Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images

Patrick Leo
George Lee
Natalie N. C. Shih
Robin Elliott
Michael D. Feldman
Anant Madabhushi

SPIE.

Patrick Leo, George Lee, Natalie N. C. Shih, Robin Elliott, Michael D. Feldman, Anant Madabhushi, "Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images," *J. Med. Imag.* **3**(4), 047502 (2016), doi: 10.1117/1.JMI.3.4.047502.

Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images

Patrick Leo,[a,*] George Lee,[a] Natalie N. C. Shih,[b] Robin Elliott,[c] Michael D. Feldman,[b] and Anant Madabhushi[a,*]
[a]Case Western Reserve University, Department of Biomedical Engineering, 2071 Martin Luther King Jr. Drive, Cleveland, Ohio 44106, United States
[b]University of Pennsylvania, Department of Pathology, 3400 Spruce Street, Philadelphia, Pennsylvania 19104, United States
[c]Case Western Reserve University, Department of Pathology, 11100 Euclid Avenue, Cleveland, Ohio 44106, United States

Abstract. Quantitative histomorphometry (QH) is the process of computerized feature extraction from digitized tissue slide images to predict disease presence, behavior, and outcome. Feature stability between sites may be compromised by laboratory-specific variables including dye batch, slice thickness, and the whole slide scanner used. We present two new measures, preparation-induced instability score and latent instability score, to quantify feature instability across and within datasets. In a use case involving prostate cancer, we examined QH features which may detect cancer on whole slide images. Using our method, we found that five feature families (graph, shape, co-occurring gland tensor, sub-graph, and texture) were different between datasets in 19.7% to 48.6% of comparisons while the values expected without site variation were 4.2% to 4.6%. Color normalizing all images to a template did not reduce instability. Scanning the same 34 slides on three scanners demonstrated that Haralick features were most substantively affected by scanner variation, being unstable in 62% of comparisons. We found that unstable feature families performed significantly worse in inter- than intrasite classification. Our results appear to suggest QH features should be evaluated across sites to assess robustness, and class discriminability alone should not represent the benchmark for digital pathology feature selection. [DOI: 10.1117/1.JMI.3.4.047502]

Keywords: quantitative histomorphometry; prostate cancer; feature stability; site variation; digital pathology; stain variability; prognosis; machine learning.

Paper 16054PRR received Apr. 3, 2016; accepted for publication Sep. 16, 2016; published online Oct. 24, 2016.

1 Introduction

Quantitative histomorphometry (QH) is the process of computerized extraction of features from digitized slide images.[1–12] These features are typically then used with automated classification methods to predict disease presence, behavior, and outcome. A major challenge for QH is the variation among pathology images across multiple sites. This variation is induced in the preparation phase prior to computational image analysis when tissue samples are stained, mounted onto a glass slide, and subsequently digitized via a whole slide scanner. Stain concentration, manufacturer, and batch effects affect the final appearance of a slide.[13] In addition, the specific whole slide scanner used to digitize a slide can affect the appearance of the final digital image. All these preparation-induced image variations can affect the automated analysis of the image and thus the calculated feature values.

Image variation affects the features computed from an image and thus poses a problem for diagnostic and predictive algorithms based on these features. A key step in using these algorithms is choosing which features to use for classification. Traditional classification-based performance measures such as accuracy and area under the receiver operating characteristic curve are typically employed in feature selection methods that aim to identify features that maximize class discriminability. But to create a robust classifier, the feature selection algorithm must consider both discrimination and stability. A feature is stable if the mean and shape of its distribution is consistent among cohorts of patients who share disease or clinical profiles or outcomes. While feature stability has been examined in the radiology space,[14–17] relatively little work has been done in the context of QH or in digital pathology.

Cross institutional color variation is a well-known problem in digital pathology as evidenced by the large number of methods developed to quantify image color and standardize images to a template.[18–21] While standardization of stains and procedures as suggested in Lyon et al.[13] could help reduce variation, logistical and physical limitations mean that digital color correction will always be needed to ensure uniform color. Color normalization (CN) is broadly the process of altering the color channel values of pixels in a source image so that its color distribution matches that of a template image. Since image color is affected by preparation procedure and thus is a possible contributor to feature instability, one needs to evaluate the effect of CN on the resulting feature expression values. Some work has been done to evaluate the effect of CN on classifier performance in applications such as mitosis detection,[18] but to our knowledge no study has been performed which examines the link between CN and feature values or feature stability. Staining and scanning

*Address all correspondence to: Patrick Leo, E-mail: patrick.leo@case.edu; Anant Madabhushi, E-mail: anant.madabhushi@case.edu

2329-4302/2016/$25.00 © 2016 SPIE

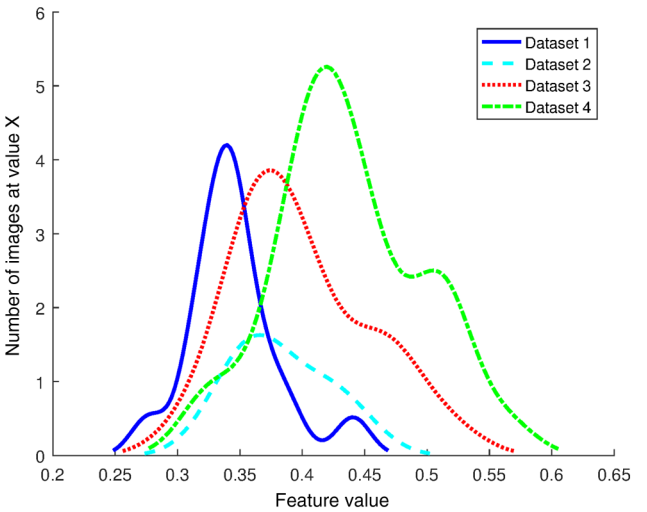


Fig. 1 Smoothed distribution of graph feature F_{48} "disorder of nearest neighbor in 20 pixel radius" from the nonrecurrence Gleason 7 patients of four datasets.

procedures should not, in an ideal setting, dramatically affect the value of a stable feature. An example of an unstable feature is shown in Fig. 1. The feature value distributions from four datasets are shown to be of similar shape but with different means and modes.

In this paper, we investigate the effect of feature instability on classification, introduce a method for quantifying feature stability across many digital pathology datasets, and determine instability specifically resulting from use of different whole-slide scanners. To examine the degree to which site variability affects feature accuracy, we investigated the degradation of classifier performance in the context of intra- and intersite classification of tumor and nontumor regions of prostate tissue for 81 patients. Stability is evaluated using two feature-based evaluation measures. The first stability evaluation measure, latent instability (LI) score, aims to evaluate the inherent randomness of a feature's distribution within a single-preparation procedure from factors such as interpatient feature variation. A low LI would indicate that intradataset variation is very low and that there is a low probability of features being different between two datasets with similar disease or clinical profiles or outcome due to random chance. We also introduce a method involving cross-dataset comparisons for quantifying the frequency at which a feature is different between datasets via a preparation-induced instability (PI) score. A high PI would indicate that a feature is affected by preparation procedures. Lastly, we apply these methods to a use case involving detecting tumor and nontumor regions on radical prostatectomy (RP) samples taken from prostate cancer patients based off QH analysis of digitized images. Our group has previously investigated the role of a number of different histomorphometric features including gland and nuclear shape, morphology, orientation, and disorder with prostate cancer presence, grade, aggressiveness, and outcome.[22–29] The goal of this study is to identify which of these classes of features, which are predictive of presence of prostate cancer, are most stable across sites and scanners. Specifically, in this paper, we examine the stability of 216 gland lumen and 26 texture features extracted from 80 whole mount prostate adenocarcinoma (CaP). Our goal was to compare the intra- and interdataset variations of the prostate histology QH features to determine if the QH variance across sites is significantly larger than might be expected due to random chance.

We applied CN to the four datasets and measure feature instability among datasets before and after normalization. The goal of the experiment was to gain some insight into the potential of CN as a technique for reducing feature instability among datasets. We scanned the 34 slides of a single dataset on three different scanners to examine the specific contribution of scanner variation to feature instability. While clearly there are multiple sources of variance affecting the stability of image features, in this paper, we focus on two critical aspects in digital pathology, color variance due to differences in site and slide digitization and the induced color variation on account of different scanners.

Thus our contributions in this paper are

- A method for evaluating precisely how histomorphometric features tend to vary across sites with varying preparation procedures.
- New quantitative measures to evaluate feature stability across and within datasets, as well as quantitatively assessing the effect of CN on the resulting feature expression.
- An evaluation of the stability of 242 QH features from five feature families in prostate histology across sites before and after CN and across whole-slide scanners.

The rest of this paper is organized as follows. Section 2 introduces the new metrics used to evaluate feature stability. Section 3 describes how feature instability affects classification and the CN, segmentation, and feature extraction techniques employed. Section 4 lays out the experimental design and quantitative results. Section 5 concludes the paper with a summary of the study and closing remarks.

2 Feature Robustness Indices

2.1 *Latent Instability Score*

To create a baseline for the expected feature value variation between images we compared feature distributions within random splits of a dataset D_k, $k \in \{1,2,3,4\}$, $D_k \cap D_r = \{\}$, $r \neq k$. Random halves of D_k were compared against each other to check for significant differences in feature distribution between the halves. The end result of this process is a calculation of a feature's LI score which represents the probability that a feature will be different among datasets due to effects not linked to the specific laboratory. Random dataset splits used to calculate LI produce subsets that contain different patients with different image content. LI is a measure of a feature's inherent variability and the degree to which interpatient variance may contribute to feature instability when measured across datasets.

We split images c_a, $a \in \{1,\ldots,N\}$, where N is the number of images in dataset D_k into two equal parts S_1^i, S_2^i, for $i \in \{1,\ldots,N_{\text{iter}}\}$, where N_{iter} refers to the number of splitting iterations. S_1 and S_2 are sets of $N/2$ images c_a such that the subsets are unique, all $S_1^i \cap S_2^i = \{\}$.

From the set of all features F, we examined one feature at a time, F_j, $j \in \{1, N_F\}$, N_F is the total number of features. We computed the LI of F_j in D_k by taking the percentage of splits where the vector of feature values $f_j(S_1^i)$ and $f_j(S_2^i)$ were statistically significantly different under the Wilcoxon rank sum test (U), such that

$$\mathrm{LI}_j^k = \frac{1}{N_{\mathrm{iter}}} \sum_{i=1}^{N_{\mathrm{iter}}} U[f_j(S_1^i), f_j(S_2^i)], \qquad U = \begin{cases} 1, & \text{if } p_i < 0.05 \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where $N_{\mathrm{iter}} = 1000$, and f_j is the feature representation of the j'th feature of all the images in S_1^i and S_2^i, and p_i is the p-value of U in iteration i. U was chosen due to its resilience to the effect of outliers and its ability to handle unknown distributions. Feature distributions were considered to be significantly different if they were different at the $p < 0.05$ confidence level. LI_j^k is equal to the percentage of splits in which feature F_j was significantly different between halves of D_k. Therefore LI $\subset [0,1]$.

2.2 *Preparation-Induced Instability Score*

The feature values of two cohorts were compared using U. A pairwise comparison of each combination of datasets D_k, D_r was performed over N_{iter} trials using a random three-quarters of each dataset, creating D_k^i and D_r^i for every iteration i for a total of $\left[N_{\mathrm{iter}} \times \binom{N_D}{2}\right]$ comparisons per feature. From these comparisons, PI was calculated as a quantitative measure of the sensitivity of a feature to staining and scanning procedures. While LI measures how often a feature would be expected to be different among datasets without any differences among laboratories, PI represents how often that feature was actually found to be different among datasets. The difference between a feature's LI and PI thus reveals the effect of laboratory preparation procedure on that feature.

$$\mathrm{PI}_j = \frac{1}{\binom{N_D}{2}} \sum_{k=1}^{N_D} \sum_{r=k+1}^{N_D} \left\{ \frac{1}{N_{\mathrm{iter}}} \sum_{i=1}^{N_{\mathrm{iter}}} U[f_j(D_k^i), f_j(D_r^i)] \right\},$$
$$U = \begin{cases} 1, & \text{if } p < 0.05 \\ 0, & \text{otherwise.} \end{cases}, \tag{2}$$

PI_j is equal to the percentage of pairwise comparisons in which f_j was significantly different among datasets. A high PI paired with a low average LI suggests that the feature is different among datasets due to fundamental differences among the datasets rather than noise inherent in the feature or interpatient variability.

3 Methods and Materials

3.1 *Significance of Feature Instability*

To determine the effect of feature instability and link the image extracted features to a specific task, we performed an experiment to test classification of tumor and benign regions in RP specimens. As all the RP patients had prostate cancer, the benign regions were selected from the noncancerous zones of those specimens. A total of 81 patients from two sites were used for this classification experiment. Dataset information is provided in Table 1. To determine the effect of using feature families of varying stability, classification experiments were performed using each feature family separately and with all 242 features. The results of cross-validation within a single site and independent validation across sites were compared with the hypothesis that unstable features would perform better in intrasite classification than in intersite tasks where feature instability would adversely affect model generalizability. Features were selected using the Wilcoxon rank-sum significance test between the tumor and benign regions of the training set at the $p < 0.05$ significance level. For patients with multiple tumor or benign regions, only the largest region of each class was used. Classification was performed using a random forest classifier[30] with 50 trees over 100 iterations. In each fold of cross validation, a random two-thirds of a site's patients were used for training with the remaining third used for validation, with all regions of every patient kept in the same set. Significant features were selected on the training data independently in each fold.

Table 1 Patients used for cancerous and noncancerous region classification.

	D_1	D_2
Total patients	40	41
With tumor regions only	5	3
With tumor and benign regions	35	38

3.2 *Statistical Analysis of Feature Robustness Indices*

The purpose of the LI experiment was twofold. First, a low LI_j^k provides confirmation that f_j is relatively consistent within D_k. By looking only within a single dataset, we were able to partially control for the staining and scanning procedures and hence potentially identify if there were features that were inherently noisy even without additional confounding sources of variation. Second, the frequency of feature difference between the halves of the datasets allowed for establishment of a baseline for how often a feature would appear different between two sets due to random chance. This baseline is the measure by which we may judge interdataset feature instability. If a distribution was rarely different between the two halves of D_k, it would appear unlikely that the feature distribution would be different across two different datasets.

A number of confounders affect the interpretation of the PI results, among them the difference in image content among the four patient cohorts. We have controlled for some clinical factors which may affect the resulting features (Gleason sum and patient outcome). The evidence of site-variation affecting stability may be assessed by comparing LI and PI results. Both LI and PI are arrived at by comparing unique sets of patients. However, while LI involves comparing patients from the same site, PI involves comparing images across sites. LI reflects the contribution of interpatient variation to feature instability while PI includes both patient variation and site-specific variation. Hence the ratio of PI to LI reflects the contribution of site variation to increased feature instability, a ratio that allows for the isolation of the site-induced variability from the image-specific variations. Further there are a number of factors that are not controlled for in the PI experiment, including image compression and original image magnification. However, in not controlling for these factors, we are following data acquisition protocols typical in digital pathology and which may be encountered in a typical clinical setting. Our findings suggest that greater

Table 2 Summary of features examined.

Family	Description	Features
Graph	Descriptors of Delaunay, Voronoi, and minimum spanning tree diagrams	51
Shape	Lumen shape, smoothness, invariant moments, and Fourier descriptors	100
CGT[6]	Entropy of gland orientation and neighborhood disorder	39
Subgraph	Local subgraph connectivity and distance between nodes	26
Texture	Relative pixel intensity, contrast, entropy, and energy	26

attention to standardization of slide preparation, digitization, and analysis may be needed.

3.3 *Color Normalization*

We employed the nonlinear stain mapping CN method described by Khan et al.[18] to normalize the color of the slide images. This method maps each stain in the source and template image to a channel, normalizes the channels, and then converts the image back to the RGB space. Other methods commonly used for hematoxylin and eosin (H&E) image normalization such as histogram color matching[31] normalize the RGB channels themselves rather than using stain channels. However, this has the disadvantage of possibly inducing artifacts in the image.[18] In our case, these artifacts severely degrade the performance of automated gland identification methods and thus necessitate the use of a stain channel approach. The images in dataset D_k were color normalized to create dataset D_k^N.

3.4 *Segmentation*

Gland lumen were automatically segmented from digital images of the cancerous regions of RP whole mount slide images using the approach described in Nguyen et al.[32] To segment lumen, the algorithm first performed k-means clustering of the colors of 10,000 randomly selected pixels in an image with $k = 4$. Pixels were given a label based on their cluster to define the prototypical color of nuclei, stroma, cytoplasm, and lumen in an image. These prototypes were then applied to the entire image to identify objects. Lumen objects surrounded by nuclei objects were deemed to be glands and the boundaries of the lumen were segmented.

3.5 *Feature Extraction and Analysis*

A total of 216 gland lumen features were extracted from segmented gland lumen after resizing segmentation results and images to 1.25× magnification. Twenty-six Haralick features were extracted from the pixel intensity values of the entire image and involved excluding pixels corresponding to the gland segmentations. The extracted features belonged to five different families and are described in Table 2. A visualization of the five feature families is shown in Fig. 2. These features were chosen for analysis based on their relevance to CaP recurrence prediction as described in previous work from our group.[6,33] Gland features are related to disease aggressiveness since more aggressive prostate cancer degrades the cohesiveness and regularity of the glands. These 216 features attempt to quantitatively capture gland morphology, shape, arrangement, and disorder of glands in the image. Since these features are intended to differentiate cancerous and noncancerous cases and images corresponding to different Gleason grades, there is a need to identify features that are stable and consistent.

The 51 graph-based features captured the spatial arrangement of the glands as calculated by using gland centroids as vertices. These include first- and second-order descriptors of Voronoi diagrams [see Figs. 2(c) and 2(i)], Delaunay triangulations, minimum spanning trees, and gland density.[6]

The 100 gland shape features measured the average shape of all the glands in an image as described by the lumen boundaries and the resulting area, perimeter, distance, smoothness, and Fourier descriptors.[23]

The 39 co-occurring gland tensor (CGT) features measured the disorder of neighborhoods of glands as measured by the entropy of orientation of the major axes of glands within a local neighborhood.[6] Gland orientations can be visualized in

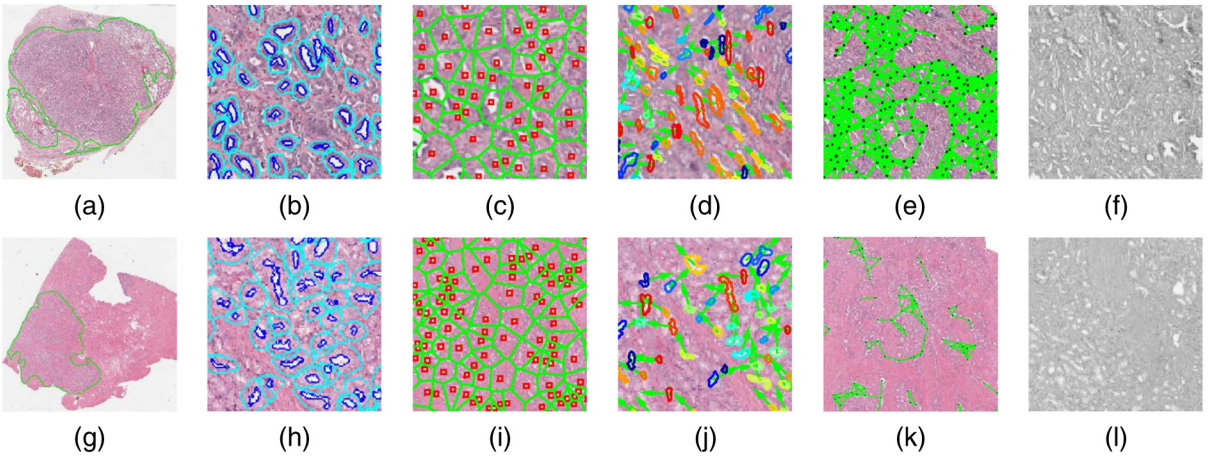


Fig. 2 Segmentation and feature visualization for (a–f) an image in D_1 and (g–l) an image in D_3. (a, g) Cancerous regions annotated by expert pathologist. Automatically extracted features corresponding to (b, h) gland shape, (c, i) global gland graphs, (d, j) gland disorder, (e, k) local gland graphs, and (f, l) Haralick intensity texture.

Figs. 2(d) and 2(j), where the gland boundaries are color coded based on the angle of their orientation.

The 26 subgraph features described the connectivity and clustering of small gland neighborhoods using gland centroids [see Figs. 2(e) and 2(k)].[5]

The 26 Haralick texture features measured second-order intensity statistics.[34] These features do not explicitly rely on gland segmentations.

3.6 *Dataset Description*

We collected 146 H&E-stained whole mount prostate tissue RGB images for the purpose of detecting cancerous regions on RP specimens. A series of QH features (as described in Sec. 3) previously shown to be predictive of biochemical recurrence of prostate cancer were extracted. Two separate cohorts from the University of Pennsylvania contained 41 and 40 patients, respectively (D_1 and D_2). The Cancer Genome Atlas[35] provided two datasets, 32 patients from the University of Pittsburgh (D_3) and 33 from Roswell Park (D_4). All images were digitized using a whole slide scanner. The images within D_3 and D_4 were available in their digital form from the TCGA. D_1 and D_2 were originally digitized at 20× magnification on an Aperio CS2 scanner. The precise make and model of the scanner used for the cases in D_3 and D_4 was not known; however, they were originally digitized at 40× magnification. For each patient, a representative cancerous region of interest was identified and annotated by an expert pathologist. Features were calculated from within the annotated region. All images in all experiments were downsized to 5× magnification (0.5 μm per pixel) for gland segmentation and 1.25× (2 μm per pixel) for feature extraction.

3.6.1 *Feature stability experiment*

To investigate the stability of these features, we controlled the populations across the datasets and matched 80 patients for Gleason score 7 (GS7) and no cancer recurrence within 5 years of surgery. Fitting these requirements were 16 patients from D_1, 14 from D_2, 28 from D_3 and 22 from D_4. This pruning was done to ensure that differences in the datasets were not due to differences in the populations.

3.6.2 *Cancerous and noncancerous region classification*

To evaluate the performance of classifiers that used varyingly stable feature families, all images of D_1 and D_2 were used. Table 1 describes the 81 patient dataset used in this experiment.

3.6.3 *Multiscanner experiment*

Thirty four slides from D_2 were scanned on a Leica Aperio CS2 (D_2^A), Phillips IntelliSite Ultra Fast Scanner (D_2^P), and Roche Ventana iScan HT (D_2^V) scanner. While D_2 contains 40 slides, some slides did not successfully scan on every scanner and hence were not used in this experiment. D_2^A was originally digitized at 20× magnification while D_2^P and D_2^V were digitized at 40×.

4 Experimental Results and Discussion

4.1 *Latent Instability to Evaluate Intradataset Feature Robustness*

Low variation in f_j within D_k appeared to suggest that the differences among cohorts were a result of variation between the staining and scanning procedures used for the slides.

All feature families across all cohorts exhibited low LI scores. No feature family in any unstandardized dataset had a mean LI score higher than 0.0522 as seen in Table 3 and Fig. 3, indicating just a 5.22% chance of a feature being significantly different in a random split. These low LI scores were indicative of low intradataset feature instability. This suggests that within the same dataset, the features do not tend to vary considerably.

4.2 *Preparation-Induced Instability to Evaluate Feature Robustness Across Cohorts and Sites*

To evaluate the effect of CN on feature value calculation and to investigate the role that CN plays in reduction of interdataset feature instability, we compared the number of feature distributions that were significantly different across datasets before and after normalization. We compared the mean preparation-induced instability score (μPI) for each family of features shown in Table 4 and Fig. 4 to the mean latent instability score (μLI) for that feature family. μPI is the mean of the PI scores of all

Table 3 Mean and standard deviation of LI score by cohort and feature family.

Site	Graph	Shape	CGT	Subgraph	Texture
D_1	0.0499 ± 0.0048	0.0498 ± 0.0118	0.0464 ± 0.0094	0.0473 ± 0.0101	0.0504 ± 0.0095
D_2	0.0334 ± 0.0066	0.0326 ± 0.0087	$\mathbf{0.0296 \pm 0.0130}$	0.0338 ± 0.0090	0.0357 ± 0.0071
D_3	0.0443 ± 0.0051	0.0422 ± 0.0107	$\mathbf{0.0522 \pm 0.0188}$	0.0435 ± 0.0117	0.0507 ± 0.0088
D_4	0.0449 ± 0.0060	0.0429 ± 0.0093	0.0410 ± 0.0158	0.0437 ± 0.0055	0.0457 ± 0.0030
D_1^N	0.0543 ± 0.0064	0.0494 ± 0.0110	0.0468 ± 0.0085	0.0490 ± 0.0095	$\mathbf{0.0558 \pm 0.0067}$
D_2^N	0.0493 ± 0.0048	0.0453 ± 0.0115	0.0395 ± 0.0161	0.0443 ± 0.0120	0.0448 ± 0.0110
D_3^N	$\mathbf{0.0363 \pm 0.0065}$	0.0436 ± 0.0099	0.0445 ± 0.0150	0.0418 ± 0.0128	0.0439 ± 0.0073
D_4^N	0.0459 ± 0.0053	0.0423 ± 0.0093	0.0411 ± 0.0147	0.0469 ± 0.0070	0.0416 ± 0.0089

Note: The bold values represent the most stable and least stable feature families amongst the D_{1-4} and D_{1-4}^N cohorts.

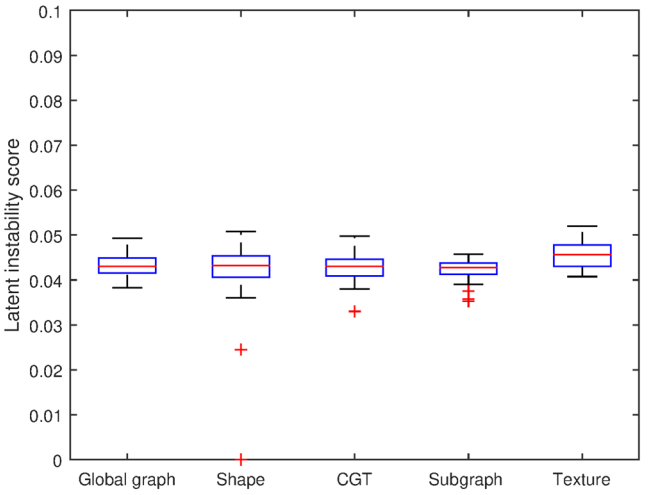


Fig. 3 Distribution of LI results by feature family.

the features in a family and μLI is the mean of a family's LI across D_{1-4}.

Shape features were found to be most resilient while CGT features were the most unstable. Shape features were different among datasets $(\mu\mathrm{PI}_{\mathrm{Shape}}/\mu\mathrm{LI}_{\mathrm{Shape}}) = 4.70$ times more often than would be expected without differences in slide preparation among datasets while CGT features were $(\mu\mathrm{PI}_{\mathrm{CGT}}/\mu\mathrm{LI}_{\mathrm{CGT}}) = 11.49$ times more likely to be different.

CGT features measure the disorder of lumen orientation in neighborhoods of glands. Missed and erroneously segmented gland objects will have a larger effect on CGT features which compare glands to their neighbors while having less of an effect on shape measurements. This is consistent with the μPI results. The families with high μPI, graph, CGT, and subgraph, all depend on accurate gland detection. In contrast, the shape features are largely based off the shape of the lumen. These, therefore, appear to be less affected by segmentation errors. Texture features, which do not rely on segmentation, fall between the gland arrangement-dependent families and the shape family in terms of μPI.

4.3 *Cancerous and Noncancerous Region Classification to Determine Effect of Unstable Feature Families on Independent Validation*

As seen in Table 5, the more unstable feature families, graph, CGT, and subgraph were significantly more accurate in intrasite classification than in the intersite task ($p = 1.6e-19$, $6.9e-62$, and $5.5e-41$). Shape, the most stable feature family, was the only feature family to perform equally well in intra- and intersite classification by AUC ($\mathrm{AUC}_{\mathrm{Intra}} = 0.96$, $\mathrm{AUC}_{\mathrm{Inter}} =$

Table 4 Mean and standard deviation of PI score in nonrecurrence Gleason 7 images of D_{1-4} (top row) and D^N_{1-4} (middle row) and in 34 images of D_2 across three scanners (bottom row).

Dataset	Graph	Shape	CGT	Subgraph	Texture
D_{1-4}	0.466 ± 0.184	$\mathbf{0.197 \pm 0.102}$	$\mathbf{0.486 \pm 0.193}$	0.429 ± 0.168	0.438 ± 0.201
D^N_{1-4}	0.479 ± 0.185	$\mathbf{0.207 \pm 0.118}$	$\mathbf{0.491 \pm 0.192}$	0.448 ± 0.168	0.428 ± 0.210
$D_2^{A,P,V}$	0.267 ± 0.161	0.411 ± 0.327	0.245 ± 0.112	$\mathbf{0.197 \pm 0.127}$	$\mathbf{0.620 \pm 0.193}$

Note: The bold values represent the most stable and least stable feature families in each experiment.

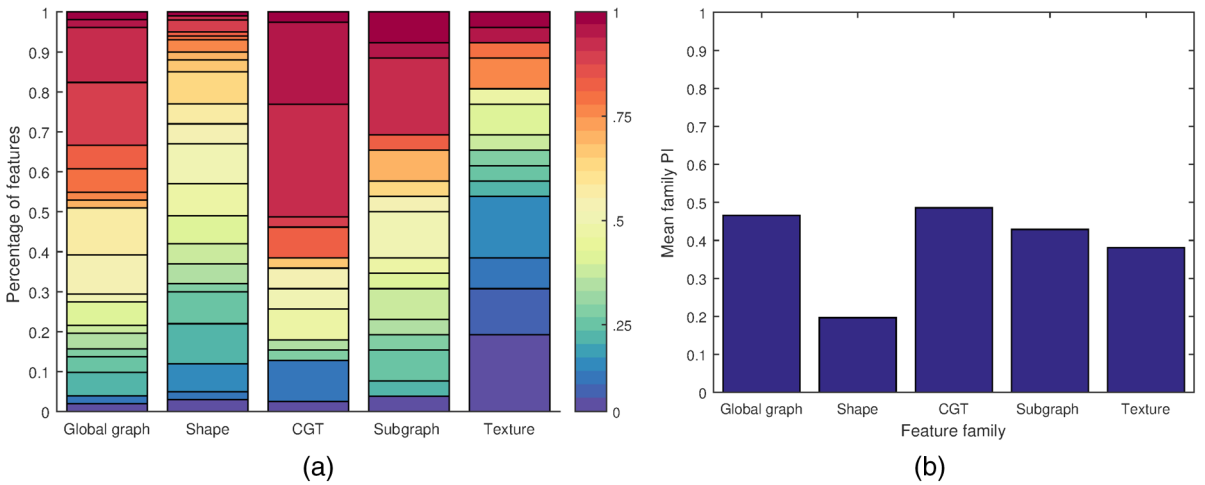


Fig. 4 Summary of intra- and interdataset experiment results. (a) Feature family PI results by score frequency. Score indicates percentage of the 6 1000-iteration subsampled pairwise comparisons (D_1 versus D_2, D_1 versus D_3, D_1 versus D_4, D_2 versus D_3, D_2 versus D_4, D_3 versus D_4) in which the given feature was significantly different. Results plotted after grouping the 239 unique feature scores into 30 bins. (b) Average PI results by feature family computed as the mean score from all the features within a family.

Table 5 Mean and standard deviation of AUC and accuracy of QH features in distinguishing tumor from benign regions over 100 iterations of a random forest classifier using Wilcoxon rank-sum test for feature selection at the $p < 0.05$ confidence level. Intrasite classification performed using threefold cross validation, intersite classification performed using independent validation. Reported as average of D_1 and D_2 results. No texture features were predictive, and the family is therefore not represented here.

	Graph	Shape	CGT	Subgraph	All families
Intrasite AUC	0.89 ± 0.12	0.96 ± 0.03	0.91 ± 0.07	0.90 ± 0.10	0.93 ± 0.05
Intersite AUC	0.83 ± 0.09	0.97 ± 0.02	0.80 ± 0.02	0.84 ± 0.04	0.95 ± 0.03
p-Value	5.7×10^{-12}	0.88	1.3×10^{-52}	2.0×10^{-8}	0.29
Intrasite accuracy	0.87 ± 0.15	0.89 ± 0.06	0.88 ± 0.06	0.88 ± 0.12	0.86 ± 0.09
Intersite accuracy	0.75 ± 0.05	0.90 ± 0.04	0.73 ± 0.03	0.75 ± 0.05	0.84 ± 0.09
p-Value	1.6×10^{-19}	3.6×10^{-4}	6.9×10^{-62}	5.5×10^{-41}	0.01

0.97, $p = 0.88$) with slightly higher accuracy in the intersite task ($\text{Acc}_{\text{Intra}} = 0.89$, $\text{Acc}_{\text{Inter}} = 0.90$, $p = 3.6e-4$). The performance drop of the more unstable features in independent validation and the consistency of the less unstable features suggests that feature instability affects the model's generalizability.

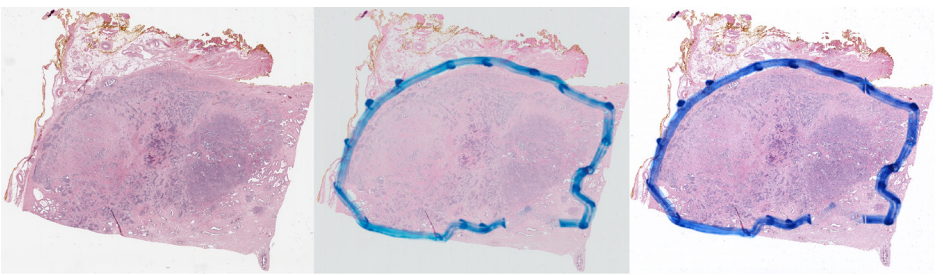
4.4 *Evaluating Effect of Scanner Variation on Feature Stability*

As seen in Table 4, automated segmentations of the same 34 D_2 slides scanned on three different scanners produced features which were less unstable in three families (global graph, CGT, and subgraph) and more unstable in one family (shape). The Haralick texture features were extremely unstable across scanners with a mean PI of 0.620 compared to 0.438 across D_{1-4}. Notably, all the images used in the D_{1-4} PI calculation were digitized using an Aperio scanner including D_1 and D_2 which were digitized on the exact same scanner. The scanner used appears to have a large effect on the texture features, which are dependent on color and contrast changes induced in the digitized slide images. It is possible the instability measurements of D_{1-4} are affected by the Aperio scanner which in turn may explain the variation in the PI of the texture features between the two experiments. Figure 5 shows these variations as well as an example of the blue annotations added onto the slides by another group; however, features were extracted from regions entirely inside the blue contour.

The lower instability of the global graph, CGT, and subgraph families in the $D_2^{A,P,V}$ experiment is not unexpected. While D_{1-4} vary in sectioning, mounting, and staining procedures in addition to scanning equipment, $D_2^{A,P,V}$ only vary in digitization hardware. It is intuitive that datasets separated only by the scanner used would be more stable than datasets prepared by entirely different laboratories. However, the large increase in shape feature PI mean and standard deviation (0.197 ± 0.102 to 0.411 ± 0.327) is somewhat surprising as the shape features were the most stable feature family by a large margin in the D_{1-4} experiment. Figure 6 shows that in some regions of interest, the automated segmentation performance varies greatly across scanners, though we did not undertake a quantitative assessment of the degree of segmentation variation. The very large standard deviation in shape feature PI as well as the instability variation among feature families suggests that some features are highly vulnerable to intersite variation while others are more robust.

4.5 *Evaluating Effect of Color Normalization on Feature Stability*

The PI and LI experiments were repeated using color normalized versions of the images of D_{1-4} to study the effects of CN on feature stability. Images were standardized to a template image, thereby creating normalized datasets D_{1-4}^N (see Fig. 7). The template image was chosen as one of the images where the automated segmentation algorithm was able to accurately extract



(a) (b) (c)

Fig. 5 A D_2 patient scanned on the (a) Aperio, (b) Phillips, and (c) Ventana scanners. The blue marker annotations were added between scans of the patient and were outside the region considered for feature analysis and thus did not affect feature values.

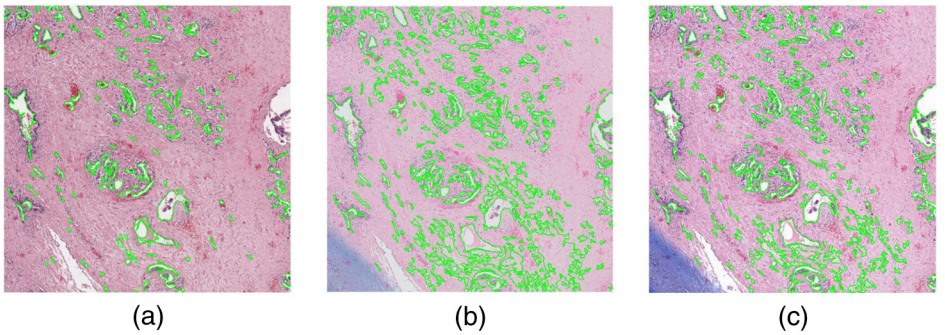


Fig. 6 Region of interest of a D_2 slide scanned on (a) Aperio, (b) Phillips, and (c) Ventana scanners with automated gland segmentation results overlaid. It is clear that the automated segmentation performed much worse on the Phillips scanner in this specific region.

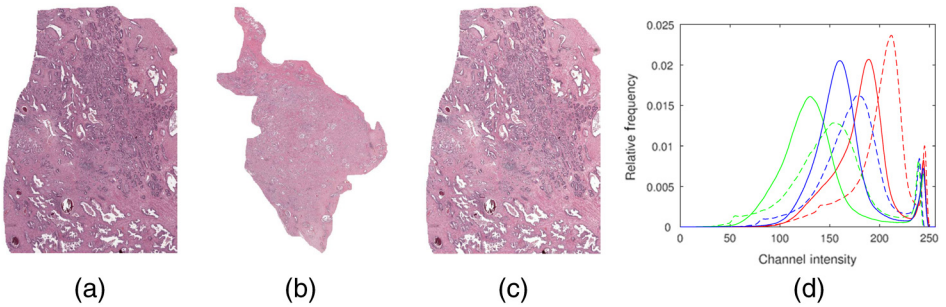


Fig. 7 Illustration of the experiment performed to evaluate the effect of CN on feature stability. (a) Source image to be normalized, (b) template image, and (c) resulting color transformed image. (d) Color histograms of pre-CN image (a) in solid lines and post-CN image (c) in dashed lines.

the gland boundaries. An example of how the feature values may change before and after normalization is shown in Fig. 8.

The goal of the CN experiments was to quantify how CN affected the number of unstable features. A decrease in feature family PI or LI after normalization would suggest that the process of CN was playing a role in reducing feature instability across or within sites. Our results suggest that CN had no effect on overall feature stability. The largest change in PI between original and normalized data was a change of 0.019 in the subgraph features with CN actually increasing intersite instability in four of the five feature families (see Table 4). The largest change in LI was in the graph features which had an increase in mean LI of 0.003, representing a 0.3% point increase in intrasite instability.

CN had a noticeable effect on the distribution of RGB values in the datasets. The histograms of eight images before and after normalization are shown in Fig. 9. After normalization, the histograms of each color channel were better aligned and had a more similar distribution. Notably, the chosen color standardization method[18] does not operate in the RGB space. Hence the changes in the histograms are merely byproducts of the

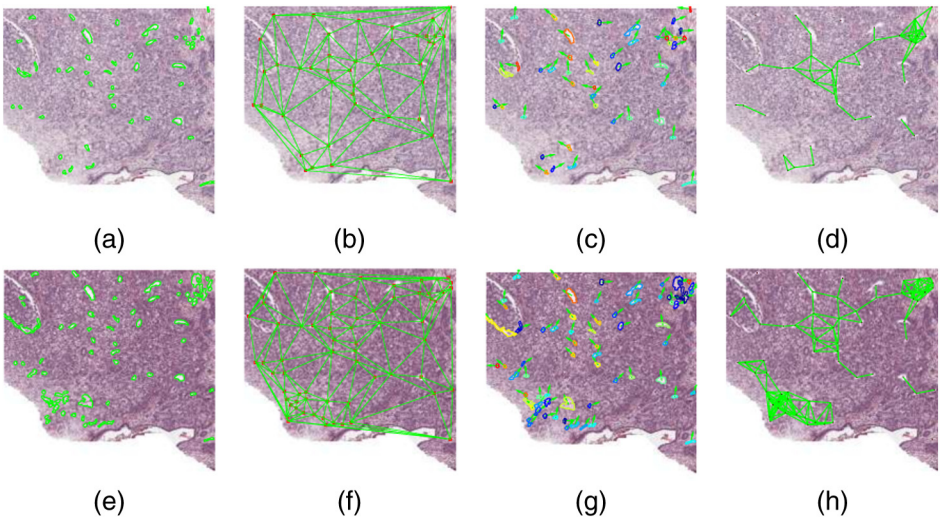


Fig. 8 Segmentation and feature visualization on image c_1 (a–d) prenormalization and (e–h) postnormalization. (a, e) Automated segmentation results. Automatically extracted features corresponding to (b, f) global gland graphs, (c, d) gland disorder, and (d, h) local gland graphs.

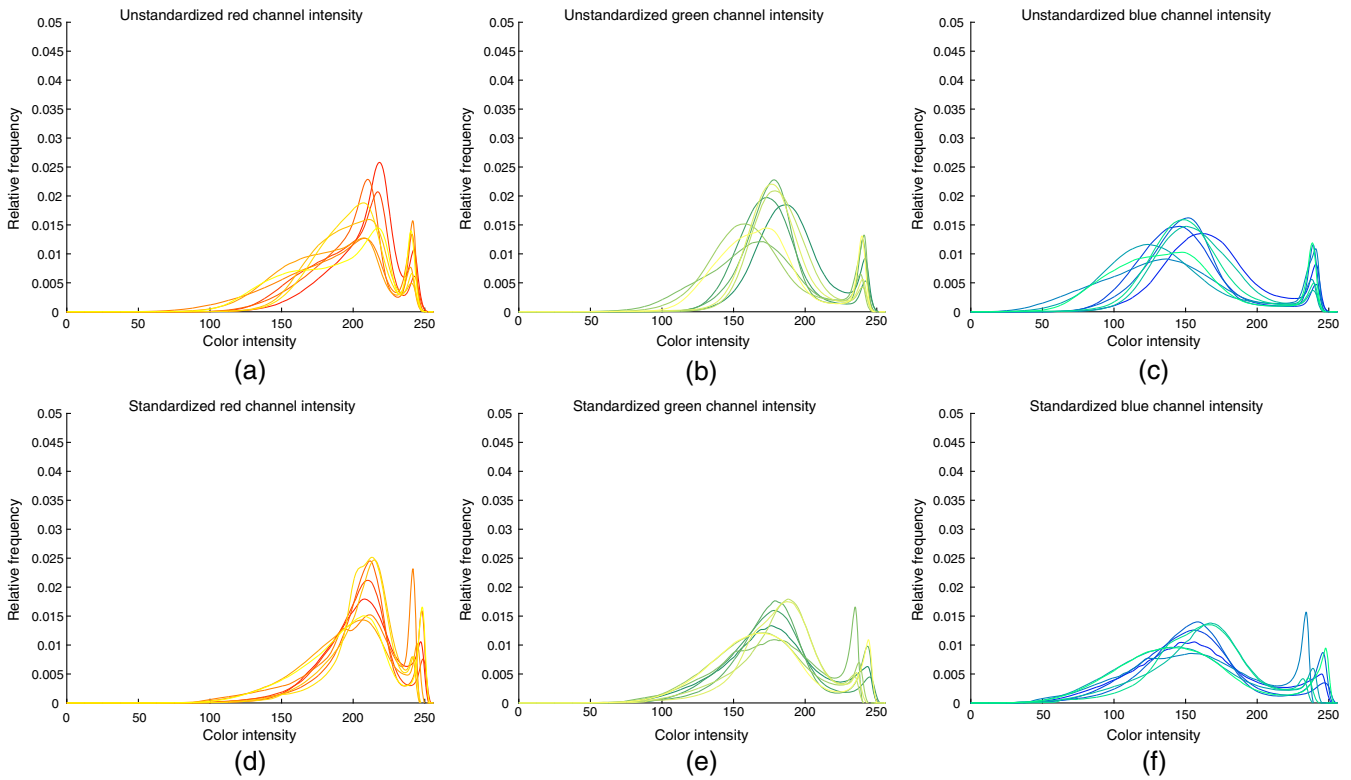


Fig. 9 Histograms of color channel intensity distributions for eight images (a–c) before and (d–f) after normalization. (a) Red, (b) green, and (c) blue channel intensity distributions in prenormalization images. (d) Red (e) green, and (f) blue channel intensity distributions in postnormalization images.

normalization of distributions in the color space. Figure 10 shows that the spread of RGB values across D_1 was reduced after normalization.

The results of the CN experiments also reveal that factors other than color appear to have played a large part in altering segmentation success and hence the resulting feature values. The chosen CN scheme is useful for this application because it attempts to normalize stain concentrations across images. Clearly the results of our experiments with CN will need to be validated via other CN schemes.

5 Concluding Remarks

In this paper, we presented a method for quantifying instability of features across multiple datasets and employed this method to determine stability in five feature families across four prostate cancer datasets with known variations due to staining, preparation, and scanning platforms. We introduced two new indices for quantifying feature instability and used them to identify which features previously shown to be predictive of cancer presence are most robust and stable. We found that all feature families exhibit differences among datasets from different institutions

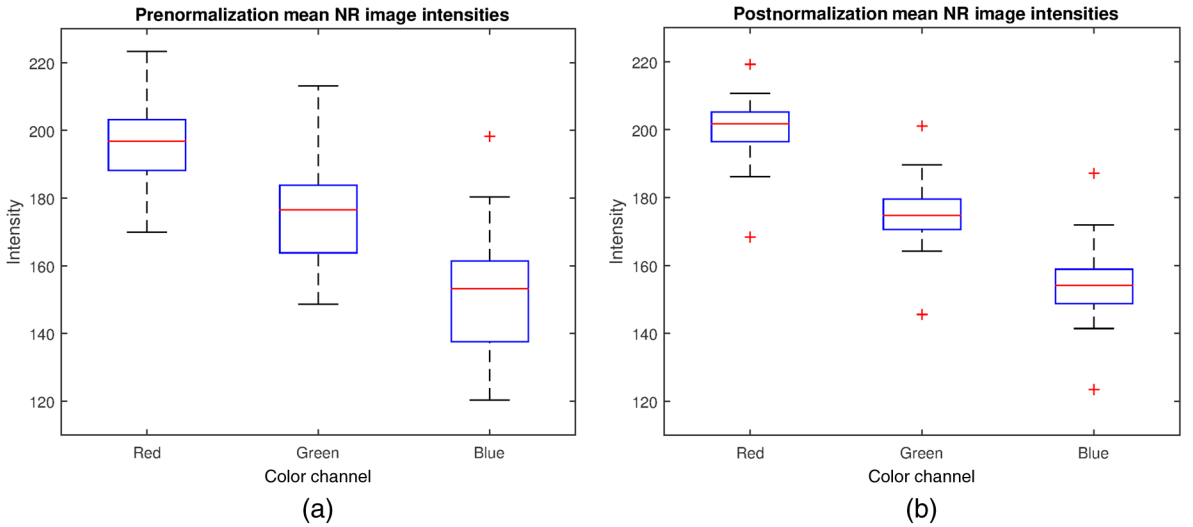


Fig. 10 Boxplots of the mean color channel intensities from each image in D_1 and D_2 (a) before normalization and (b) after normalization.

at a rate nearly 5 to 12 times what would be expected based on random chance. Although shape features were most resilient across datasets, all feature families exhibited sensitivity to staining and scanning procedure variations and the shape features themselves were especially vulnerable to scanner variation. With intra- and intersite classification, we have demonstrated that feature instability affects classifier performance and shown that a feature family's mean PI may indicate the degree to which the accuracy of those features may degrade with site variability. We have demonstrated that while CN can definitely affect the feature values and their stability, CN alone cannot solve the problem of interdataset feature instability. We found that the number of features different among datasets can change after CN, but in our case this did not resolve feature instability. Limitations to this study include that we did not consider other sources of variation that might be affecting the stability of the feature values, e.g., variations in segmentation. While we qualitatively observed that the automated gland segmentations varied among different scanners, we did not quantitatively determine the extent or cause of this variation. We did not examine how these instability scores may vary when using data from other sites or how sensitive our scores were to the specific tissue images used to calculate them. Additionally, we only considered one CN scheme in this work and the results obtained need to be validated in conjunction with other CN methods. We have not investigated how stability information may be incorporated in feature selection or classifier construction and our use case was confined to a specific application involving cancer detection on prostate histopathology images. In future work, we will seek to look at the trade-off between feature discriminability and stability more comprehensively and in the context of other use cases. This framework for quantifying feature instability may be useful in designing and developing future digital pathology-based computer-aided diagnostic algorithms which will need robust and discriminating features in order to be generalizable and consistent across sites.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers (R21CA167811-01, R21CA179327-01, R21CA195152-01, U24CA199374-01, and K01ES026841), National Institute of Diabetes and Digestive and Kidney Diseases (R01DK098503-02), DOD Prostate Cancer Synergistic Idea Development Award (PC120857), DOD Lung Cancer Idea Development New Investigator Award (LC130463), DOD Prostate Cancer Idea Development Award, Ohio Third Frontier Technology development Grant, CTSC Coulter Annual Pilot Grant, Case Comprehensive Cancer Center Pilot Grant, VelaSano Grant from the Cleveland Clinic, Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. J. Xu et al., "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Trans. Med. Imaging* **35**(1), 119–130 (2016).
2. S. Ginsburg et al., "Feature importance in nonlinear embeddings (FINE): applications in digital pathology," *IEEE Trans. Med. Imaging* **35**(1), 76–88 (2016).
3. A. Sridhar, S. Doyle, and A. Madabhushi, "Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces," *J. Pathol. Inf.* **6**, 41 (2015).
4. S. Ali et al., "Selective invocation of shape priors for deformable segmentation and morphologic classification of prostate cancer tissue microarrays," *Comput. Med. Imaging Graphics* **41**, 3–13 (2015).
5. G. Lee et al., "Supervised multi-view canonical correlation analysis (sMVCCA): integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE Trans. Med. Imaging* **34**, 284–297 (2015).
6. G. Lee et al., "Co-occurring gland angularity in localized subgraphs: predicting biochemical recurrence in intermediate-risk prostate cancer patients," *PLoS One* **9**, e97954 (2014).
7. A. Cruz-Roa et al., "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," *Lect. Notes Comput. Sci.* **8150**, 403–410 (2013).
8. G. Lee et al., "Cell orientation entropy (COrE): predicting biochemical recurrence from prostate cancer tissue microarrays," *Lect. Notes Comput. Sci.* **8151**, 396–403 (2013).
9. J. S. Lewis et al., "A quantitative histomorphometric classifier (QuHbIC) identifies aggressive versus indolent p16-positive oropharyngeal squamous cell carcinoma," *Am. J. Surg. Pathol.* **38**, 128–137 (2014).
10. R. Sparks and A. Madabhushi, "Explicit shape descriptors: novel morphologic features for histopathology classification," *Med. Image Anal.* **17**, 997–1009 (2013).
11. R. Sparks and A. Madabhushi, "Statistical shape model for manifold regularization: Gleason grading of prostate histology," *Comput. Vision Image Understanding* **117**, 1138–1146 (2013).
12. A. Basavanhally et al., "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides," *IEEE Trans. Biomed. Eng.* **60**, 2089–2099 (2013).
13. H. Lyon et al., "Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains and chromogenic reagents," *Histochem. J.* **26**(7), 533–544 (1994).
14. J. A. Oliver et al., "Variability of image features computed from conventional and respiratory-gated PET/CT images of lung cancer," *Transl. Oncol.* **8**(6), 524–534 (2015).
15. X. Fave et al., "Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer?" *Med. Phys.* **42**(12), 6784–6797 (2015).
16. M. J. Nyflot et al., "Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards," *J. Med. Imaging* **2**(4), 041002 (2015).
17. R. T. Leijenaar et al., "Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability," *Acta Oncol.* **52**(7), 1391–1397 (2013).
18. A. M. Khan et al., "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Trans. Biomed. Eng.* **61**(6), 1729–1738 (2014).
19. E. Reinhard et al., "Color transfer between images," *IEEE Comput. Graphics Appl.* **21**, 34–41 (2001).
20. M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," in *IEEE Int. Symp. on Biomedical Imaging: From Nano to Macro (ISBI '09)*, pp. 1107–1110 (2009).
21. S. Kothari et al., "Automatic batch-invariant color segmentation of histological cancer images," in *IEEE Int. Symp. on Biomedical Imaging: From Nano to Macro (ISBI '11)*, pp. 657–660 (2011).
22. F. Ghaznavi et al., "Digital imaging in pathology: whole-slide imaging and beyond," *Annu. Rev. Pathol. Mech. Dis.* **8**, 331–359 (2013).
23. S. Doyle et al., "Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer," *BMC Bioinf.* **13**, 282 (2012).
24. J. Hipp et al., "Integration of architectural and cytologic driven image algorithms for prostate adenocarcinoma identification," *Anal. Cell. Pathol.* **35**(4), 251–265 (2012).
25. E. Yu et al., "Detection of prostate cancer on histopathology using color fractals and probabilistic pairwise Markov models," in *IEEE Int. Conf. of Engineering in Medicine and Biology Society (EMBS '11)*, pp. 3427–3430 (2011).
26. A. Golugula et al., "Supervised regularized canonical correlation analysis: integrating histologic and proteomic measurements for predicting

biochemical recurrence following prostate surgery," *BMC Bioinf.* **12**, 483 (2011).
27. A. Madabhushi et al., "Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data," *Comput. Med. Imaging Graphics* **35**, 506–514 (2011).
28. S. Doyle et al., "A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. Biomed. Eng.* **59**, 1205–1218 (2012).
29. J. P. Monaco et al., "High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models," *Med. Image Anal.* **14**, 617–629 (2010).
30. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
31. S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Comput. Vision Graphics Image Process.* **39**(3), 355–368 (1987).
32. K. Nguyen et al., "Structure and context in prostatic gland segmentation and classification," *Lect. Notes Comput. Sci.* **7510**, 115–123 (2012).
33. S. B. Ginsburg et al., "Novel PCA-VIP scheme for ranking MRI protocols and identifying computer-extracted MRI measurements associated with central gland and peripheral zone prostate tumors," *J. Magn. Reson. Imaging* **41**, 1383–1393 (2015).
34. R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst, Man Cybern.* **3**(6), 610–621 (1973).
35. National Cancer Institute, "Home—The Cancer Genome Atlas-Cancer Genome-TCGA," 2016, http://cancergenome.nih.gov/.

Biographies for the authors are not available.