VECTORIZED PERSISTENT HOMOLOGY REPRESENTATIONS FOR CHARACTERIZING GLANDULAR ARCHITECTURE IN HISTOLOGY IMAGES

Deepak Roy Chittajallu [⋆], *Neal Siekierski*[⋆], *Sanghoon Lee*[†], *Samuel Gerber*[⋆], *Jonathan Beezley*[⋆], David Manthey[⋆], David Gutman[†] and Lee Cooper[†]

[⋆] Kitware Inc., Carrboro, NC, USA [†] Emory University, Atlanta, GA, USA

ABSTRACT

Characterizing glandular architecture in histology images of adenocarcinomas is a fundamental problem in digital pathology, with important implications for computer-assisted diagnosis and grading. In this paper, we present a new set of features for encoding the glandular epithelium architecture based on two recently developed vectorized persistent homology representations called *persistence images* and *persistence landscapes* and demonstrate their application to colorectal cancer diagnosis. On the MICCAI 2015 Gland Segmentation Challenge Contest dataset with 165 images (85 training, 80 test images), we obtained a benign vs malignant classification accuracy of 85% and 83% using persistence image and persistence landscape based features, respectively.

Index Terms— Histopathology, Cancer Grading, Persistent homology, Persistence images, Persistence landscapes, Machine learning, Computer aided diagnosis

1. INTRODUCTION

Histopathology is the study of the presence, extent, and progression of a disease through microscopic examination of thin sections of biopsied tissue that are chemically processed and fixed onto glass slides and dyed with one or more stains to highlight different cellular/tissue components (e.g. cell nuclei or membranes) and antigens/proteins (e.g. Ki-67 indicating cell proliferation) of interest. It is regarded as the gold standard in clinical diagnosis and grading of several diseases including most types of cancer.

In clinical practice, histologic evaluation still largely depends on manual assessment of glass slides by a pathologist with a microscope, although improvements in whole-slide imaging devices and subsequent regulatory approval of whole-slide-imaging and computational algorithms are rapidly paving the way for increased clinical use of digital imaging and computational interpretation. Algorithmic evaluation of tissue specimens may eventually improve the efficiency, objectivity, reproducibility, and accuracy of the diagnostic process.

Pathologists integrate information across scales from subcellular to macro when evaluating histology. For adenocarcinomas, lesions that originate in the epithelium of glandular structures including lung, prostate, pancreatic, and colorectal cancers, the architecture of glandular structures conveys significant information about the presence and degree of malignancy. Normal appearing structures with organized epithelium become disorganized with the unchecked growth and aberrant signaling in cancer (Figure 1). There have been several efforts to develop quantitative features for characterizing glandular structures for computer-aided grading [1, 2, 3, 4].

In this paper, we present a new set of features for encoding the glandular epithelium architecture based on two recently developed vectorized persistent homology representations called *persistence images* [5] and *persistence landscapes* [6] and demonstrate their application to colorectal cancer diagnosis. To the best of our knowledge, this is the first application of these representations for cancer diagnosis.

2. BACKGROUND

In this section, we present a brief background on persistence homology. Given a dataset in the form of a point cloud (e.g. set of nuclei centroids in histology images), persistent homology can be seen as a theoretical tool to detect and characterize prominent topological features (e.g. connected components, loops, voids) at multiple scales [7]. These topological descriptors can then be used as features for building machine learning models to solve predictive problems.

Simplicial homology: The foundational concepts of persistent homology are simplices, simplicial complexes, filtration, and homology groups. A p-simplex σ_p is defined as the convex hull of $p+1$ affinely independent points/vertices. For example, a single vertex is a 0-simplex, an edge is a 1-simplex, a triangle is a 2-simplex, a tetrahedron is a 3-simplex, and so on. A face of a p-simplex is defined as a subset of its $p+1$ points/vertices. For example, the tetrahedron which is a 3-simplex has 4 triangular faces, 6 edge faces, and 4 vertex faces each of which are simplices themselves. A simplicial complex K is a finite collection of simplices subject to two conditions: (i) if a simplex σ is in K then any face of σ is also in K, and (ii) if two simplices σ and σ' are in K then $\sigma \cap \sigma'$ must either be empty or a face of both σ and σ' i.e. they must either be glued together along whole faces or be separate. Given a simplicial complex K, a simplicial complex L formed by a subset of its simplices is re-

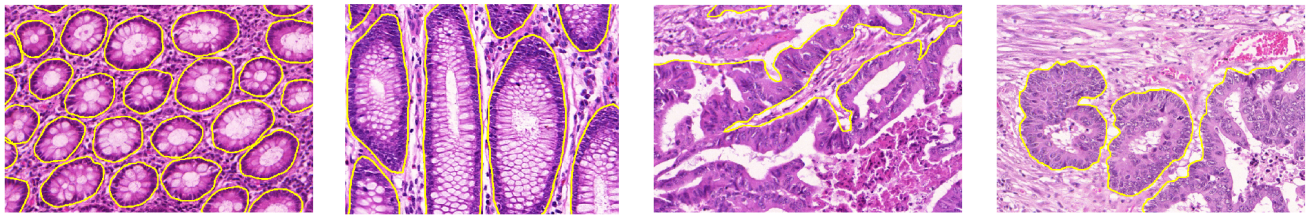
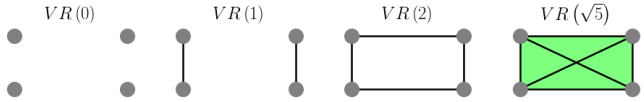


Fig. 1: Sample benign (first-two) and malignant (last-two) images overlaid (yellow) with manually annotated gland boundaries.

ferred to as the sub-complex of K denoted symbolically as $L \subset K$. A nested sequence of simplicial sub-complexes $K_0 = \phi \subset K_1 \subset K_2 \subset ... \subset K_n = K$ that ascends from an empty set all the way up to K is called a filtration of K denoted as $\mathbb{F}(K)$. A d-dimensional homology group $H_d(K)$ of a simplicial complex K is the set of all d-dimensional void in it. For example, the 0-dimensional homology group $H_0(K)$ is the set of all connected components, the 1-dimensional homology group $H_1(K)$ is the set of all 2D loops, the 2-dimensional homology group $H_2(K)$ is the set of all 3D cavities and so on. The rank or the number of voids in a d-dimensional homology group $H_d(K)$ is referred to as the d-dimensional Betti number $\beta_d(K)$.

Vectoris-Rips filtration: Given a dataset in the form of a point cloud of n points, how can we derive a simplicial complex that encodes the underlying topological structure? One approach for generating it is to examine all subsets of $p+1$ points, and add the p-simplex made up of those points to the simplical complex if the distance between all pairs of points in the simplex is less than a present distance ϵ. Such a complex is called a Vectoris-Rips complex of diameter ϵ which we will henceforth denote as $VR(\epsilon)$. Note that, if a simplex is in $VR(\epsilon)$, then all of its faces are also in $VR(\epsilon)$. The schematic below shows the Vectoris-rips complex for different diameter values for a dataset of four points corresponding to the corners of a rectangle with a width of 2 and a height of 1.



A natural question that arises now is how to choose the best diameter ϵ for a given dataset. Persistent homology examines all diameters within a range of interest to see how the system of voids change and provides a topological characterization at multiple scales. An increasing sequence of diameters/scales $\epsilon_1 < \epsilon_2 < ... < \epsilon_n$ results in a nested sequence of Vectoris-Rips simplicial complexes $VR(\epsilon_1) \subset VR(\epsilon_2) \subset ... \subset VR(\epsilon_n)$ referred to as Vectoris-Rips filtration.

Persistence diagram (PD) representation: Given a filtration $\mathbb{F}$ through a series of n scales, the idea of persistent homology is to track the scales at which each void appears and disappears. This information can be summarized in the form of a multi-set $BD_d(\mathbb{F}) = \{(b_i, d_i) \mid b_i, d_i \in \{1, 2, ..., n\} \wedge b_i < d_i\}$ of 2D points representing the birth-death scales of each d-dimensional void i in the filtration. Considering these pairs as points in $\mathbb{R}^2$ we obtain the persistence diagram representation and considering them as birth-death intervals $[b_i, d_i]$ we obtain the barcode representation. This summarization of topological information as a multi-set of points is not amicable for statistics and machine learning, wherein a finite-dimensional vector representation is more convenient. Persistence landscapes [5] and persistence images [6], described below, are two recently developed vectorized persistent homology representations to address this problem.

Persistence landscape (PL) representation: Given a birth-death pair (b, d), let $f_{(b,d)} : \mathbb{R} \to [0, \infty]$ be a piece-wise linear triangle shaped function defined as follows:

$$f_{(b,d)}(x) = \begin{cases} 0 & if \quad x \notin (b, d) \\ x - b & if \quad x \in \left(b, \frac{b+d}{2}\right] \\ d - x & if \quad x \in \left(\frac{b+d}{2}, d\right) \end{cases} \tag{1}$$

Given a multi-set of m birth-death points $\{(b_i, d_i)\}_{i=1}^m$ from a PD, persistence landscape is defined as a 2D function $\lambda : \mathbb{N} \times \mathbb{R} \to [0, \infty]$ where $\lambda(k, x)$ is equal to the k-th largest value of $\{f_{(b_i,d_i)}(x)\}_{i=1}^m$ if $k \leq m$ and zero otherwise. This function can be discretized over a grid to obtain a finite-dimensional vector representation for machine learning.

Persistence image (PI) representation: Given a multi-set of m birth-death points $BD = \{(b_i, d_i)\}_{i=1}^m$ from a PD, a linear transform $T(b, d) = (b, d - b)$ is applied to obtain a multi-set of birth-persistence points $BP = \{(b_i, p_i)\}_{i=1}^m$. Based on this, a 2D real-valued function $\rho_{bp} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ called a persistence surface is defined as weighted sum of isotropic bi-variate gaussian/normal probability density functions $\mathcal{N}\left(x, y; \sigma^2\right)$ with variance σ^2 centered at each of the birth-persistence pairs as follows:

$$\rho_{bp}(x, y) = \sum_{i=1}^{m} w(b_i, p_i) * \mathcal{N}\left(x - b_i, y - p_i; \sigma^2\right) \tag{2}$$

where $w(b, p)$ is a weighting function critical to the stability of the persistence surface. A natural choice is to pick a weighting function that assigns higher weights to points with higher persistence values. However, in certain applications, points of medium persistence may be more important. Hence, the weighting function is defined more generally as a piecewise linear function as follows:

$$w(b, p; c) = \begin{cases} 0 & if\ p \leq 0 \\ p/c & if\ p \leq c \\ 1 & otherwise \end{cases} \tag{3}$$

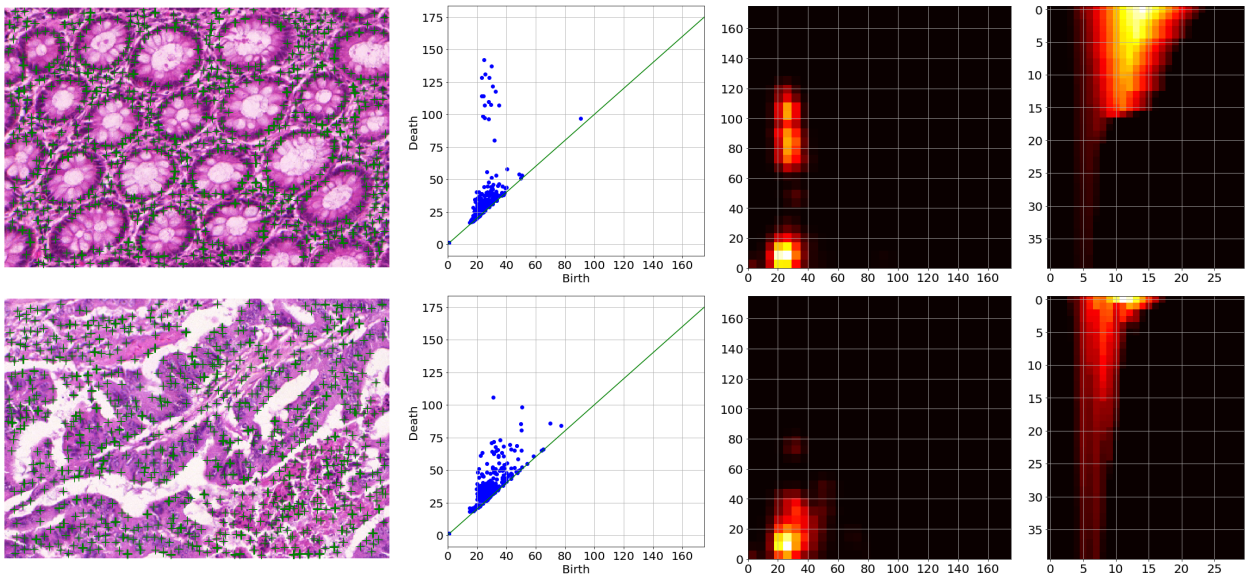


Fig. 2: Illustration of persistent homology representations for sample benign (top-row) and malignant (bottom-row) images: (column-1) Input image overlaid with nuclei centroids, (column-2) persistence diagram of the Vectoris-rips filtration of the nuclei centroid point cloud, (column-3) Persistence image computed with $c = 175$ and discretized onto a 30x30 grid resulting in 900 features, and (column-4) Peristence landscape representation discretized onto a 40x30 grid resulting in 1200 features.

where c is the maximum persistence value of all important topological features. Lastly, the persistence image is generated by defining a discrete grid in the domain of the 2D persistence surface function $\rho_{bp}(x, y)$ and computing its integral in each grid box. In case the birth values of all the points is zero, as is the case with 0-dimensional homology group H_0 of connected components, then both persistence surface and persistence image can be represented compactly in 1D.

3. METHOD

In this section, we present our approach for using persistence image and persistence landscape representations to characterize the glandular epithelium architecture and train a machine learning model for cancer diagnosis.

Detecting nuclei centroids: Given a histology image, we first pre-process it using the color normalization method of Reinhard *et al*. [8]. Next, we use the unsupervised color deconvolution method of Macenko *et al*. [9] to extract the nuclear stain and minimum cross entropy thresholding [10] to segment the nuclear foreground. Lastly, we use a fast Difference-of-Gaussian implementation of the scale-adaptive Laplacian-of-Gaussian filter of Al-Kofahi *et al*. [11] to detect nuclei centroids. This pipeline was implemented using an open-source toolkit called HistomicsTK [1].

Extracting topological features using persistent homology: Considering the set of nuclear centroids as a point cloud, we compute the persistence diagram of its Vectoris-Rips filtration for the homological dimension-1 corresponding to 2D loops using a fast multiscale approach [2]. We then compute the persistence landscape and persistence image representations, described in Section 2 and use them as features characterizing the 2D voids/loops formed by glandular epithelial cell nuclei.

Training machine learning model for classification: Given a training set of images with benign/malignant labels, we train a random forest classifier based on the topological features. We use principal component analysis (PCA) to reduce the dimensionality of each of the feature groups such that 99% of the variance is preserved. We optimize the hyper-parameters of the classification model via cross-validation using a sequential model-based optimization technique called tree-structured parzen estimator. We used the open-source python toolkits scikit-learn and hyperopt for machine learning and hyper-parameter optimization, respectively.

4. RESULTS

We used the MICCAI 2015 Gland Segmentation Challenge Contest dataset [12] to evaluate the proposed method. This dataset contains 165 images derived from 16 hematoxylin-eosin stained histological sections of normal, and stage T3 and stage T4 colorectal adenocarcinomas digitized using a Zeiss MIRAX MIDI SlideScanner with a pixel resolution of $0.620\mu m$ equivalent to a 20x objective magnification. An expert pathologist delineated the boundary of all the glands in each image and graded it as either $benign$ or $malignant$ based on the overall glandular architecture. The dataset was divided by the challenge organizers into two independent parts: a training set of 85 images (37 benign, 48 malignant) and a test set of 80 images (37 benign, 43 malignant).

We used the approach described in Section 3 to train a random forest model to distinguish between benign and malignant images on the training set of 85 images and validated it on the test set of 80 images. Figure 2 shows the persis-

[1] https://github.com/DigitalSlideArchive/HistomicsTK

[2] https://bitbucket.org/suppechasper/homology

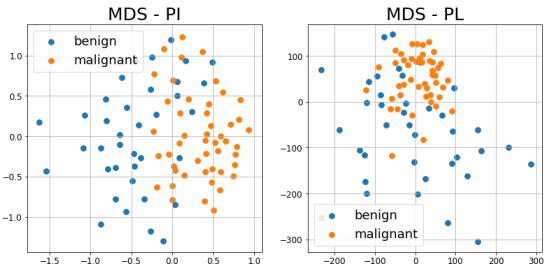


Fig. 3: Visualization of persistence image (left) and landscape (right) features of training samples projected to 2D using a dimensionality reduction technique called multidimensional scaling (MDS) and color coded by class.

tence image and persistence landscape feature representations derived from the nuclei centroid point cloud for sample benign and malignant image from the training set. Figure 3 shows a visualization of 2D MDS-projections of the persistence image and persistence landscape feature representations of all the training samples color coded by grade. Notice that the benign and malignant classes form separate clusters for both the feature sets. Table 1 reports the benign vs malignant classification performance on the test set for different feature sets: (i) persistence image (PI) features only, (ii) persistence landscape (PL) features only, (iii) both persistence image and landscape (PI + PL) features, and (iv) state-of-the-art cell graph features proposed by Doyle *et al*. [2].

Features	**Acc**	**AUC**	**Precision**	**Recall**
PI	0.85	0.85	0.78	0.95
PL	0.83	0.84	0.77	0.92
PI + PL	0.85	0.85	0.78	0.95
Cell graph properties	0.81	0.81	0.75	0.89

Table 1: Benign vs Malignant classification performance

5. CONCLUSION

Vectorized representations of persistence homology can encode common topological patterns observed in histology images. For tissues with glandular structures, topological features can encode important information about the epithelium that undergoes changes with malignant transformation. Using persistence image and persistence landscape based features to characterize glandular architecture in colorectal tissues, we were able to classify between benign and malignant images with a high degree of accuracy. Our preliminary experiments indicate that the performance of these features is better than state-of-the-art features based on cell graphs. Future work will focus on improving the scalability of these features and improved dimensionality reduction to deal with their high-dimensionality. Furthermore, considering the vectorized nature of persistence image and landscape representations, we will evaluate their effectiveness in a deep learning setting.

6. REFERENCES

[1] S. Doyle, M. Hwang, K. Shah, et al., "Automated grading of prostate cancer using architectural and textural image features," in *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2007, pp. 1284–1287.

[2] S. Doyle, S. Agner, A. Madabhushi, et al., "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, may 2008, pp. 496–499.

[3] D. Altunbay, C. Cigir, C. Sokmensuer, et al., "Color graphs for automated cancer diagnosis and grading," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 665–674, mar 2010.

[4] N. Singh, H. D. Couture, J. S. Marron, et al., "Topological Descriptors of Histology Images," in *International Workshop on Machine Learning in Medical Imaging*, 2014, pp. 231–239.

[5] H. Adams, S. Chepushtanova, T. Emerson, et al., "Persistence Images: A Stable Vector Representation of Persistent Homology," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 218–252, 2015.

[6] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *Journal of Machine Learning Research*, vol. 16, pp. 77–102, 2012.

[7] X. Zhu, "Persistent homology: An introduction and a new text representation for natural language processing," in *International Joint Conference on Artificial Intelligence*, 2013, pp. 1953–1959.

[8] E. Reinhard, M. Ashikhmin, B. Gooch, et al., "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.

[9] M. Macenko, M. Niethammer, J. S. Marron, et al., "A method for normalizing histology slides for quantitative analysis," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, jun 2009, pp. 1107–1110.

[10] C.H. Li and P.K.S. Tam, "An iterative algorithm for minimum cross entropy thresholding," *Pattern Recognition Letters*, vol. 19, no. 8, pp. 771–776, jun 1998.

[11] Y. Al-Kofahi, W. Lassoued, W. Lee, et al., "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, 2010.

[12] K. Sirinukunwattana, J. P. W. Pluim, H. Chen, et al., "Gland segmentation in colon histology images: The glas challenge contest," *Medical Image Analysis*, vol. 35, pp. 489–502, jan 2017.