voronoi, Delaunay, and minimum spanning tree graphs connecting nuclei on digital pathology (see Figure 2); architectural features describing cell clusteredness; Fourier descriptors quantifying nuclear morphology; and texture features computed based on co-occurrence matrices. FINE is leveraged to identify which of the QH features are most useful for predicting risk of biochemical recurrence within five years of radical prostatectomy since biochemical recurrence is a major risk factor for poor outcome in prostate cancer patients [40].

5) Predicting Biochemical Recurrence of Prostate Cancer Based on Tissue Microarrays: Tissue microarrays were acquired from 44 subjects with biopsy-confirmed prostate cancer participating in an active surveillance protocol. Tissue was stained with hematoxylin and eosin, and a total of 242 graph-based features describing glandular arrangements on pathology were extracted from the microarrays (see Table III, Figure 2). These features included quantitative descriptors of Voronoi, Delaunay, and minimum spanning tree graphs connecting glands on digital pathology; architectural features describing gland arrangements; Fourier descriptors quantifying gland morphology; and texture features computed based on co-occurrence matrices. FINE is leveraged to identify which of the QH features are most useful for predicting risk of biochemical recurrence.

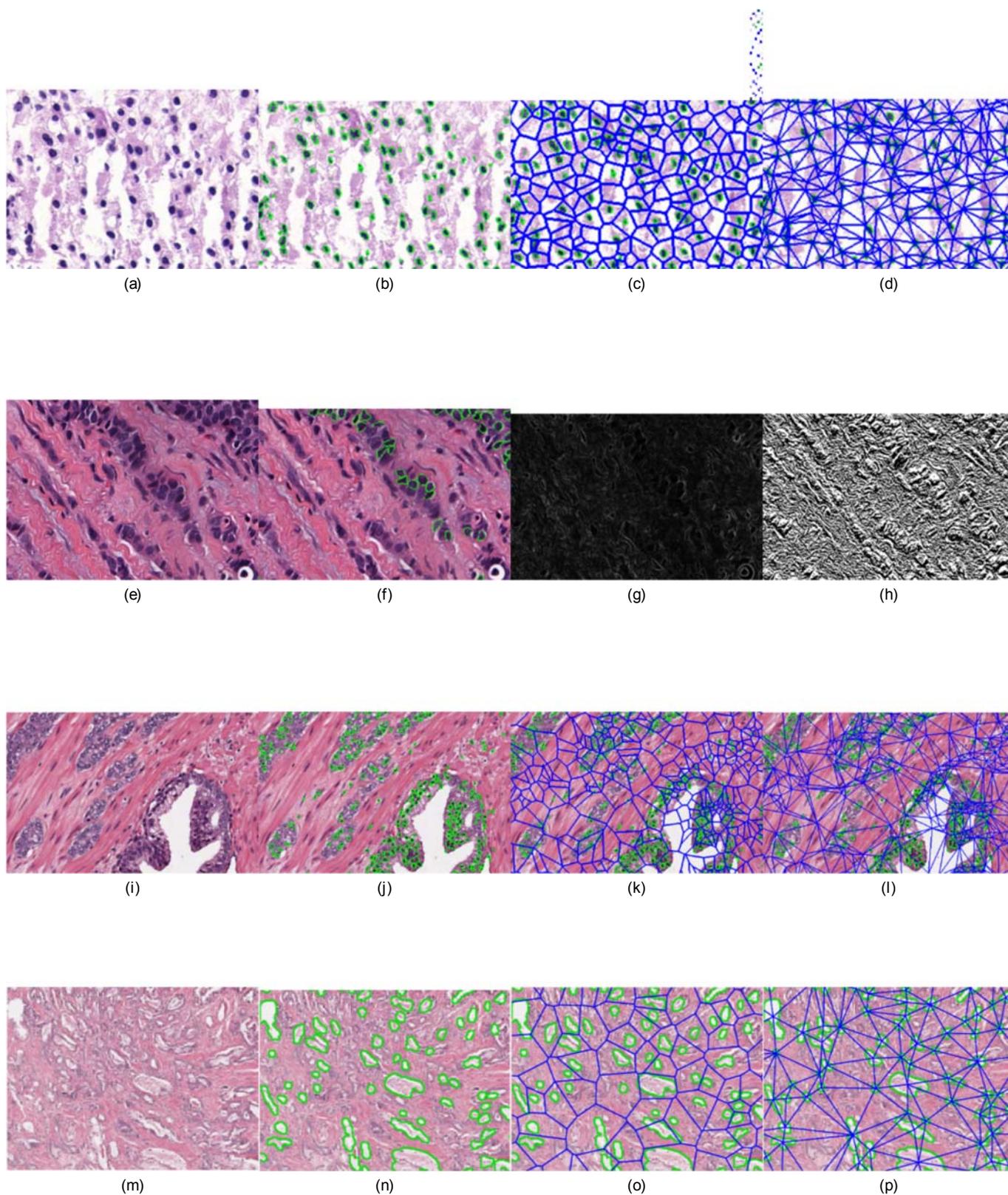


Fig. 2: Digital pathology and feature representations. For dataset S<sub>1</sub> a representative patch of histology is shown in (a); segmented nuclei are illustrated in (b); and Voronoi and Delaunay graphs constructed based on the nuclei are shown in (c) and (d), respectively. For dataset S<sub>2</sub> a representative patch of histology is shown in (e); segmented nuclei are illustrated in (f); and texture representations of the histology are shown in (g) and (h). For dataset S<sub>3</sub> a representative patch of histology is shown in (i); segmented nuclei are illustrated in (j); and Voronoi and Delaunay graphs constructed based on the nuclei are shown in (k) and (l), respectively. For dataset S<sub>4</sub> a representative patch of histology is shown in (m); segmented glands are illustrated in (n); and Voronoi and Delaunay graphs constructed based on the glands are shown in (o) and (p), respectively.

## B. Parameter Estimation and Embedding Construction

Some of the publicly available datasets evaluated in this paper contained hundreds or thousands of instances. As a result, it was necessary to randomly sample a small number of instances to construct the embeddings, which can be computationally intractable when the number of instances is too high. Consequently, 50 rounds of bootstrapping were performed. For the NIPS 2003 Feature Selection Challenge datasets 75% of the instances, up to a maximum of 100 instances, were randomly selected during each round of bootstrapping to construct the embedding and tune parameters. For the digital pathology datasets, a class-balanced set of approximately 75% of the instances were sampled during each round of bootstrapping to train the classifier, and the remaining instances were used to evaluate the classifier.

FINE was implemented to identify key contributors to classification on embeddings obtained via PCA, Isomap, LE, and LLE. Parameters associated with these DR methods—the intrinsic dimensionality parameter  $h$ , which is used in all three embeddings;  $K$ , which is used to create a neighborhood graph for Isomap, LE, and LLE; and  $\alpha$ , which is needed to compute the graph Laplacian for LE—were chosen to reduce residual variance [23]:

$$1 \square \square (D^E; D^G): \quad (19)$$

Here  $D^E$  is a matrix of Euclidean distances between points in the low-dimensional embedding,  $D^G$  is the DR algorithm's estimate of point-wise distances (e.g., for Isomap  $D^G = K$ ), and  $\square$  denotes the linear correlation coefficient. Finally, PCA, Isomap, LE, and LLE embeddings were constructed using the parameter values chosen by minimizing eq. (19).

## C. Classifier Training

Once embeddings were constructed, FINE scores were computed, and the features associated with the highest FINE scores were selected. Because a samples-to-features ratio of 10:1 is generally recommended to build a robust classifier [10], the number of selected features was limited to one tenth of the number of samples in the training dataset. Then, within the same bootstrapping rounds, classifiers were trained in conjunction with the features selected by FINE. It is important to note that the objective of FINE is not merely to select features that maximize classification accuracy, but rather to identify a stable set of features that also provide good class discriminability. Consequently, our focus was less on constructing the most accurate classifier and more on identifying features that could yield stable and reproducible classifiers. The logistic regression classifier employs linear regression when the outcome variable is binomially distributed; thus, it allows the features themselves to drive classification. Due to its simplicity, the logistic regression classifier was chosen to evaluate classifier accuracy associated with features selected by FINE. Additionally, we evaluated the classification accuracy associated with the selected features in conjunction with two other classifiers: a support vector machine (SVM) with a radial basis function kernel (scale factor = 1) and a random forest (RF) with 50 trees.

## D. Performance Evaluation Measures

The feature selection performance of FINE was evaluated in comparison to three filter methods commonly used for FS: the t-test, Fisher score, and Gini index [41]. Evaluation was performed to assess (a) stability and (b) classification accuracy associated with selected feature subsets. Stability was evaluated by the Jaccard index [42], which measures the degree of overlap between feature sets:

$$J(J_1; J_2) = \frac{|J_1 \cap J_2|}{|J_1 \cup J_2|}, \quad (20)$$

where  $J_1 \subseteq \{1, \dots, mg\}$  and  $J_2 \subseteq \{1, \dots, mg\}$  are two feature subsets. If a feature selection algorithm is stable, repeated implementations will lead to similar feature subsets and  $J$  close to 1.

Additionally, classification accuracy associated with feature subsets selected by the FINE method was evaluated in conjunction with LR, SVM, and RF classifiers. Based on the prediction results obtained from these classifiers, receiver operating characteristic (ROC) curves were generated, and the area under the ROC curve (AUC) was used to evaluate classifier accuracy in conjunction with different feature subsets. For the NIPS 2003 Feature Selection Challenge datasets, the classifiers were evaluated on a dedicated validation set provided by the Challenge. For the digital pathology datasets, a class-balanced set of up to 75% of the instances were sampled during each round of bootstrapping to train the classifier, and the remaining instances were used as independent data to evaluate the classifier.

## VI. EXPERIMENTAL RESULTS AND DISCUSSION

### A. Publicly Available Datasets

1) Madelon: Figure 3 displays the AUC and  $J$  values associated with feature subsets selected by the t-test, Fisher score, Gini index, and FINE. Feature subsets obtained via the t-test, Fisher score, and Gini index were highly unstable ( $J = 0.04$  in all cases), varying considerably between rounds of bootstrapping. In contrast,  $J$  associated with  $FINE_{PCA}$ ,  $FINE_{Isomap}$ , and  $FINE_{LE}$  was slightly higher (0.06–0.07), and the features obtained via  $FINE_{LLE}$  were moderately stable, yielding  $J = 0.22$ . Regardless of classification method (LR, SVM or RF), the top ranking features selected by  $FINE_{PCA}$ ,  $FINE_{Isomap}$ ,  $FINE_{LE}$  and  $FINE_{LLE}$  outperform the embedding vectors themselves, providing AUC values between 0.55 and 0.65 while leveraging only 13–16 features. In contrast, the AUC values yielded by feature sets selected via the Gini index, Fisher score, or t-test ranged from 0.54 to 0.56, with the highest AUC values relying on 97 features.

2) Arcene: In spite of the high dimensionality of this protein expression dataset, all feature selection methods provided the same perfect stability ( $J = 1$ ), although the top-ranking features were different for each feature selection algorithm. When the embedding vectors were used in conjunction with classifiers, PCA provided AUC values as high as 0.77–0.79, higher than any other method. However, features selected based on  $FINE_{PCA}$  were no better than random guessing. Whereas the Gini index and t-test provided AUC values

Dataset	$h_{PCA}$	$h_{ISO}$	$h_{LE}$	$h_{LLE}$
Madelon	36	4	24	1
Arcene	15	5	40	1
Dexter	6	2	7	19
Gisette	6	2	14	47

TABLE IV: Average value of the intrinsic dimensionality parameter  $h$ , selected via residual variance minimization, for four NIPS datasets and four dimensionality reduction algorithms.

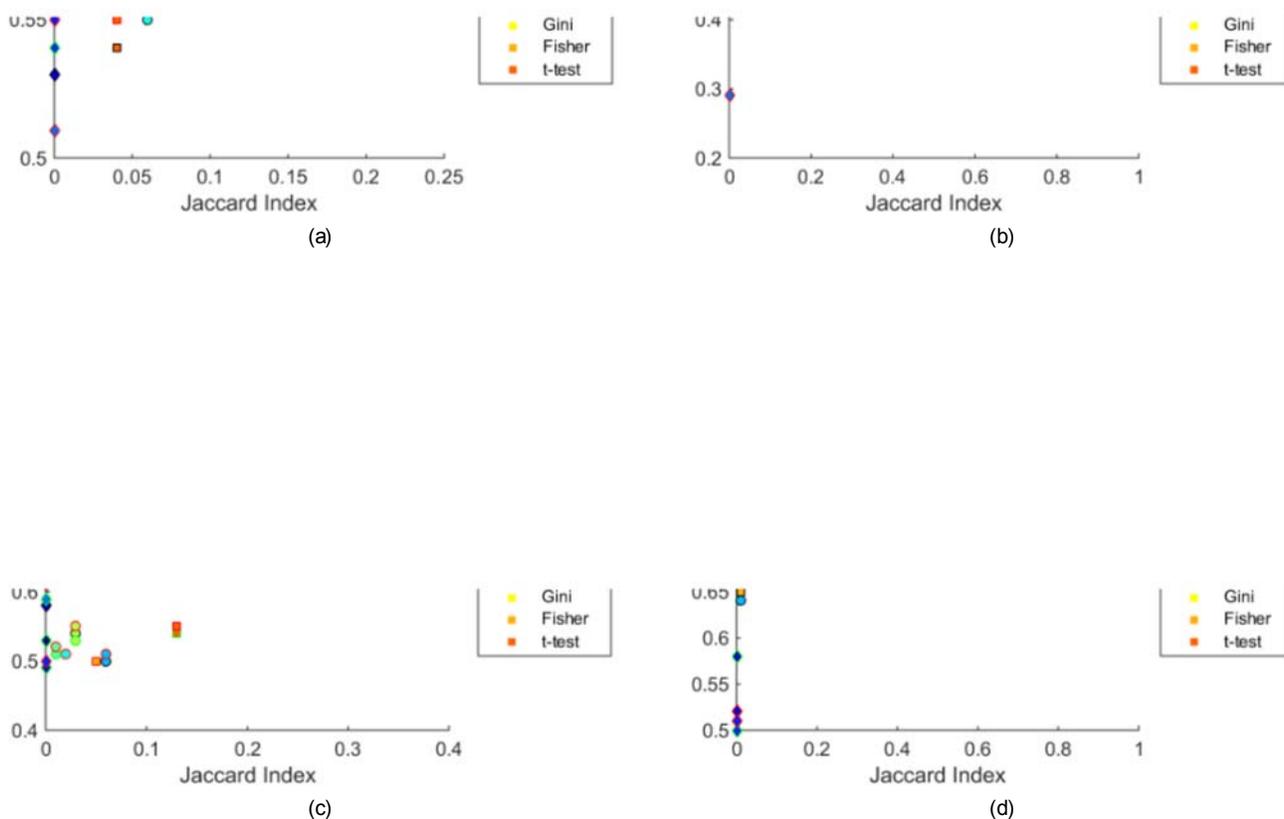


Fig. 3: AUC and Jaccard index associated with four dimensionality reduction algorithms, the top five features selected based on FINE scores, and three common feature selection algorithms for NIPS datasets (a) Madelon, (b) Arcene, (c) Dexter, and (d) Gisette. Data points are shown for three classification methods: logistic regression (black outline), support vector machine (green outline), and random forest (red outline) classifiers.

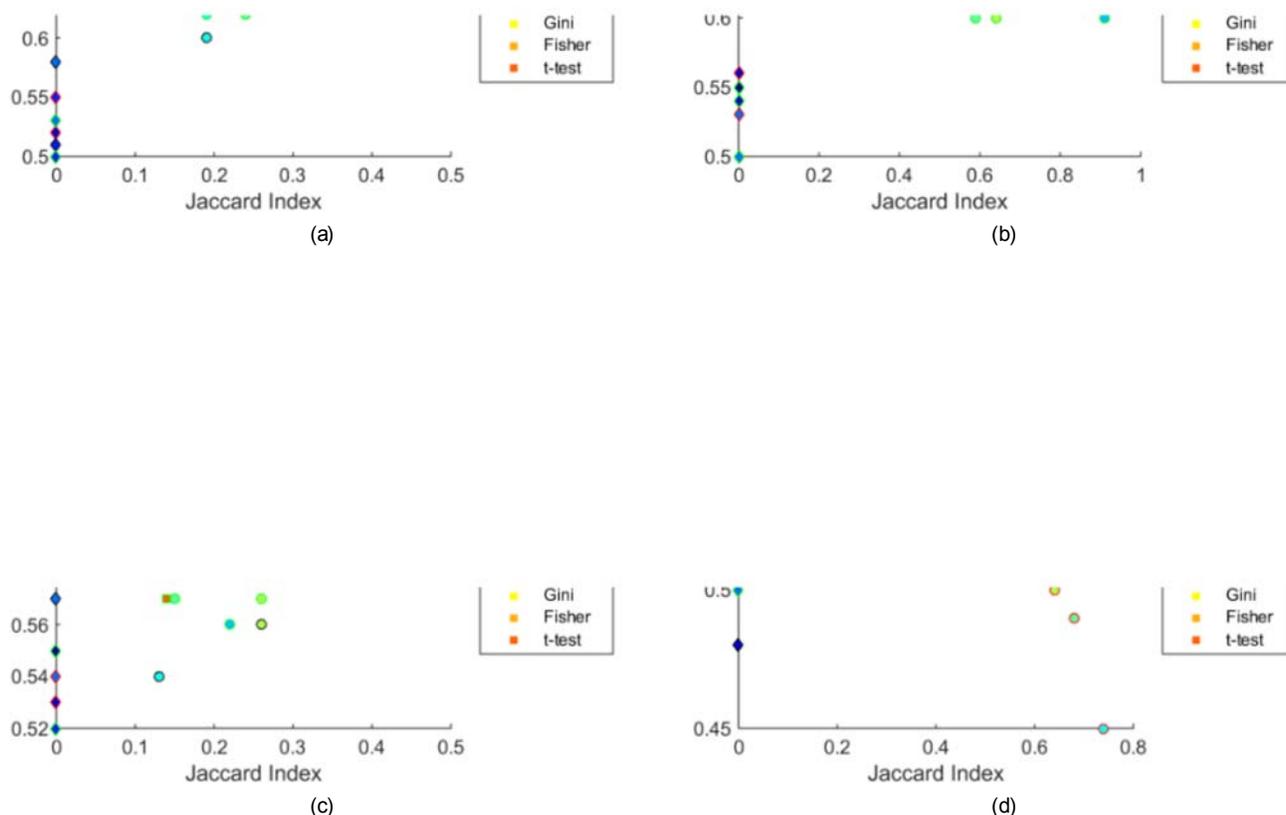


Fig. 4: AUC and Jaccard index associated with four dimensionality reduction algorithms, the top five features selected based on FINE scores, and three common feature selection algorithms for pathology datasets (a)  $S_1$ , (b)  $S_2$ , (c)  $S_3$ , and (d)  $S_4$ . Data points are shown for three classification methods: logistic regression (black outline), support vector machine (green outline), and random forest (red outline) classifiers.

between 0.69 and 0.77, the Fisher score,  $FINE_{\text{Isomap}}$ , and  $FINE_{\text{LE}}$  yielded AUC values between 0.6 and 0.7.

3) Dexter: Feature subsets selected via all four FINE methods and the Fisher score were highly unstable, probably due to the fact that this is a sparse dataset with 20,000 features. Surprisingly, both the Gini index and the t-test yielded fairly stable datasets ( $J = 0.39$  and  $J = 0.13$ , respectively). Due to the constraint of a 10:1 samples-to-features ratio, a maximum of 30 features were selected. As a result, AUC values ranged between 0.5 and 0.55 for almost all feature selection algorithms. The only exception was the Gini index, which led to feature sets associated with AUC values as high as 0.87–0.92, depending on the classifier used.

4) Gisette: The feature subset selected based on  $FINE_{\text{LE}}$  was associated with  $J = 0.99$ , whereas feature subsets

selected via  $FINE_{\text{PCA}}$ ,  $FINE_{\text{Isomap}}$ , and  $FINE_{\text{LLE}}$ , as well as the Fisher score, were highly unstable. When up to 300 features were selected, AUC values ranging from 0.91 to 0.95 were achieved for all four FINE methods, as well as for the Gini index, Fisher score, and t-test. Regardless of the DR algorithm, FINE-based feature subset selection led to higher AUC values than the embedding vectors themselves.

#### B. Predicting Breast Cancer Recurrence Based on Tissue Microarrays

Figure 4 displays the values of  $J$  associated with feature subsets selected by the t-test, Fisher score, Gini index, and FINE. For dataset  $S_1$ , the t-test, Fisher score, and Gini index provided moderately stable feature sets. Whereas  $FINE_{\text{LLE}}$ ,

Dataset	DR Method	h	Top 5 Features	Dataset	DR Method	h	Top 5 Features
S <sub>1</sub>	PCA	5	Low Hosoya index High Hosoya index Intermediate Hosoya index Mean NN in 50-pixel radius Disorder of distance to 7 NN	S <sub>2</sub>	PCA	5	Area Filled area Convex area Gray level 10 SD RGB Gray level 55 SD RGB
	Isomap	5	Disorder of distance to 7 NN Mean NN in 50-pixel radius SD of NN in 50-pixel radius High Hosoya index Low Hosoya index		Isomap	5	Area Filled area Convex area Gabor 90 mean RGB Gabor 88 mean RGB
	LE	5	Disorder of distance to 7 NN Mean NN in 50-pixel radius SD of NN in 50-pixel radius High Hosoya index Intermediate Hosoya index		LE	7	Area Filled area Convex area Gray level 10 SD RGB Gray level 55 SD RGB
	LLE	23	Low Hosoya index High Hosoya index Intermediate Hosoya index Delaunay triangle area min/max Delaunay triangle side length min/max		LLE	5	Area Filled area Convex area Gray level 10 SD RGB Gray level 55 SD RGB
S <sub>3</sub>	PCA	5	Fourier descriptor 5 SD Delaunay side length min/max Mean tensor correlation Invariant moment 1 min/max SD of distance to 3 NN	S <sub>4</sub>	PCA	6	Invariant moment 3 min/max Invariant moment 4 min/max Mean connected component size Invariant moment 7 mean Invariant moment 5 min/max
	Isomap	5	Mean information measure 1 SD of contrast variance SD of intensity variance Mean NN in 40-pixel radius Perimeter ratio SD		Isomap	5	Invariant moment 5 SD Fourier descriptor 8 mean Delaunay triangle area min/max Invariant moment 5 mean Mean connected component size
	LE	5	Fourier descriptor 1 min/max Tensor contrast energy SD Tensor contrast energy range SD of entropy Invariant moment 6 min/max		LE	5	Invariant moment 3 min/max Invariant moment 4 min/max Invariant moment 6 SD Fourier descriptor 10 mean Fourier descriptor 8 SD
	LLE	5	Fourier descriptor 7 mean Fourier descriptor 9 min/max Fourier descriptor 7 SD Invariant moment 6 min/max Invariant moment 5 min/max		LLE	21	Fourier descriptor 8 median Invariant moment 7 median Fourier descriptor 9 median Invariant moment 2 min/max Delaunay triangle area min/max

TABLE V: Average h and top features selected by FINE in conjunction with four DR methods for each of the pathology datasets (obtained by voting across 50 bootstrapping rounds). NN=nearest neighbors, SD=standard deviation.

$FINE_{Isomap}$ , and  $FINE_{LE}$  provided low to moderately stable feature subsets,  $FINE_{PCA}$  provided higher stability ( $J = 0.5$ ).

AUC values obtained by using the selected features in conjunction with LR, SVM, and RF classifiers are shown in Figure 4. The top ranking features selected by  $FINE_{PCA}$  and  $FINE_{LLE}$  perform on par with features selected by the Gini index, t-test and Fisher score, yielding AUC values between 0.8 and 0.83. For all four DR methods, the top-ranking features selected by FINE provided higher AUC values than the embedding vectors themselves. The top-ranking features selected by  $FINE_{PCA}$ ,  $FINE_{Isomap}$ ,  $FINE_{LE}$ , and  $FINE_{LLE}$  were similar, including the number of subgraphs with low, intermediate, and high Hosoya indices and local measures of cell clusteredness (see Table V). Hosoya features are closely associated with cell clusteredness since the Hosoya index measures connectedness of subgraphs representing clusters of cells. Since highly proliferative tumors tend to manifest more closely clustered cells, cell clusteredness is closely related with proliferation, which is highly predictive of recurrence risk for breast cancer [43].

### C. Predicting OncotypeDX Risk Categories on Whole Slides

For dataset S<sub>2</sub>, J associated with feature subsets selected by the t-test, Fisher score, Gini index, and FINE remains consistently below 0.27 (see Figure 4). Nevertheless,  $FINE_{PCA}$ ,  $FINE_{Isomap}$ ,  $FINE_{LE}$ , and  $FINE_{LLE}$  all provide higher stability than the comparative strategies, with J ranging from 0.64–0.91. The stability of feature rankings obtained via FINE can be attributed to the robustness of the data embedding in the low dimensional space.  $FINE_{PCA}$  provides the most stable feature subset, followed by  $FINE_{ISO}$ ,  $FINE_{LE}$ , and  $FINE_{LLE}$ .

AUC values obtained by using the selected features in conjunction with three types of classifiers are shown in Figure 4. By selecting feature subsets comprised of three features or less, FINE provided AUCs ranging between 0.75–0.76, higher than the t-test, Fisher score, and Gini index. As was the case with dataset S<sub>1</sub>, the histomorphometric features ranked highest by FINE were relatively independent of DR scheme. In fact, the top three features were always nuclear area, convex area, and filled area, as nuclear area tends to be higher in patients with high OncotypeDX risk scores (see Figure 5). Two Gabor features were found useful in conjunction with Isomap, while two attributes of gray level histograms were useful in con-

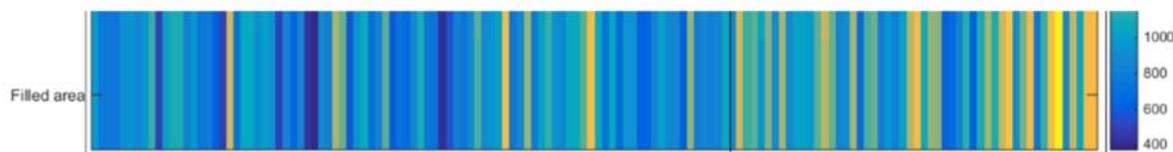


Fig. 5: Heatmaps of the three top-ranking features selected via FINE for dataset  $S_2$ .

junction with PCA, LE, and LLE (see Table V). Interestingly, these features are all part of the Bloom Richardson grading lexicon (nuclear pleomorphism and chromatin patterns) [44], and grade has been shown to be one of the strongest predictors of disease outcome in ER+ breast cancers [45]. Hence, the results from FINE are biologically intuitive.

#### D. Predicting Biochemical Recurrence of Prostate Cancer Based on Tissue Whole Mounts

For dataset  $S_3$ , feature subsets selected based on FINE scores provided high stability, with  $J$  values ranging from 0.64–0.74. In contrast, feature subsets selected based on the  $t$ -test, Fisher score, and Gini index were substantially lower (see Figure 4). The top ranking features selected by  $\text{FINE}_{\text{PCA}}$ ,  $\text{FINE}_{\text{isomap}}$ ,  $\text{FINE}_{\text{LE}}$ , and  $\text{FINE}_{\text{LLE}}$  perform on par with features selected by the  $t$ -test, Fisher score, and Gini index and the embedding vectors themselves. For this dataset the AUC values ranged from 0.52–0.64 depending on the classifier used. These AUC values are similar to AUC values previously reported for this problem. For example, Lee et al. [3] reported AUC values of 0.56–0.67 for predicting biochemical recurrence risk after radical prostatectomy for treating prostate cancer. Only when histomorphometric features were combined with proteomic features obtained via mass spectrometry to predict biochemical recurrence risk following radical prostatectomy were higher AUC values of 0.74–0.93 achieved [11], [46].

The five top-ranking histomorphometric features were similar for  $\text{FINE}_{\text{PCA}}$ ,  $\text{FINE}_{\text{LE}}$ , and  $\text{FINE}_{\text{LLE}}$  and included primarily morphological features: Fourier descriptors, invariant moments, and tensor features (see Table V).  $\text{FINE}_{\text{isomap}}$ , however, returned a very different feature set that included three texture features and two features describing glandular architecture. These features were previously found to be useful for predicting prostate cancer aggressiveness on pathology [47]. Prostate cancer aggressiveness is a part of the Kattan nomogram for predicting the risk of biochemical recurrence of prostate cancer [48], so it is reasonable that these features would be useful for predicting the risk of biochemical recurrence of prostate cancer.

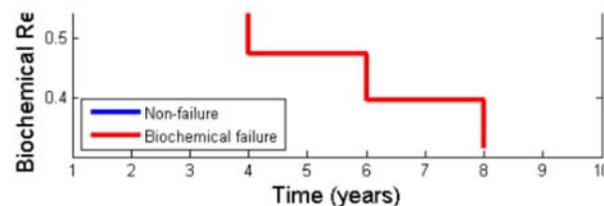


Fig. 6: Kaplan–Meier survival curves ( $p = 0.02$ ) obtained using the five top-ranking features selected by FINE to predict the likelihood of recurrence-free survival in cohort  $S_4$ .

#### E. Predicting Biochemical Recurrence of Prostate Cancer Based on Tissue Microarrays

For dataset  $S_4$ ,  $\text{FINE}_{\text{PCA}}$  and  $\text{FINE}_{\text{LE}}$  are associated with  $J$  ranging between 0.22–0.26 (see Figure 4), although  $J$  associated with feature subsets selected based on the Fisher score were more stable. As was the case with datasets  $S_1$  and  $S_2$ ,  $\text{FINE}_{\text{PCA}}$  and  $\text{FINE}_{\text{LLE}}$  provided more stable feature subsets than  $\text{FINE}_{\text{isomap}}$  and  $\text{FINE}_{\text{LE}}$ .

The top ranking features selected by  $\text{FINE}_{\text{PCA}}$ ,  $\text{FINE}_{\text{isomap}}$ ,  $\text{FINE}_{\text{LE}}$  and  $\text{FINE}_{\text{LLE}}$  provided AUC values ranging from 0.54–0.64, on par with the  $t$ -test, Fisher score, and Gini index (AUC range: 0.57–0.67). The top-ranking histomorphometric features were relatively similar for all DR schemes; they included a combination of Fourier descriptors, invariant moments, and several morphological features computed via a cell graph (see Table V). These features were previously shown to be useful for predicting biochemical recurrence of prostate cancer [49].

In addition to assessing AUC values, Kaplan–Meier survival analysis [50] was also performed for this dataset (see Figure

6) since the time until recurrence was available for all subjects who experienced biochemical recurrence. Kaplan–Meier survival probabilities, calculated as the ratio of the number of subjects who remain event–free to the total number of study subjects, are useful for measuring the fraction of subjects who remain recurrence–free at any given time after treatment. Kaplan–Meier survival curves that stratify patients based upon their risk of developing biochemical recurrence are shown in Figure 6. These survival curves were computed using a logistic regression classifier trained on the five top–ranking histomorphometric features. The difference between the two survival curves is statistically significant ( $p = 0.02$ ), according to the log–rank test.

## VII. CONCLUDING REMARKS

In this paper we present FINE, a novel approach for overcoming the curse of dimensionality and data sparsity. FINE enables performing dimensionality reduction without compromising on the interpretability of the classifier in the reduced feature space. By ranking features based on the strengths of their contributions to classification in a reduced feature subspace obtained via nonlinear dimensionality reduction, FINE can be used for feature selection, much as any other filter method is used for feature selection. However, it is important to note that FINE is very different from other filters. Firstly, FINE is the only filter that ranks features based on their roles in (a) defining the geometry of a non–linearly derived embedding and (b) driving accurate classification. Secondly, the feature subsets provided by FINE tended to be more stable than feature subsets selected using other filter methods. Whereas PCA is relatively robust to small data perturbations, NLDR algorithms minimize the contribution of outliers by basing the geometry of the embedding on locally connected neighborhoods. As a result, the stability of the low–dimensional embeddings tends to translate into more stable feature subset selection. Thirdly, whereas many filters can be used for feature selection, FINE can also be used for feature discovery post facto. That is, once DR has already been done and/or a classifier has already been constructed in a low dimensional embedding space, FINE provides insight into which of the original, high dimensional features actually contributed most prominently to classification in the embedding space.

We note that our study did have its limitations. Firstly, since FINE ranks features based on their contributions to classification on an embedding, it is possible and even probable that highly correlated or redundant features may contribute greatly to classification on an embedding. As a result, features selected based on their FINE scores may be correlated and redundant. This problem can be overcome by the introduction of a redundancy term that penalizes the selection of redundant features. Secondly, we did not leverage FINE to quantify the roles that groups of features derived from digital pathology (e.g., architectural features versus texture features or graph–based features) play in predicting biochemical recurrence [46]. Such an analysis would provide insight into what types of features should be focused on in quantitative histomorphometry. Another possible avenue for future work is to extend

FINE to embeddings obtained by solving a generalized eigenvalue problem (e.g., canonical correlation analysis, weighted extensions of principal component analysis). In spite of these limitations, we have presented a new method for performing DR without compromising on classifier interpretability, and we have demonstrated its role in the context of building robust classifiers for prognosis prediction for an array of digital pathology problems.

## ACKNOWLEDGMENT

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers R01CA136535-01, R01CA140772-01, R21CA167811-01, R21CA179327-01, R21CA195152-01; the National Institute of Diabetes and Digestive and Kidney Diseases under award number R01DK098503-02; the DOD Prostate Cancer Synergistic Idea Development Award (PC120857); the DOD Lung Cancer Idea Development New Investigator Award (LC130463); the DOD Prostate Cancer Idea Development Award; the Ohio Third Frontier Technology development Grant; the CTSC Coulter Annual Pilot Grant; the Case Comprehensive Cancer Center Pilot Grant; VelaSano Grant from the Cleveland Clinic; the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University; and the National Science Foundation Graduate Research Fellowship Program.

## REFERENCES

- [1] M. Guran, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: a review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.
- [2] H. Wang, A. Cruz-Roa, A. Basavanahally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *Journal of Medical Imaging*, vol. 1, 2014.
- [3] G. Lee, R. Sparks, S. Ali, N. Shih, M. Feldman, E. Spangler, T. Rebbeck, J. Tomaszewski, and A. Madabhushi, "Co-occurring gland angularity in localized subgraphs: predicting biochemical recurrence in intermediate-risk prostate cancer patients," *PLoS ONE*, vol. 9, 2014.
- [4] P. Huang and Y. Lai, "Effective segmentation and classification for HCC biopsy images," *Pattern Recognition*, vol. 43, pp. 1550–1563, 2010.
- [5] S. Petushi, F. Garcia, M. Haber, C. Katsinis, and A. Tozeren, "Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer," *BMC Medical Imaging*, vol. 6, pp. 14–24, 2006.
- [6] A. Basavanahally, S. Ganesan, S. Agner, J. Monaco, M. Feldman, J. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lympholymph infiltration in HER2+ breast cancer histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 642–653, 2010.
- [7] M. Donovan, S. Hamman, M. Clayton, F. Khan, M. Sapir, V. Bayer-Zubek, G. Fernandez, R. Mesa-Tejada, M. Teverovskiy, V. Reuter, R. Scardino, and C. Cordon-Cardo, "Systems pathology approach for the prediction of prostate cancer progression after radical prostatectomy," *Journal of Clinical Oncology*, vol. 26, pp. 3923–3929, 2008.
- [8] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. IT-14, pp. 55–63, 1968.
- [9] R. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinfo*, vol. 19, pp. 1484–1491, 2003.
- [10] L. Kanal and B. Chandrasekaran, "On dimensionality and sample size in statistical pattern classification," *Pattern Recognition*, vol. 3, pp. 225–234, 1971.

- [11] A. Golugula, G. Lee, S. Master, M. Feldman, J. Tomaszewski, D. Speicher, and A. Madabhushi, "Supervised regularized canonical correlation analysis: integrating histologic and proteomic measurements for predicting biochemical recurrence following prostate surgery," *BMC Bioinformatics*, vol. 12, p. 483, 2011.
- [12] P. Tiwari, M. Rosen, and A. Madabhushi, "A hierarchical spectral clustering and nonlinear dimensionality reduction scheme for detection of prostate cancer from magnetic resonance spectroscopy (MRS)," *Medical Physics*, vol. 36, pp. 3927–3939, 2009.
- [13] P. Tiwari, S. Viswanath, J. Kurhanewicz, A. Sridhar, and A. Madabhushi, "Multimodal wavelet embedding representation for data combination (MaWERiC): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection," *NMR in Biomedicine*, vol. 25, pp. 607–919, 2012.
- [14] A. Motsinger and M. Ritchie, "Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies," *Human Genomics*, vol. 2, pp. 318–328, 2006.
- [15] G. Lee, C. Rodriguez, and A. Madabhushi, "Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 5, pp. 368–384, 2008.
- [16] R. Sparks and A. Madabhushi, "Statistical shape model for manifold regularization: Gleason grading of prostate histology," *Computer Vision and Image Understanding*, vol. 117, pp. 1138–1146, 2013.
- [17] —, "Explicit shape descriptors: novel morphologic features for histopathology classification," *Medical Image Analysis*, vol. 17, pp. 997–1009, 2013.
- [18] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.
- [19] P. Estevez, M. Tesmer, C. Perez, and J. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, pp. 189–201, 2009.
- [20] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.
- [21] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2001.
- [22] S. Ginsburg, S. Viswanath, B. Bloch, N. Rofsky, E. Genega, R. Lenkinski, and A. Madabhushi, "Novel PCA-VIP scheme for ranking MRI protocols and identifying computer-extracted MRI measurements associated with central gland and peripheral zone prostate tumors," *Journal of Magnetic Resonance Imaging*, vol. Epub ahead of print, 2014.
- [23] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.
- [24] M. Bekin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.
- [25] F. D. la Torre, "A least-squares framework for component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1041–1055, 2012.
- [26] I. Guyon, A. Hur, S. Gunn, and G. Dror, "Result analysis of the NIPS 2003 feature selection challenge," in *Advances in Neural Information Processing Systems 17*, 2004.
- [27] R. Martis, U. Acharya, H. Adeli, H. Prasad, J. Tan, K. Chua, C. Too, S. Yeo, and L. Tong, "Computer aided diagnosis of atrial arrhythmia using dimensionality reduction methods on transform domain representation," *Biomedical Signal Processing and Control*, vol. 13, pp. 295–305, 2014.
- [28] M. Nagarajan, M. Huber, T. Schlossbauer, G. Leinsinger, A. Krol, and A. Wismuller, "Classification of small lesions on dynamic breast MRI: integrating dimension reduction and out-of-sample extension into CADx methodology," *Artificial Intelligence in Medicine*, vol. 60, pp. 65–77, 2014.
- [29] I. Illan, J. Gorriz, J. Ramirez, D. Salas-Gonzalez, M. Lopez, F. Segovia, R. Chaves, M. Gomez-Rio, and C. Puntonet, "18F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis," *Information Sciences*, vol. 181, pp. 903–916, 2011.
- [30] S. Wang, J. Yao, and R. Summers, "Improved classification for computer-aided polyp detection in CT colonography by nonlinear dimensionality reduction," *Medical P*, vol. 35, pp. 1377–1386, 2008.
- [31] M. Masaeli, G. Fung, and J. Dy, "From transformation-based dimensionality reduction to feature selection," in *Proceedings of the 27th International Conference on Machine Learning*, 2010.
- [32] X. Fang, Y. Xu, X. Li, Z. Fan, H. Liu, and Y. Chen, "Locality and similarity preserving embedding for feature selection," *Neurocomputing*, vol. 128, pp. 304–315, 2014.
- [33] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognition*, vol. 48, pp. 10–19, 2015.
- [34] Y. Ren, G. Zhang, G. Yu, and X. Li, "Local and global structure preserving based feature selection," *Neurocomputing*, vol. 89, pp. 147–157, 2012.
- [35] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 619–632, 2013.
- [36] D. Wei, S. Li, and M. Tan, "Graph embedding based feature selection," *Neurocomputing*, vol. 93, pp. 115–125, 2012.
- [37] I. Chong and C. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and Intelligent Laboratory Systems*, vol. 78, pp. 103–112, 2005.
- [38] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.
- [39] J. Ham, D. D. Lee, S. Mika, and B. Scholkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proceedings of the 21st International Conference on Machine Learning*, 2004.
- [40] J. Ross, C. Sheehan, H. Fisher, R. Kaufman, P. Kaur, K. Gray, I. Webb, G. Gray, R. Mosher, and B. Kallakury, "Correlation of primary prostate-specific membrane antigen expression with disease recurrence in prostate cancer," *Clinical Cancer Research*, vol. 9, pp. 6357–6362, 2003.
- [41] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley & Sons, 2001.
- [42] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, pp. 37–50, 1912.
- [43] S. Han, K. Park, B. Bae, K. Kim, H. Kim, Y. Kim, and H. Kim, "E2F1 expression is related with poor survival of lymph node-positive breast cancer patients treated with fluorouracil, doxorubicin, and cyclophosphamide," *Breast Cancer Research and Treatment*, vol. 82, pp. 11–16, 2003.
- [44] H. Bloom and W. Richardson, "Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years," *British Journal of Cancer*, vol. 11, pp. 359–377, 1957.
- [45] E. Rakha, J. Reis-Filho, F. Baehner, D. Dabbs, T. Decker, V. Eusebi, S. Fox, S. Ichihara, J. Jacquemier, S. Lakhani, J. Palacios, A. Richardson, S. Schnitt, F. Schmitt, P. Tan, G. Tse, S. Badve, and I. Ellis, "Breast cancer prognostic classification in the molecular era: the role of histological grade," *Breast Cancer Research*, vol. 12, pp. 207–218, 2010.
- [46] G. Lee, A. Singanamalli, H. Wang, M. Feldman, S. Master, N. Shih, E. Spangler, T. Rebbeck, J. Tomaszewski, and A. Madabhushi, "Supervised multi-view canonical correlation analysis (smvcca): Integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 284–297, 2015.
- [47] A. Tabesh, M. Teverovskiy, H. Pang, V. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and gleason grading of histological images," *IEEE Transactions on Medical Imaging*, vol. 26, pp. 1366–1378, 2007.
- [48] M. Kattan, T. Wheeler, and P. Scardino, "Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer," *Journal of Clinical Oncology*, vol. 17, pp. 1499–1507, 1999.
- [49] G. Lee, S. Ali, R. Veltri, J. Epstein, C. Christudass, and A. Madabhushi, "Cell orientation entropy (COE): Predicting biochemical recurrence from prostate cancer tissue microarrays," in *MICCAI*, 2013.
- [50] J. Rich, J. Neely, R. Paniello, C. Voelker, B. Nussenbaum, and E. Wang, "A practical guide to understanding Kaplan-Meier curves," *Otolaryngology - Head and Neck Surgery*, vol. 143, pp. 331–336, 2010.