PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

# Persistent homology for the automatic classification of prostate cancer aggressiveness in histopathology images

Peter Lawson, Jordan Schupbach, Brittany Terese Fasy, John W. Sheppard

Peter Lawson, Jordan Schupbach, Brittany Terese Fasy, John W. Sheppard, "Persistent homology for the automatic classification of prostate cancer aggressiveness in histopathology images," Proc. SPIE 10956, Medical Imaging 2019: Digital Pathology, 109560G (18 March 2019); doi: 10.1117/12.2513137

SPIE.

Event: SPIE Medical Imaging, 2019, San Diego, California, United States

# Persistent Homology for the Automatic Classification of Prostate Cancer Aggressiveness in Histopathology Images

Peter Lawson[a], Jordan Schupbach[b], Brittany Terese Fasy[b,c], and John W. Sheppard[c]

[a]Department of Biomedical Engineering, Tulane University, 500 Lindy Boggs Center, New Orleans, LA 70118, USA
[b]Department of Mathematical Sciences, Montana State University, 2-214 Wilson Hall, Bozeman, MT 59717, USA
[c]Gianforte School of Computing, Montana State University, 357 Barnard Hall, Bozeman, MT 59717, USA

## ABSTRACT

In this study, we present an automated approach to classify prostate cancer (PCa) whole slide images (WSIs) as high or low cancer aggressiveness using features derived from persistent homology, a tool of topological data analysis (TDA). This extends previous work on the use of these features for representing the characteristics of prostate cancer architecture in region of interest (ROI) images, and demonstrates the value of features derived from persistent homology to predict cancer aggressiveness of WSIs on an ROI basis. We compute persistence on ROI images and summarize persistence as a persistence image. Using this summary we construct a random forest classifier to predict cancer aggressiveness. We demonstrate the potential of persistent homology to capture the architectural differences between low and high grade prostate cancers in a feature representation that lends itself well to machine learning approaches.

## 1. INTRODUCTION

The Gleason grading system is a powerful prognostic predictor of patient outcomes in prostate cancer, and the most powerful predictor in routine clinical use. However, Gleason grading requires the subjective evaluation of architectural patterns present in prostate histopathology whole slide images (WSIs). As a consequence, Gleason scores exhibit high inter-observer variability.[1] An effort to increase the reproducibility of Gleason grading has led to the development of computational approaches, especially deep learning approaches, to enable the automatic grading of WSIs. In this study we present an automated approach to grade WSIs using features derived from the field of topological data analysis (TDA).

With the rise of digital pathology enabled by the increase in the digitization of histopathology WSIs, computer vision approaches, and in particular, deep learning approaches, are becoming the standard for building computer aided diagnostic (CAD) models to automatically identify and grade cancer in histopathology images.[2] Deep learning models, including CNNs, require training on large image datasets, often consisting of millions of images. These images need to be manually annotated by expert pathologists prior to training, a task that is both time-consuming and cost-prohibitive. While these deep learning models may represent the state of art in digital pathology, curation of large, expertly annotated datasets serves as a major impedance to their development and adoption. In addition, adoption of deep learning approaches is limited by their 'black box' nature, which prevents an intuitive mapping of the features learned by the model to features in the histopathology image.[3]

The overall goal of this work is the automatic prediction of cancer aggressiveness of whole slide prostate cancer images using features derived from persistent homology. Previous work has focused on training of CNNs on WSIs with whole slide annotations in order to predict the slide's Gleason score by averaging across ROI level predictions.[4] This work extends that approach with features derived from persistent homology, a field of topological data analysis (TDA), that offer greater intuition by mapping to features in histopathology images.

Further author information: (Send correspondence to P.L.)
P.L.: E-mail: plawson@tulane.edu

Medical Imaging 2019: Digital Pathology, edited by John E. Tomaszewski, Aaron D. Ward,
Proc. of SPIE Vol. 10956, 109560G · © 2019 SPIE · CCC code:
1605-7422/19/$18 · doi: 10.1117/12.2513137

## 2. RELATED WORK

Automatic prostate cancer grading is an area of interest in digital pathology.[5–9] However, few have focused on explicitly rendering a Gleason score or binary classification of cancer aggressiveness for WSIs. Automatic grading can be classified broadly into two approaches: grading with handcrafted features and grading with unsupervised (or learned) features.[3] Handcrafted features are those that correspond to a measurable feature in the histopathology image, such as gland size, shape, and orientation, and offer more interpretability than other approaches, as they often correspond to features pathologists are already trained to identify. Classifiers trained on histomorphometric features, like nuclei and gland shape, are often impeded by the amount of effort necessary by pathologists to hand demarcate features of interest in order to effectively train the segmentation approach. Approaches applying handcrafted features leverage simple classifiers, such as support vector machines (SVM), trained on textural and morphometric features including one such work that achieved 77% accuracy classifying Gleason 3 vs Gleason 4 images.[10]

Conversely, deep learning approaches utilize large histopathology annotated datasets (often millions of ROIs) to learn features in an entirely unsupervised approach. These methods perform well but have been criticized in the pathology community for lacking interpretability, serving as a "black-box" where it is difficult to understand learned features, or map them back to traditional histopathological features. Deep learning approaches, in particular CNNs, have shown excellent performance in automated grading of PCa histopathology images.[4,7,9,11] Some approaches seek to combine the explanatory power of handcrafted features with the predictive power of CNN approaches. One such methodology leverages a hybrid approach involving identifying handcrafted features (shearlet transforms), and applying CNNs to generate learned features from the handcrafted features.[6] This hybrid approach involved training a CNN with both the shearlet transformed images and original RGB images, and outperformed an SVM using shearlet transforms alone in classifying Gleason grade 2 to 5 images (88 % accuracy).[6]

TDA can be viewed as a hybrid approach, a middle ground between traditional handcrafted feature approaches that require an *a priori* understanding to compute morphometric feature, and deep learning approaches that learn representations of features over extremely large training datasets. We have previously demonstrated the power of TDA, and persistent homology specifically, to capture PCa architecure both within and between Gleason grades.[12] TDA has a long history in describing the shape of geometric objects and is progressing in area of automating shape analysis. The analysis of handcrafted features in the context of shape analysis requires the calculation of shape characteristics (e.g., volume, area, eccentricity), which can be computationally challenging, and often requires laborious hand annotation. In the case of prostate cancer images, this might correspond to the volume of individual prostate nuclei, cells, and glands. Extraction of handcrafted features increases in difficulty in prostate cancer images with increasing Gleason grade, as glands become less well delineated. TDA overcomes these challenges by providing an automated means of generating handcrafted shape features. Turner *et al.*[13] recently developed a topological summary descriptor for capturing the shape of objects. Crawford *et al.*[14] extended their method to map to functional inner-product space that is more conducive to statistical analysis. Crawford *et al.* used their method to capture the topological features of *glioblastoma multiforme* tumors for use in predicting disease free and overall survival. Niclau *et al.*[15] used TDA in the discovery of a new type of breast cancer associated with longer survival.

## 3. BACKGROUND

In this paper, we use tools from TDA to construct a classifier for WSI Gleason score prediction using sampled ROI level grayscale images. TDA provides a means of measuring features that are invariant under continuous deformation of an underlying space. We use techniques from TDA to capture architectural information contained in the ROI level images and then use these derived features in machine learning techniques to classify cancer aggressiveness. In particular, we represent ROI level images as a persistence image and using this representation, we construct a random forest classifier.

One tool of TDA is persistent homology (PH), for which we provide necessary background. PH is a way we can encode topological features changing in resolution as a multiset, which can be visualized either by a barcode or diagram.[16,17] These can then be summarized in in a variety of ways that are more useful for constructing

statistical and machine learning models. For instance, persistence landscapes developed by Bubenik[18] and persistence images described by Adams et al.[19] The methods chosen for this paper are with persistence diagrams as summarized by persistence images. We give the required background on how to use these methods to extract topological features from image data.

We begin with a brief introduction to homology and PH, but for a more thorough introduction to the topic we refer the reader to the books on TDA by Edelsbrunner and Harer [20] on algebraic topology by Munkres[21] and on computational homology by Kaczynski et al.[22] Our description of homology assumes some knowledge of abstract algebra, but we refer the unfamiliar reader to the text by Dummit and Foote.[23] We then introduce persistence diagrams,[16] persistence intensity functions[24] and the persistence images[19] derived from them. Last, we give a brief introduction on random forest[25] models, but refer to a book by Hastie and Tibshirani[26] for a more in-depth discussion.

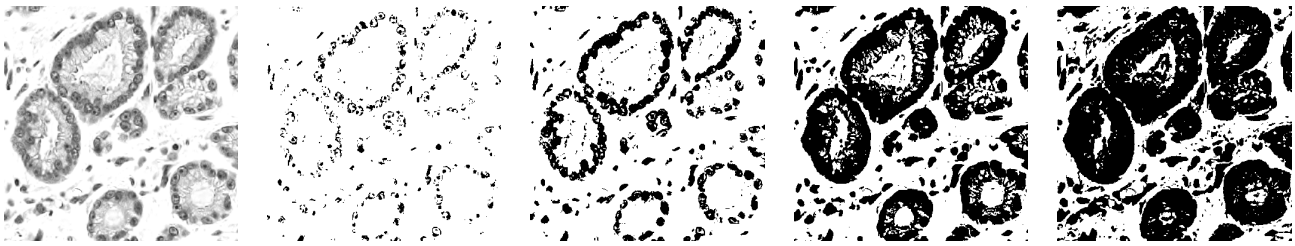
### 3.1 Homology

Homology is a tool from algebraic topology that allows one to characterize a topological space according to how many $d$-dimensional holes are in each dimension $d$. A typical approach for computing homology is to restrict ourselves to homology with coefficients in the integers modulo two, denoted $\mathbb{Z}_2$, and on the space of simplicial complexes. A geometric *$q$-simplex* is the convex hull of $q+1$ affinely independent points and can be intuitively thought of as a $q$-dimensional generalization of a triangle. That is, a zero-simplex is a point, a one-simplex a line, a two simplex a triangle, a three-simplex a tetrahedra, and so on. One way to denote this simplex is to write it as a list of its vertices $\sigma = [u_1, \ldots, u_q]$. Note that the boundary of a $q$-simplex is comprised of $(q+1)$ $(q-1)$-simplices (e.g. the boundary of the one simplex is two points, is three edges for a two-simplex and four triangles for a three-simplex).

A *simplicial complex*, $K$, is a set of simplicies such that every face of the simplicies in $K$ is also in $K$, and the intersection of any two simplicies $\sigma$ and $\tau$ in $K$ is either empty or a face of both. A face $\sigma$ of a simplex $K$ is a subset of the simplex that is also a simplex. Now, given a simplicial complex $K$ of dimension $k$, a *$q$-chain* is the formal sum of simplices in $K$. That is, for the set of $q$ simplices in $K = \{\sigma_1, \ldots, \sigma_q\}$, a $q$-chain can be formed as $c = \sum_{i=1}^{q} a_i \sigma_i$, where $a_i \in \mathbb{Z}_2$. The space of $q$-chains on $K$, denoted $C_q(K)$, is the set whose elements are formal sums of $q$-simplices of $K$. Note that $C_q(K)$ forms a group under addition and also forms a vector space with coefficients in $\mathbb{Z}_2$.

The boundary of a $q$-simplex $\sigma = [u_1, \ldots, u_q]$ of dimension $q$ is the sum of its $(q-1)$-dimensional faces, denoted $\partial_q \sigma = \sum_{j=0}^{q} [u_0, \ldots, \hat{u}_j, \ldots, u_q]$ where the vertex $\hat{u}_j$has been removed. Here, the boundary homomorphism maps a $q$-chain to a $(q-1)$-chain so that we can write $\partial_q : C_q \to C_{q-1}$. A *$q$-cycle* is now a $q$-chain with an empty boundary (i.e. the kernel of the boundary mapping $\partial$, $Ker(\partial_q)$). Because $\partial$ commutes under addition, a $q$-cycle forms a group, which we denote $Z_q(K)$. Note that $Z_q(K)$ is a subgroup of $C_q(K)$, which we denote $Z_q(K) \leq C_q(K)$. A *$q$-boundary* is a $q$-chain that is the boundary of a $(q+1)$-chain (i.e., the image of the boundary map $Im(\partial_{q+1})$. The $q$-boundary similarly forms a group which we denote $B_q$, where we $B_q$ is a normal subgroup of $C_q$, denoted $B_q \trianglelefteq C_q$. Now, the quotient group formed as the $q$-cycle group modulo the $q$-boundary group forms the homology group $H_q(K)$ (i.e. $H_q(K) = Z_q(K)/B_q(K)$) which is also a vector space.

Intuitively, we can think of the rank of $H_q(K)$ as measuring the number of elements in a $q$-dimensional topological subspace that have no boundary and are not a boundary of a $(q+1)$-dimensional subspace. This rank is referred to as the $(q-1)^{\text{st}}$ Betti number. Another way to think about this is that the $(q-1)^{\text{st}}$ Betti number corresponds to the number of generating $q$-dimensional homological features in a topological space. A one-dimensional connected component corresponds to zero-dimensional topological feature, a two-dimensional cycle corresponds to a one-dimensional feature, a three-dimensional void corresponds to a two-dimensional feature, continuing in this fashion for $q$ dimensions. The $(q-1)^{\text{st}}$ Betti number of a complex serves as a measure of $H_q(K)$. In the context of a binary thresholded image, the connected components corresponds to disjoint groups of blacwhite k pixels and a cycle components corresponds to a loop formed with the pixels (e.g. the ring formed by a gland).



(a) (b) (c) (d) (e)

Figure 1: A visual description of a filtration obtained by varying a binary mask on an ROI. The first image (a) is the raw image and images (b)-(e) give the binary masks of the image for several threshold values. As the filtration increases (left to right) new components are born when they appear as new disconnected components and die when they merge into an older component.

### 3.2 Persistent Homology

*Persistent Homology* (PH), provides a robust means of measuring the scale of topological features by tracking changes in homology through a nested sequence of simplicial complexes, known as a *filtration*, that evolves with respect to a size parameter $t \in \mathbb{T}$. For simplicial complex $K$, let $f : K \to \mathbb{T}$ be a monotonic function in the sense that for $a \leq b$, $K_a \subset K_b$ , where $K_b$ is a subcomplex of $K$ and $K_s$ is a subcomplex of $K_b$. By varying our "size" parameter $t \in \mathbb{T}$, we induce an ordered sequence of simplicial complexes $\{K_t\}_{t \in T}$.

For a two-dimensional digital image, we can build a filtration in the following way. We represent each image as a simplicial complex generated through a triangulation over the image pixel intensities. A simplicial complex can be induced using the lower-star filtration, where in the function above we assign a complex to the maximum over all its vertices. With an 8-bit grayscale image, we have the filtration $\{K_r\}_{r \in T} = \{K_0, \ldots, K_{255}\}$. One can visualize what is occurring throughout this filtration as a sequence of images with varying levels of a binary mask being applied, an example of which is given in Figure 1. Notice that as the threshold for the binary mask increases, new connected components are born and eventually are absorbed into other connected components. Similarly, cycles are born and eventually are absorbed by other cycles.

### 3.3 Persistence Diagrams

PDs were first introduced by Edelsbrunner, Letscher, and Zomorodian.[16] A PD summarizes the birth and death times of topological features through a filtration of a topological space. Specifically, given a filtration $f : K \to \mathbb{R}$, a topological feature is born or dies at critical points in $f$, where changes in Betti numbers occur with respect to the scale parameter. For a simplicial complex, this may be when an edge gets connected to a vertex or when a cycle is born. An example of the persistence diagrams for the connected and loop components of a particular grayscale image of PCa tissue is given in Figure 2.

It is important to note that the function $f$ respects what is known as the "elder rule".[20] That is, topological features that have earlier birth times absorb ones that are born later. For example, consider two connected components, one that is born at time $t = 10$ and another that is born at time $t = 20$. Further suppose that these two connected components merge into one connected component at time $t = 30$. This event is represented on the persistence diagram as the birth-death pair $(b, d) = (20, 30)$. The topological feature that is born at $t = 10$ will either persist forever ($t = \infty$) or eventually be merged into a connected component that was born prior to $t = 10$. The lifetime of a particular topological feature is $d - b$ and topological features with longer lifetimes tend to be thought as more important features of the topological space, as short lived topological features can more easily be born from random variation.

Unfortunately, PDs are not amenable to a statistical analysis. For instance, one cannot compute a unique mean for a PD. Typically, PDs are summarized by some summary measure such as persistence landscapes [18] or persistence instensity functions (PIFs). We use the latter in our analysis.

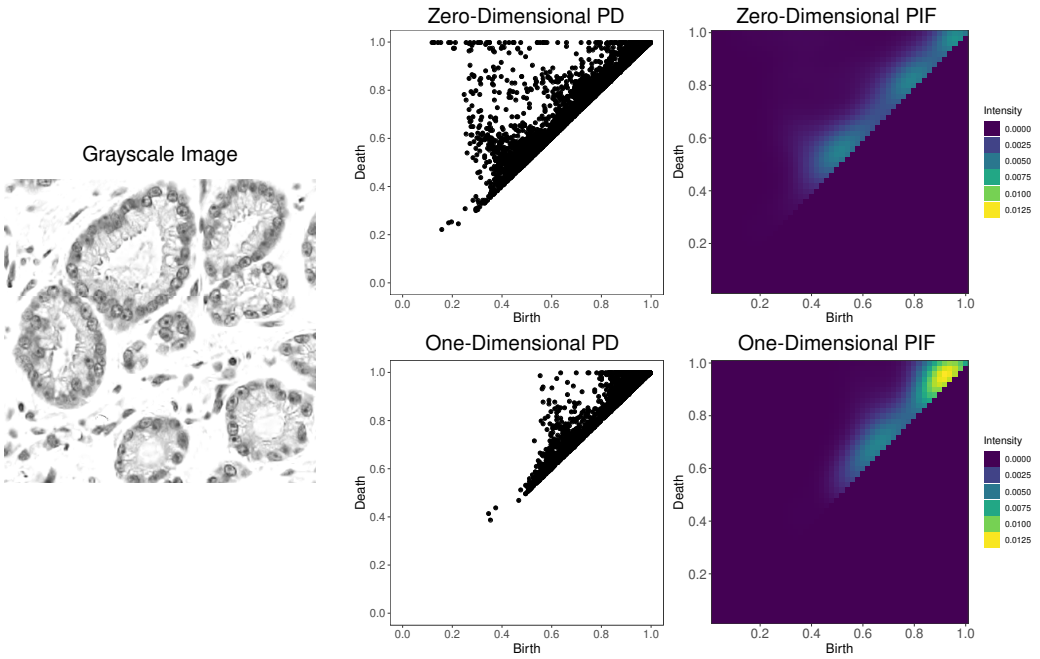



Figure 2: An example of persistence diagrams and persistence intensity functions generated by an ROI for a WSI image labeled as having a Gleason grade of 3.

### 3.4 Persistence Intensity Functions

Persistence intensity functions (PIFs), or persistence surfaces, first described by Edelsbrunner *et al.*[27] and further developed by Chen *et al.*[24] and Adams *et al.*[19] give a representation of PDs as a 2-D function. More specifically, the PIF is a weighted intensity measure of the PD. That is, it is a weighted measure of intensity for the presence of a topological feature for some birth-death combination. The *persistence image* is the pixel evaluation of the PIF over a grid on its domain and represents the space of features used in machine learning approaches. We construct an estimate of the PIF in the same way as given in Chen *et al.*[24] We assume our sampling process induces a spatial point process on the space defined by the PD, which in turn induces a random intensity function.

Following Chen *et al.*, we estimate this intensity function using a weighted 2-D kernel estimate of the density; however, we write it in its more general form. For a given PD, $\mathcal{D} = \{b_i, d_i\}_{i=1}^{K}$, the estimated PIF is defined as a weighted 2-D kernel density estimate

$$\hat{\kappa}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x})|\mathbf{H}|^{-1/2} K_{\mathbf{H}}(|\mathbf{H}|^{-1/2}(\mathbf{x} - \mathbf{x}_i))$$

where $\mathbf{x}$ is a birth-death pair $(b, d)$, $w(\mathbf{x})$ is a weight function, $K_{\mathbf{H}}(\mathbf{x})$ is a 2-D symmetric kernel function, and $\mathbf{H}$ is a $2 \times 2$ positive definite symmetric bandwidth (smoothing) matrix. That is, $\hat{\kappa}_{\mathbf{H}}$ is simply a smoothed version of the persistence measure $\Phi(x, y)$. We give an example of a persistence image computed on an ROI in Figure 2. We can see that the PIF gives a measure of how many points we expect to see in a given region of the persistence diagram.

### 3.5 Random Forests

The random forest, developed by Breiman,[25] is an extension of bagging[28] (Bootstrap AGGregating) classification and regression trees[29] that performs well with little tuning of parameters required.[26] In the case of binary trees, classification and regression trees predict a response by traversing down a decision tree $T$ where each split into child nodes, $t_L$ and $t_R$, corresponds to a split on a covariate in the dataset. Predicted responses for a set of

covariates is the mean value of the root node corresponding to that set of covariates in regression. Alternatively, probability of class membership is estimated by the proportion of each class in the root node for classification. Trees are built in a greedy fashion, where the optimal split $s^*$ is determined at each node $t$ of the tree by maximizing the decrease of node impurity (e.g. Gini index, Shannon entropy, etc.).

Bagging is a general ensemble procedure in which predictors are generated for a large number of bootstrap subsamples of some dataset, and then predictions are averaged across the ensemble to produce a predictor with less variability .[26] In the case of classification, the predicted class is the one that receives the largest number of plurality votes from the ensemble of classifiers. That is, for an ensemble of tree classifiers, $\{T_b(\mathbf{x})\}_{b=1}^{B}$, where each tree $T_b$ in the ensemble is built on a bootstrap subsample of the larger dataset, the bagging estimate of the predicted response would be

$$\hat{\mathbf{y}} = majority\ vote\left\{\{T_b(\mathbf{x})\}_{b=1}^{B}\right\}.$$

Random forests improve on the concept bagged classification and regression trees by taking $m \leq p$ random subsamples of the $p$ predictors when building the trees in the ensemble. This has the effect of reducing the correlation among the trees built using the bootstrap samples. For instance, if a very strong predictor exists in the training set, we might expect that the first split will always use this variable inducing a correlation in the trees in the ensemble. The number of random predictors to use at each split, $m$, is an important tuning parameter for random forests (usually tuned via cross validation).

## 4. METHODS

WSIs, and corresponding pathology reports, were extracted from The Cancer Genome Atlas (TCGA) prostatic adenocarcinoma public dataset[30] . Regions of probable tumor content were extracted into ROIs based on the blue ratio of the WSI. Whole slide Gleason scores were applied to corresponding ROIs and statistical representations of persistence diagrams were computed for each ROI. A series of random forest classifiers were constructed and applied to the statistical representations of persistent homology, varying parameters such as number of trees and feature count per tree. We follow a similar approach to Jimenez-del-Toro *et al.*[7] in the image acquisition and image processing steps of our analysis.

### 4.1 Image Acquisition

All WSIs were acquired from the TCGA[30] database, a publicly available database released by the National Cancer Institute comprising over two petabytes of clinical data. A total of 235 prostate adenocarcinoma (PRAD) WSIs of radical prostatectomies along with the corresponding pathology reports for each case were selected from a subset of 500 cases, matching the patient subset utilized by Jimenez-del-Toro *et al.*[4] This subset represents a balanced dataset of high and low grade cancer WSIs. Each WSI was converted from the proprietary Aperio SVS format to 8-bit PNG for ease of analysis and storage. Optical character recognition (OCR) was performed on each corresponding pathology report. After digitization of pathology reports, each report was mined for the corresponding Gleason score using a series of regular expressions, matching each WSI to a Gleason score.

### 4.2 ROI Extraction

ROIs are extracted from the WSIs through a rejection sampling process. Following the approach of Jimenez-del-Toro *et al.*[4] a binary mask (e.g. see Figure 3) is applied to the WSI by thresholding the blue-ratio ($BR$) image represented by the following equation:

$$BR = \frac{100 \times B}{1 + R + G} \times \frac{256}{1 + B + R + G} \tag{1}$$

where $B$, $R$, and $G$ are the blue, red, and green color channels of an RGB image. The $BR$ seeks to find regions of high pixel intensity with a predominantly blue coloration, indicating the presence of nuclei, as the nuclei specific stain hematoxylin stains a bluish purple color. The first fraction in Equation (1) corresponds to the relative contribution of the blue channel by hue, and the second fraction corresponds to the relative contribution of the blue channel by pixel intensity. Dilation and then erosion morphological operations were then applied to the binary masks with a window size of five pixels to fill in small holes of the binary mask. Downsampling by a factor

(a) Raw Image (b) Blue Ratio as a PDF

Figure 3: **Blue Ratio as a Probability Density Function.** a) A whole-slide histopathology image. b) Visualization of the same image colored by the computed blue ratio: red indicates a high probability (blue ratio greater than 0.55) of cancer tissue, and blue indicates to a low probability of cancer tissue.

of 10 was applied to each WSI in order to reduce memory and processing requirements of generating the binary masks. A $BR$ of 0.55 was found to segment WSIs adequately into regions with high proportions of cancerous tissue. Next, 3,000 ROIs of size $256 \times 256$, comprising three-channel RGB images, were sampled randomly from each WSI by sampling an ROI from the WSI and accepting the sample if the entire ROI falls within the binary mask. In total 705,000 ROIs were generated.

### 4.3 Image Processing

Color deconvolution was performed on each sampled ROI in order to separate the hematoxylin color channel, corresponding to nuclei, in accordance with the method described by Macenko *et al.*[31] using the Matlab[32] Stain Normalization Toolbox developed by the University of Warwick. Then deconvolved ROIs were converted to grayscale using the luminance method as this method has been found to exceed the performance of other methods in feature extraction tasks for texture-based images.[33] In addition, in order enable a comparison of models trained on unprocessed images and deconvolved images, the original ROIs were converted from RGB to grayscale using the previously described luminance method.

### 4.4 Persistent Homology Derived Features

PDs were computed for each color-deconvolved grayscale ROI. For each $256 \times 256$ ROI, a lower-star filtration $\{K_\tau\}_{\tau=0}^{255}$ was constructed from an 8-bit grayscale representation of the cubical complex, and then the associated persistence diagram was constructed. This was done using the R package TDA[34] with interface to the Dionysus C++ library. The diagram output corresponds to two groups, the zero-dimensional homology group, $H_0(\{K_\tau\})$, which is generated by connected components, and the one-dimensional homology group, $H_1\{K_\tau\})$, which is generated by cycles.

To summarize information given in the two persistence diagrams, PIFs were then calculated. PIFs were evaluated on a $50 \times 50$ grid using a standard deviation of $h = 0.2$ for an independent multivariate normal distribution kernel, resulting in $(50)(49) = 2450$ evaluation points over the two diagrams. Note that we only use pixels above the birth = death line, as there is no intensity in the PIF below this line and hence no information. Each evaluation point was then used as a feature in a random forest model.

### 4.5 Random Forest Approach

Following the approach of Jimenez-del-Toro *et al.*,[4] binary classification was applied to the ROIs, divided into two classes of PCa aggressiveness, with low for Gleason 6 and 7 scores (eg. 3+3, 3+4, or 4+3), or high for

Table 1: Model characteristics for random forest model trained on zero, one, and zero and one-dimensional features, for deconvolved and unprocessed ROIs respectively.

| | Unprocessed Images | | | Deconvolved Images | | |
|---|---|---|---|---|---|---|
| Measure | 0 & 1-Dim | 0-Dim | 1-Dim | 0 & 1-Dim | 0-Dim | 1-Dim |
| **ROI AUC** | 0.7489 | .7415 | 0.7198 | 0.7348 | 0.7008 | 0.7264 |
| **ROC Youden's Index** | 0.5373 | 0.5592 | 0.5758 | 0.5571 | 0.5428 | 0.5414 |
| **ROI Accuracy** | 0.6953 | 0.6897 | 0.6719 | 0.6753 | 0.6477 | 0.6696 |
| **ROI F-Score** | 0.6800 | 0.6820 | 0.6692 | 0.6739 | 0.6371 | 0.6532 |
| **WSI-AUC** | 0.8135 | 0.8056 | 0.8000 | 0.8095 | 0.7873 | 0.8175 |
| **WSI Youden's Index** | 0.7122 | 0.7242 | 0.6905 | 0.4270 | 0.5147 | 0.5862 |
| **WSI Accuracy** | 0.7606 | 0.7606 | 0.7606 | 0.7465 | 0.7324 | 0.7606 |
| **WSI F-Score** | 0.7792 | 0.7792 | 0.7848 | 0.6897 | 0.7077 | 0.7606 |
| **OOB Error** | 0.1006 | 0.1092 | 0.1210 | 0.1407 | 0.1544 | 0.1582 |

Gleason scores 8, 9 and 10 (eg. 4+4, 4+5, 5+4 or 5+5), represented as $[0, 1]$ with 0 for low and 1 for high grade cancer. A random forest classifier was constructed, with 500 trees utilized in order to minimize out-of-bag error without adversely impacting computational time. Training and tuning was conducted on the 164 WSI subset of the total 235 with validation being conducted on the remaining 71 WSI cases.

In order to make predictions at the WSI level we implement a multiple step procedure. We first use the average of the probability of class 1 membership (high PCa aggressiveness), given by the random forest classifier trained on the 164 WSI subset, over the ROIs in a WSI to estimate the probability of the WSI belonging to that class. Using these estimated probabilities, we construct a receiver operating characteristic (ROC) curve to tune the threshold for predicted class membership with respect to sensitivity and specificity. We define a threshold as the Youden's J statistic, $J = sensitivity + specificity - 1$, estimated from the ROC area under the curve (AUC). We then assign class membership relative to the defined threshold, with WSIs with an average probability of less than or equal to the threshold assigned to class 0 (low PCa aggressiveness) and WSIs above the threshold assigned to class 1 (high PCa aggressiveness). A ROC curve is then generated for the predicted WSI classes in order to evaluate the performance of the model. 95% confidence intervals were computed for the AUC of all models.

## 5. RESULTS

The highest performing final random forest classifier was trained on one-dimensional features extracted from deconvolved images, with an AUC of 0.8175. This provided only a marginal increase in performance over the random forest classifiers trained unprocessed images, with a maximum AUC of 0.8135 on zero and one-dimensional features combined. The minimal increase in performance gained by combining features of zero-dimensional and one-dimensional homology may indicate that they are capturing redundant features, or that any potential benefit is offset by the "curse of dimensionality". WSI prediction performance across all models was comparable, the details of which are shown in Table. 1. The corresponding ROC curves, with computed confidence intervals, are shown in Fig. 4 and Fig. 5 for the models trained on deconvolved and unprocessed images respectively. The PIFs averages by class are shown in Fig. 6a for the averages by true high and low PCa aggressiveness classes, and in Fig. 6b for the ROI level class predictions of high and low PCa aggressiveness.

Finally variable importance is displayed as a $50 \times 50$ matrix for both one and zero-dimensional random forest models trained on deconvolved ROIs (see Fig. 7). Fig 7a., zero-dimensional variable importance, indicates that zero-dimensional features (connected components) that are born early in the filtration, and persist over the course of the filtration, are the most powerful features in capturing variance between zero-dimensional PIFs (indicated by the high intensity region in the upper left of the zero-dimensional feature importance diagram). The zero-dimensional feature importance diagram indicates that connected components born approximately 20% of the way through the filtration, and persisting midway through the filtration, account for another proportion of

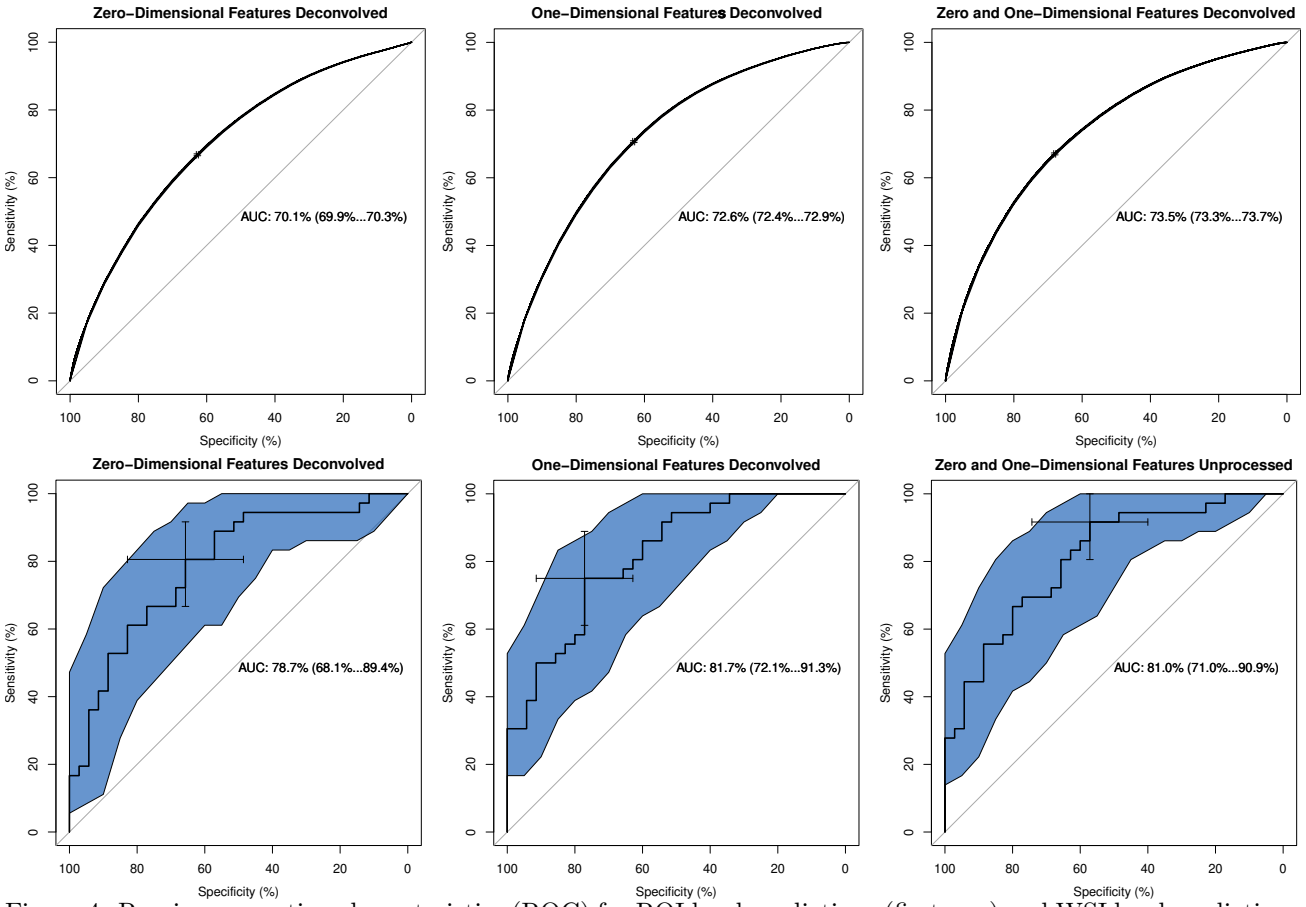



Figure 4: Receiver operating characteristics (ROC) for ROI level predictions (first row) and WSI level predictions (second row). ROC curves are shown for random forest models trained on zero-dimensional features, one-dimensional features, and zero and one-dimensional feature combined. All random forest models were trained on deconvolved images.

features that explain the variability between zero-dimensional PIFs. Fig 7b, one-dimensional variable importance, indicates that one-dimensional features (cycles, potentially corresponding to prostate glands) are focused into three main high intensity regions: features that are born early, and die early, features that are born approximately three-quarters through the filtration, and die approximately 90% through the filtration, and finally features that persist through the filtration (corresponding to the upper right region of the variable importance map). The features that persist midway through the one-dimensional filtration may indicate larger cycles, to include glands, or inter-glandular features, present in the PCa ROIs. Attempting to map feature generators to the original image space is beyond the scope of this work, but may be considered in future work.

## 6. THREATS TO VALIDITY

Although we have achieved positive results in our predictions at the WSI level, the approach is not ideal. Specifically, inheriting labels from the case level (Gleason score from the pathology report) to the ROI level, and using the learned ROI labels to predict back to the WSI level has the potential for introducing model bias. Tissue architecture on each WSI is comprised of a relative contribution of at least two Gleason grades, in varying proportion, as well as benign tissue, and potentially cancer mimickers. Thus, inheriting global labels from the WSI level and applying them to local, heterogeneous regions at the ROI level may bias prediction. However, without validating on another dataset, it is hard to assess the degree to which this is a problem. To mitigate this issue, we would need a large number of hand annotated ROIs to in order to build a robust classifier at the

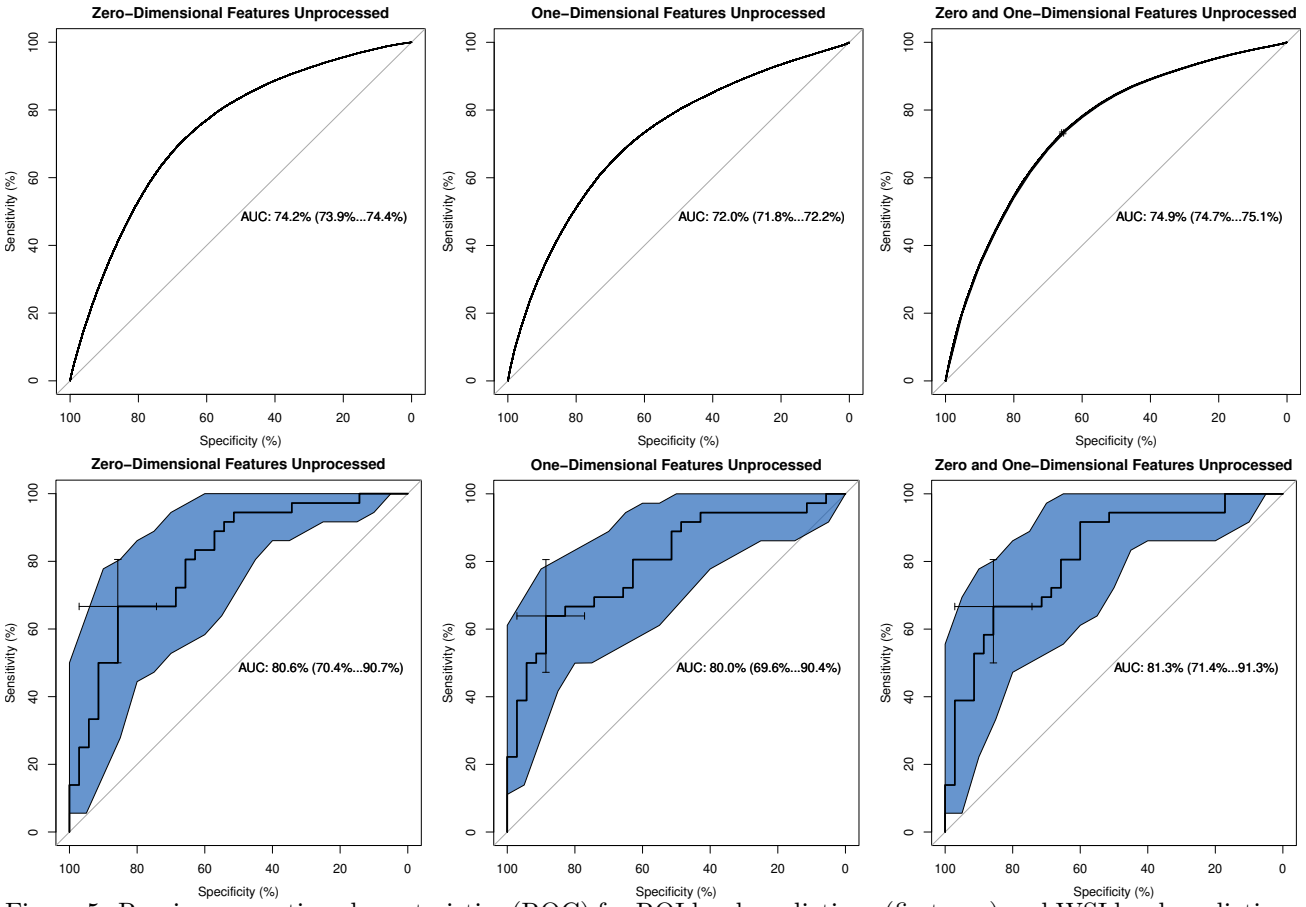


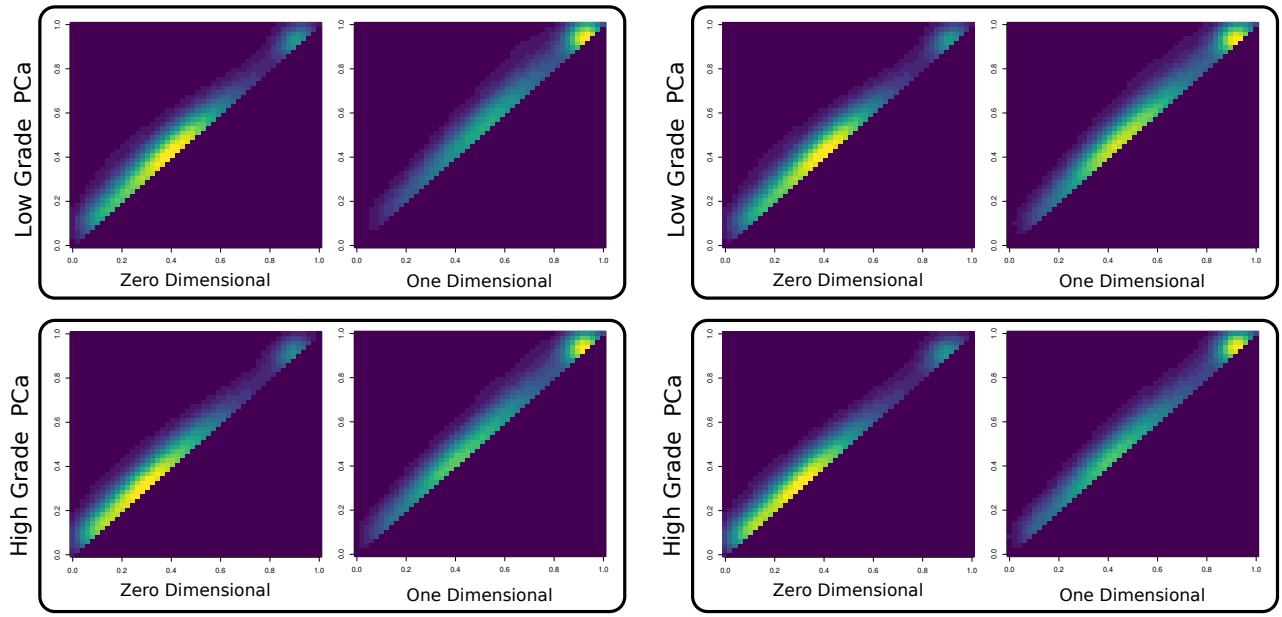

Figure 5: Receiver operating characteristics (ROC) for ROI level predictions (first row) and WSI level predictions (second row). ROC curves are shown for random forest models trained on zero-dimensional features, one-dimensional features, and zero and one-dimensional feature combined. All random forest models were trained on unprocessed images.

ROI level. However, we currently only have a small number of such labeled ROIs. There are other approaches to dealing with a sparse set of labeled examples at the ROI level, which we propose as future work.

There are added concerns with mapping a mixture of Gleason scores into the two classes of high vs low grade PCa. The clinical relevance in the division of Gleason scores into two classes, high and low, is not sufficiently demonstrated. The dividing between classes between Gleason score 7 (low) and Gleason score 8 (high) may not be suited to evaluating prognostic outcomes. In particular, clear prognostic differences between Gleason 3+4 and 4+3 have been demonstrated, with 4+3 PCa associated with a three-fold increase in lethality with respect to 3+4 PCa.[35] An optimal approach would involve classifying WSIs at a higher granularity than a binary classification scheme. We, however, limit ourselves to the binomial case as this work is only an initial investigation into the efficacy of using features derived from TDA at the ROI level for the representation of PCa WSIs. Future work will involve prediction beyond a binary class using WSI Gleason grades or Grade groups.

## 7. FUTURE WORK

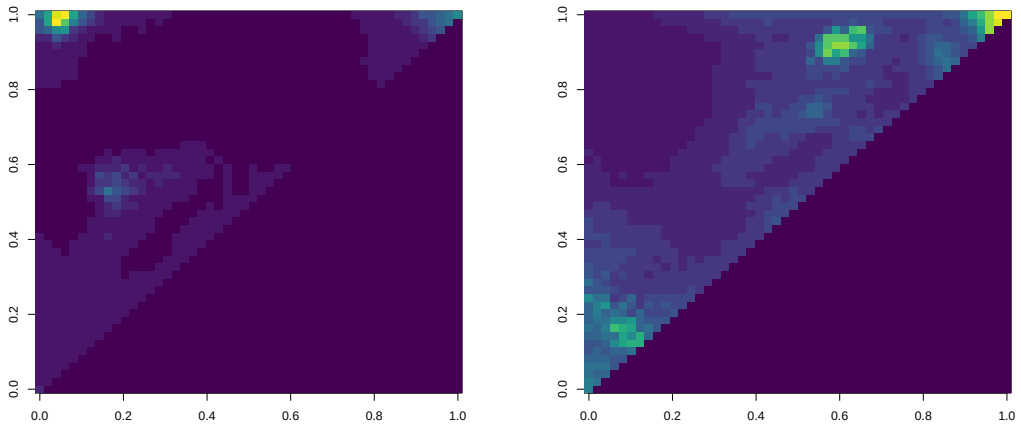
The results of this work are promising, having successfully achieved accuracy comparable to CNNs applied to the data generated in a similar manner.[4] This indicates that topological features may offer a more explainable alternative to texture based CNN features and points to the possibility of building multi-modal approaches that combine the advantages of CNNs and PH derived features. A natural extension of this future work is the




(a) True PCa high and low aggressiveness classes

(b) ROI level predicted PCa high and low aggressiveness classes

Figure 6: Average PIFs for the (a) true high and low PCa aggressiveness classes and (b) average PIFs for the ROI level predicted high and low PCa aggressiveness. Average PIFs are shown for zero and one-dimensional homology features.



(a) Variable importance zero-dimensional

(b) Variable importance one-dimensional

Figure 7: Variable importance for random forest trained on zero and one-dimensional features, respectively.

exploration of the extent CNNs are capturing similar textural and architectural features to persistent homology based features, or whether these features are unique.

While we have demonstrated one method of automatically categorizing ROIs PCa aggressiveness, given a paucity or absence of ROI level annotations, other approaches exist for developing models given sparsely labeled ROI images datasets, namely using semi-supervised [36] and transfer learning [37] techniques. In semi-supervised learning, we can use a small number of labeled ROI level images to predict labels on a much larger number of unlabled ROI level images. Transfer learning techniques would involve taking labeled ROIs from a similar area

and using them to aid in learning the relationship between topological features and a relatively smaller number of labeled ROIs.

Finally, training only high-probability cancer regions indicated by a high $BR$ inherently biases perfomance of trained classifiers towards high-grade cancer. Future work will focus on integrating benign tissue, as well as PCa cancer mimickers such as benign prostatic hyperplasia using an inverse binary mask generated from the $BR$, augmented with automated segmentation approaches in order to ensure a higher quality training dataset. Future work will explore these approaches for automated WSI scoring.

## 8. SUMMARY

The results of this work are promising, having successfully achieved a relatively high performing classification of high vs low-grade cancer in WSIs. This suggests value in the representation of shape based features in image data using methods from TDA. Using only these types of features, our top classifier achieved an AUC of 0.82 for predicting high vs low grade annotated WSIs. In addition the performance of this classifier was largely unaffected by color deconvolution applied to the input histopathology image, indicating that additional image processing may be unnecessary to maximizing performance, thereby decreasing the amount of pre-processing necessary. This indicates that TDA provides a robust means of generating shaped based features that perform well in decision tree based classifiers, and while this work explored the automated grading of WSIs, TDA may find success in ROI level prediction, either in lieu of, or ensembled with, CNNs.

## 9. ACKNOWLEDGEMENTS

Research was supported by the National Science Foundation and National Institutes of Health under grants NIH/NSF DMS-1664848 and 1557750 (Tulane), NIH/NSF DMS-1664858 and 1557716 (MSU), and the National Cancer Institute of the National Institutes of Health under NIH NCI R33CA196457 (Tulane). P.L. was partially supported by NSF IGERT funding (DGE-1144646) through Tulane's Interdisciplinary PhD Program in Bioinnovation.

## 10. AUTHOR CONTRIBUTIONS

P. Lawson and J. Schupbach acquired and processed the data, performed the experiments, analyzed the result, and wrote the manuscript. J. Sheppard edited the manuscript and assisted in defining the experiments and analyzing the result. B.T. Fasy assisted in writing the introduction to TDA, and edited the manuscript.

## REFERENCES

[1] Engers, R., "Reproducibility and reliability of tumor grading in urological neoplasms," *World Journal of Urology* **25**, 595–605 (Dec. 2007).

[2] Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-van de Kaa, C., Bult, P., van Ginneken, B., and van der Laak, J., "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific Reports* **6**, 26286 (2016).

[3] Madabhushi, A. and Lee, G., "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical Image Analysis* **33**, 170–175 (Oct. 2016).

[4] Jiménez del Toro, O., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., and Müller, H., "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score," in [*Proc. SPIE Medical Imaging: 2017*], Gurcan, M. N. and Tomaszewski, J. E., eds., 101400O (Mar. 2017).

[5] Niazi, M. K. K., Yao, K., Zynger, D., Clinton, S., Chen, J., Koyuturk, M., LaFramboise, T., and Gurcan, M., "Visually Meaningful Histopathological Features for Automatic Grading of Prostate Cancer," *IEEE Journal of Biomedical and Health Informatics* **PP**(99), 1–1 (2016).

[6] Rezaeilouyeh, H., Mollahosseini, A., and Mahoor, M. H., "Microscopic medical image classification framework via deep learning and shearlet transform," *Journal of Medical Imaging* **3**, 044501 (Nov. 2016).

[7] Gummeson, A., Arvidsson, I., Ohlsson, M., Overgaard, N. C., Krzyzanowska, A., Heyden, A., Bjartell, A., and Aström, K., "Automatic Gleason grading of HE stained microscopic prostate images using deep convolutional neural networks," 101400S (Mar. 2017).

[8] Huang, C.-H. and Racoceanu, D., "Automated high-grade prostate cancer detection and ranking on whole slide images," 101400A (Mar. 2017).

[9] Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Rüschoff, J. H., and Claassen, M., "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific Reports* **8** (Dec. 2018).

[10] Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., and Tomaszeweski, J., "Automated Grading of Prostate Cancer Using Architectural and Textural Image Features," in [*2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*], 1284–1287, IEEE, Arlington, VA, USA (2007).

[11] Nagpal, K., Foote, D., Liu, Y., Po-Hsuan, Chen, Wulczyn, E., Tan, F., Olson, N., Smith, J. L., Mohtashamian, A., Wren, J. H., Corrado, G. S., MacDonald, R., Peng, L. H., Amin, M. B., Evans, A. J., Sangoi, A. R., Mermel, C. H., Hipp, J. D., and Stumpe, M. C., "Development and Validation of a Deep Learning Algorithm for Improving Gleason Scoring of Prostate Cancer," *arXiv:1811.06497 [cs]* (Nov. 2018). arXiv: 1811.06497.

[12] Lawson, P., Sholl, A. B., Brown, J. Q., Fasy, B. T., and Wenk, C., "Persistent Homology for the Quantitative Evaluation of Architectural Features in Prostate Cancer Histology," *Scientific Reports* **9**, 1139 (Feb. 2019).

[13] Turner, K., Mukherjee, S., and Boyer, D. M., "Persistent homology transform for modeling shapes and surfaces," *Information and Inference: A Journal of the IMA* **3**(4), 310–344 (2014).

[14] Crawford, L., Monod, A., Chen, A. X., Mukherjee, S., and Rabadán, R., "Functional data analysis using a topological summary statistic: The smooth euler characteristic transform," *arXiv preprint arXiv:1611.06818* (2016).

[15] Nicolau, M., Levine, A. J., and Carlsson, G., "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival," *Proceedings of the National Academy of Sciences* , 201102826 (2011).

[16] Edelsbrunner, H., Letscher, D., and Zomorodian, A., "Topological persistence and simplification," *Discrete & Computational Geometry* **28**, 511–533 (Nov 2002).

[17] Zomorodian, A. and Carlsson, G., "Computing persistent homology," *Discrete & Computational Geometry* **33**(2), 249–274 (2005).

[18] Bubenik, P., "Statistical topological data analysis using persistence landscapes," *arXiv:1207.6437 [cs, math, stat]* (July 2012).

[19] Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L., "Persistence Images: A Stable Vector Representation of Persistent Homology," *Machine Learning* , 35 (2017).

[20] Edelsbrunner, H. and Harer, J., [*Computational Topology: An Introduction*].

[21] Munkres, J. R., [*Elements of algebraic topology*], CRC Press (2018).

[22] Kaczynski, T., Mischaikow, K., and Mrozek, M., [*Computational homology*], vol. 157, Springer Science & Business Media (2006).

[23] Dummit, D. S. and Foote, R. M., [*Abstract algebra*], vol. 3, Wiley Hoboken (2004).

[24] Chen, Y.-C., Wang, D., Rinaldo, A., and Wasserman, L., "Statistical Analysis of Persistence Intensity Functions," *arXiv:1510.02502 [stat]* (Oct. 2015).

[25] Breiman, L., "Random Forests," *Machine Learning* **45**, 5–32 (Oct. 2001).

[26] Hastie, T., Tibshirani, R., and Friedman, J. H., [*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*], Springer series in statistics, Springer, New York, NY, 2nd ed ed. (2009).

[27] Edelsbrunner, H., Ivanov, A., and Karasev, R., "Open Problems in Discrete and Computational Geometry," *Modeling and Analysis of Information Systems* , 13 (2000).

[28] Breiman, L., "Bagging predictors," *Machine Learning* **24**, 123–140 (Aug. 1996).

[29] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A., [*Classification and Regression Trees (The Wadsworth Statistics/Probability Series)*], Wadsworth Publishing (jun 1983).

[30] Network, C. G. A. R. et al., "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature* **455**(7216), 1061 (2008).

[31] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Xiaojun Guan, Schmitt, C., and Thomas, N. E., "A method for normalizing histology slides for quantitative analysis," in [*2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*], 1107–1110, IEEE, Boston, MA, USA (June 2009).

[32] MATLAB, [*9.5.0.942161 (R2018b)*], The MathWorks Inc., Natick, Massachusetts (2018).

[33] Kanan, C. and Cottrell, G. W., "Color-to-Grayscale: Does the Method Matter in Image Recognition?," *PLoS ONE* **7**, e29740 (Jan. 2012).

[34] Fasy, B. T., Kim, J., Lecci, F., and Maria, C., "Introduction to the R package TDA," *arXiv preprint arXiv:1411.1830* (2014).

[35] Stark, J. R., Perner, S., Stampfer, M. J., Sinnott, J. A., Finn, S., Eisenstein, A. S., Ma, J., Fiorentino, M., Kurth, T., Loda, M., Giovannucci, E. L., Rubin, M. A., and Mucci, L. A., "Gleason Score and Lethal Prostate Cancer: Does $3 + 4 = 4 + 3$?," *Journal of Clinical Oncology* **27**, 3459–3464 (July 2009).

[36] Chapelle, O., Schölkopf, B., and Zien, A., eds., [*Semi-Supervised Learning*], Adaptive Computation and Machine Learning, MIT Press, Cambridge, Mass (2006).

[37] Torrey, L. and Shavlik, J., "Transfer learning," in [*Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*], Soria, E., Martin, J., Martinez, R. M. M., and Serrano, A., eds., ch. 11, 242–264, IGI Global (2010).