Supplemental-1

INFORMATIVE FEATURE SELECTION AND REDUCTION

Identification of informative image features is necessary for WSI prediction modeling. Dimensionality reduction in WSI prediction modeling is beneficial for the following reasons: (1) prediction modeling after dimensionality reduction can result in simpler models with higher prediction performance and (2) dimensionality reduction can provide insights about the data by highlighting important features or dimensions [1].To identify informative and robust image features, we can apply one of two techniques: (1) feature selection or (2) feature reduction. These techniques reduce the dimensionality of the feature space by removing irrelevant and redundant features to improve the performance of prediction modeling.

Feature selection methods can be broadly classified into three categories: filter, wrapper, and embedded methods [2]. Filter methods include univariate methods that filter features based on statistical properties (e.g., t-test, Wilcoxon rank sum test, ANOVA, and chi-square) [3] as well as multivariate methods that consider the effects of multiple interacting features (e.g., minimum redundancy maximum correlation (mRMR) [4], and relief-F [5]). Because filter methods are fast and scalable to high-dimensional data, they are often used in pathology informatics [6-9]. However, filter methods select features independent of the classifier; as such, they may not select optimal feature sets for a particular classifier. In contrast, wrapper methods generate various subsets of features using a deterministic or randomized search method and directly evaluate them with a classifier. Common wrapper methods are often coupled with a search method and include sequential forward search (SFS) [10], sequential backward elimination (SBE), randomized hill climbing [11], genetic algorithm [12], and simulated annealing [13]. Sequential search methods are commonly used in pathology informatics systems [14-16]. The drawbacks of wrapper methods include over-fitting and computational cost. Thus, embedded methods identify important features as intrinsic properties of a classifier (e.g., the weight vector of a SVM classifier [17] and the nodes of a random forest or tree classifier [18]). DiFranco et al. used random forest feature selection in their system for detecting regions of prostate tumor in WSIs [19].

Feature reduction techniques transform high-dimensional data into meaningful low-dimensional data. Ideally, reduced dimensionality should correspond to intrinsic dimensionality of data. In comparison to feature selection methods, feature reduction methods transform the original features instead of selecting an optimal feature subset. Moreover, they are unsupervised with the exception of linear discriminant analysis (LDA) methods. Feature reduction methods can be divided into two groups: (1) linear feature reduction techniques (e.g., principal component analysis (PCA), independent component analysis (ICA), factor analysis, and LDA) and (2) nonlinear feature reduction techniques including multidimensional scaling (MDS), ISOMAP, kernel PCA, local linear embedding (LLE), Laplacian Eigenmaps, and graph embedding [20, 21]. Because of the intuitive interpretation of PCA transformed features, it is one of the most commonly used feature reduction techniques in pathology informatics [16, 22, 23]. Besides PCA, researchers have also used graph embedding [24], ISOMAP [25], and MDS [26] for feature transformation in pathology informatics systems.

CLASSIFICATION

Classification methods commonly used in pathology informatics include k-nearest neighbors (k-NN) [14, 15, 23, 25, 27-29], support vector machines (SVM) [7, 9, 14-16, 23-25, 27, 30-35], Bayesian methods [8, 14-16, 23, 25, 27, 34, 36-38], neural networks [39, 40], decision trees [25, 41] and logistic regression [41]. Researchers often evaluate image features using multiple classifiers and report the best-performing classifiers [14, 15, 23, 25, 27]. In addition to basic classifiers, researchers in pathology informatics use boosting algorithms (i.e., combining a weighted set of weak classifiers to produce a robust classifier [25, 29, 36]) and ensemble methods that combine the decisions of multiple classifiers [15]. It is important to note that feature selection/reduction and classification should be conducted within a cross-validation framework, especially when evaluating systems for clinical prediction [42].

As the predictive accuracy of supervised learning methods depends on the quality of the training data, researchers are investigating methods for collecting and combining training data for predictive modeling. Training data can be collected using the following methods: (1) one-time annotation by a single

pathologist, (2) one-time annotation by multiple pathologists, and (3) run-time continuous annotation. Most informatics systems use the first method. However, the performance of these systems is subjective to the pathologist's knowledge. The second method for annotation requires a method for combining annotations from multiple experts [43]. The third method of annotation falls in the field of active learning or relevance feedback, where one or more pathologists provide active feedback to the learning algorithm in order to iteratively improve its knowledge [43-46]. Although active learning based algorithms may need a longer training phase, they have the potential to evolve into useful decision support systems for clinical applications.

Information Extraction for Case Study 1

WSIs provided by TCGA often have image artifacts. For quality control, we detect tissue-fold [47] and pen-mark [48] artifacts from the lowest resolution WSIs. Algorithm for information extraction from quality-controlled WSIs includes the following steps:

1. Crop the highest-resolution, *H*x*W*-pixel WSI of sample s of patient l, $\mathbf{I}^{s,l}$, into a matrix of non-overlapping, 512x512-pixel tiles. In total, a WSI is cropped into $J^{s,l}$ tiles, where $J^{s,l} = \left\lfloor \frac{H}{512} \right\rfloor \times \left\lfloor \frac{W}{512} \right\rfloor$.
2. Select tissue (excluding pen-mark and blank) tiles with less than 10% tissue-fold artifact.
3. Extract a vector of 461 image features, $\mathbf{x}_k^{s,l}$, capturing various pixel- and object-level image features (**Table S1**) from the $K^{s,l}$ tiles (usually $K^{s,l} << J^{s,l}$) that passed quality control. Therefore, the WSI is represented by $\mathbf{I}^{s,l} \approx \left\{ \mathbf{x}_k^{s,l} \middle| k \in \left[1, K^{s,l}\right] \right\}$.
4. Annotate tiles as tumor or non-tumor using a supervised classification model $f(\mathbf{x})$, trained using image features $\mathbf{x}_k^{s,l}$ of manually annotated tumor $\left(A_k^{s,l} = 1\right)$ and non-tumor $\left(A_k^{s,l} = 0\right)$ tiles from 17 WSIs (**Figure 5B**) [48].

$$A_k^{s,l} = f\left(\mathbf{x}_k^{s,l}\right),$$

where $A_k^{s,l} = \{0,1\}$, $f(\mathbf{x})$ is a linear SVM classifier model that returns 1 if $\mathbf{x}_{461\times1}$ is similar to tumor tile features in the training data.

5. Represent patients by combining image features, $\mathbf{x}_k^{s,l}$, of all tumor tiles $\left(A_k^{s,l}=1\right)$ in all of the samples $\left(s \in [1, S^l]\right)$ of a patient l.

   $p_i^l = C_i\left(\left\{x_{i,k}^{s,l} \middle| s \in [1,S^l], k \in [1,K^{s,l}], A_k^{s,l}=1\right\}\right)$, where p_i^l is the feature i for patient l.

   The tile-feature combination function, C_i, depends on the type of feature i. As tile size is constant, we represent a WSI by simply averaging all pixel-level tile features except for Haralick and fractal features, where we sum co-occurrence matrices and histograms for tiles, respectively, and then calculate the statistics. For combining object-level tile features, we assume that objects in a tile are a subset of all objects in a patient's samples and combine features using group statistics accounting for the number of objects in each tile. For comparison, we also extract and combine image features of all tissue tiles including non-tumor regions (without step 3).

Table S1: Image features extracted from WSIs

Feature Subset	**Number of Features**	**Feature Description**
Color	73	RGB histograms, histogram statistics, and stain co-occurrence
Global texture	138	Haralick, gray-level histogram statistics, fractal, GHM multiwavelet, and Gabor
Eosinphilic-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Eosinphilic-region texture	18	Haralick and gray-level histogram statistics

No-stain-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Basophilic-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Basophilic-region texture	18	Haralick and gray-level histogram statistics
Nuclear shape	26	Count, elliptical area, major-minor axes lengths, eccentricity, and cluster size
Nuclear topology	35	Delaunay triangle, Voronoi diagram, minimum spanning tree, and closeness

Multi-Resolution Representation for Exploratory Analysis of WSIs

In this case study, we present a visualization framework for exploring spatial patterns in WSIs at multiple resolutions using an unsupervised segmentation algorithm, Statistical Region Merging (SRM) [49, 50]. SRM quantizes the color space of an image while simultaneously merging similar neighboring regions. With a low number of quantization levels, it can provide a low-detail representation of an image with key landmarks. As the quantization level is increased, we can observe a more detailed representation of an image. Thus, this SRM-based visualization follows Ben Schneiderman's mantra: "overview first, zoom in and filter, details on demand" [51]. In this case study, we process a WSI of an ovarian serous carcinoma sample from TCGA [52] (**Figure S1**). **Figures S1A-S1D** illustrates SRM results for four scenarios where WSIs at different resolutions were processed with different SRM-quantization levels. Such exploration can help in selecting the appropriate resolution and quantization of a WSI for a particular image processing application. For example, we found that the four scenarios would be useful for the following applications. The scenario in **Figure S1D** is useful for separation of tissue from background. The scenario in **Figure S1C** is useful for identification of tissue folds, tumor regions, and non-tumor regions. The scenario in **Figure S1B** is useful for cell counting. The scenario in **Figure S1A** is useful for shape analysis. While

applying SRM on WSIs, we observe the following: (1) SRM can process WSIs in a reasonable time, (2) SRM-processed images are compressed and thus, are faster to load, and (3) WSIs can be quickly annotated by Scalable Vector Graphics (SVG) files produced by the SRM software, and (4) processing WSIs at higher resolution is computationally intensive (**Figure S1E**). Many researchers start with full-resolution data to solve a problem (red point to the right of D in **Figure S1E**). However, starting at a different resolution and/or level of detail (points A-C in **Figure S1E**) can save computation time. We observe considerable decrease in file size and processing time for the four scenarios in **Figure S1E**. The source data for this case study can be found on a tissue imaging wiki site [53].

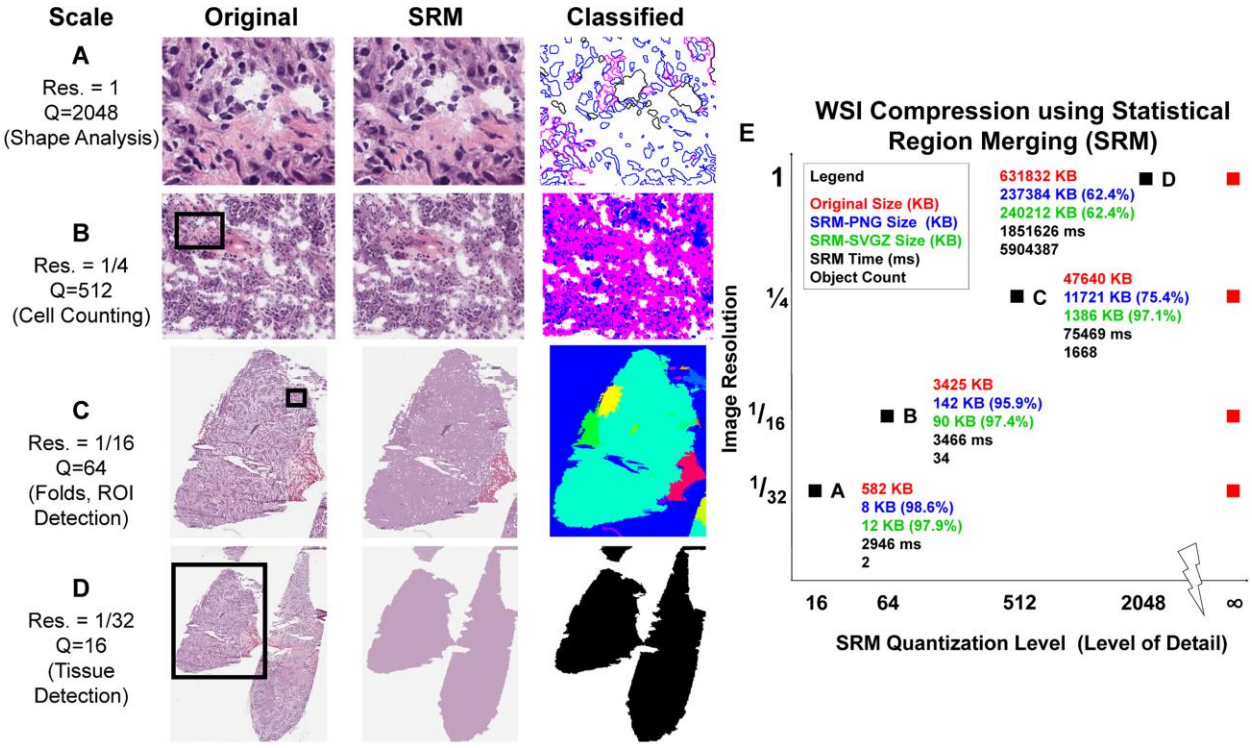

Figure S1: TCGA ovarian cancer image segmented using SRM at various resolutions matched to level of detail (Q). (A-D) The SRM column shows the similarity of the color quantized image to the original image. The 'Classified' outlines distinct shapes in the images. The suggested purpose of each processing result is given in parentheses in the 'Scale' column. (E) Plot of the relationships between resolution and Q for the case study (Both axes are on logarithmic scale). The first two file sizes given (red and blue) are the sizes of PNG-format images, and the third (green) is a compressed ZIP format of a descriptive vector graphics XML file.

REFERENCES

1. Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003;3:1157-82.

2. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507-17.

3. Hall MA. Correlation-based feature selection for machine learning: Waikato University, New Zealand; 1999.

4. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol. 2005;3(2):185-205.

5. Kononenko I. Estimating attributes: Analysis and extensions of relief. In: Bergadano F, De Raedt L, editors. Machine learning: Ecml-94: Springer Berlin / Heidelberg; 1994. p. 171-82.

6. Kothari S, Phan JH, Stokes TH, et al. Pathology imaging informatics for quantitative analysis of whole-slide images. J Am Med Inform Assoc. under review.

7. Cruz-Roa A, Caicedo JC, González FA. Visual pattern mining in histology image collections using bag of features. Artif Intell Med. 2011;52(2):91.

8. Muthu Rama Krishnan M, Pal M, Paul RR, et al. Computer vision approach to morphometric feature analysis of basal cell nuclei for evaluating malignant potentiality of oral submucous fibrosis. J Med Syst. 2012;36(3):1746-56.

9. Bilgin CC, Bullough P, Plopper GE, et al. Ecm-aware cell-graph mining for bone tissue modeling and classification. Data Min Knowl Discov. 2009;20(3):416-38.

10. Kittler J. Feature set search algorithms. Pattern recognition and signal processing. 1978:41-60.

11. Skalak DB. Prototype and feature selection by sampling and random mutation hill climbing algorithms. Proc conf on machine learning. 1994:293-301.

12. Holland JH. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. Ann Arbor: Michigan University; 1975.

13. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983;220(4598):671.

14. Po-Whei H, Cheng-Hsiung L. Automatic classification for pathological prostate images based on fractal analysis. IEEE Trans Med Imaging. 2009;28(7):1037-50.

15. Kong J, Sertel O, Shimada H, et al. Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. Pattern Recognition. 2009;42(6):1080-92.

16. Srivastava S, Rodríguez JJ, Rouse AR, et al. Computer-aided identification of ovarian cancer in confocal microendoscope images. J Biomed Opt. 2008;13(2):024021.

17. Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46(1):389-422.

18. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.

19. DiFranco MD, O'Hurley G, Kay EW, et al. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. Comput Med Imaging Graph. 2011;35:629-45.

20. Yan S, Xu D, Zhang B, et al. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell. 2007;29(1):40-51.

21. van der Maaten L, Postma E, van den Herik J. Dimensionality reduction: A comparative review. Tilburg University Technical Report, TiCC-TR 2009-005 [serial on the Internet]. 2009: Available from: file:///var/www/html/librarian/library/00277.pdf.

22. Rahman M, Bhattacharya P, Desai BC. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. IEEE Trans Inf Technol Biomed. 2007;11(1):58-69.

23. Sertel O, Kong J, Catalyurek U, et al. Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading. J of Signal Processing Syst. 2009;55(1):169-83.

24. Basavanhally AN, Ganesan S, Agner S, et al. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. IEEE Trans Biomed Eng. 2010;57(3):642-53.

25. Yang L, Chen W, Meer P, et al. Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. IEEE Trans Inf Technol Biomed. 2009;13(4):636-44.

26. Lei Z, Wetzel AW, Gilbertson J, et al. Design and analysis of a content-based pathology image retrieval system. IEEE Trans Inf Technol Biomed. 2003;7(4):249-55.

27. Tabesh A, Teverovskiy M, Ho-Yuen P, et al. Multifeature prostate cancer diagnosis and gleason grading of histological images. IEEE Trans on Med Imaging. 2007;26(10):1366-78.

28. Jafari-Khouzani K, Soltanian-Zadeh H. Multiwavelet grading of pathological images of prostate. IEEE Trans Biomed Eng. 2003;50(6):697-704.

29. Foran DJ, Yang L, Chen W, et al. Imageminer: A software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. J Am Med Inform Assoc. 2011;18(4):403-15.

30. Caicedo JC, González FA, Romero E. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. J Biomed Inform. 2011;44(4):519-28.

31. Raza S, Parry R, Moffitt R, et al. An analysis of scale and rotation invariance in the bag-of-features method for histopathological image classification. In: Fichtinger G, Martel A, Peters T, editors. Med image comput comput assist interv 2011: Springer Berlin / Heidelberg; 2011. p. 66-74.

32. Cooper LAD, Jun K, Gutman DA, et al. An integrative approach for in silico glioma research. IEEE Trans Biomed Eng. 2010;57(10):2617-21.

33. Yang L, Tuzel O, Chen W, et al. Pathminer: A web-based tool for computer-assisted diagnostics in pathology. IEEE Trans Inf Technol Biomed. 2009;13(3):291-9.

34. Boucheron L. Object-and spatial-level quantitative analysis of multispectral histopathology images for detection and characterization of cancer: PhD thesis, University of California, Santa Barbara; 2008.

35. Celebi ME, Kingravi HA, Uddin B, et al. A methodological approach to the classification of dermoscopy images. Comput Med Imaging Graph. 2007;31(6):362-73.

36. Doyle S, Feldman M, Tomaszewski J, et al. A boosted bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies. IEEE Trans Biomed Eng. 2010;59(5):1205-18.

37. Chaudry Q, Raza S, Young A, et al. Automated renal cell carcinoma subtype classification using morphological, textural and wavelets based features. J of Signal Processing Syst. 2009;55(1):15-23.

38. Sudbo J, Bankfalvi A, Bryne M, et al. Prognostic value of graph theory-based tissue architecture analysis in carcinomas of the tongue. Lab Invest. 2000;80(12):1881-9.

39. Tang HL, Hanka R, Ip HHS. Histological image retrieval based on semantic content analysis. IEEE Trans Inf Technol Biomed. 2003;7(1):26-36.

40. Schnorrenberg F, Pattichis CS, Schizas CN, et al. Content-based retrieval of breast cancer biopsy slides. Technol Health Care. 2000;8(5):291-7.

41. Fuchs T, Wild P, Moch H, et al. Computational pathology analysis of tissue microarrays predicts survival of renal clear cell carcinoma patients. In: Metaxas D, Axel L, Fichtinger G, Székely G, editors. Med image comput comput assist interv 2008: Springer Berlin / Heidelberg; 2008. p. 1-8.

42. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. Int J Med Inform. 2008;77(2):81-97.

43. Fuchs TJ, Buhmann JM. Computational pathology: Challenges and promises for tissue analysis. Comput Med Imaging Graph. 2011;35(7-8):515-30.

44. Doyle S, Monaco J, Feldman M, et al. An active learning based classification strategy for the minority class problem: Application to histopathology annotation. BMC Bioinformatics. 2011;12:424.

45. Cosatto E, Miller M, Graf HP, et al. Grading nuclear pleomorphism on histological micrographs. Proc Int Conf Pattern Recogn. 2008:1-4.

46. Begelman G, Pechuk M, Rivlin E, et al. A microscopic telepathology system for multiresolution computer-aided diagnostics. J Multimed. 2006;1(7).

47. Palokangas S, Selinummi J, Yli-Harja O. Segmentation of folds in tissue section images. Conf Proc IEEE Eng Med Biol Soc. 2007;2007:5642-5.

48. Kothari S, Phan JH, Osunkoya AO, et al. Biological interpretation of morphological patterns in histopathological whole-slide images. Proceedings of the 3rd ACM Conference on Bioinformatics, Computational Biology and Biomedicine. 2012:218-25.

49. Nock R. Fast and reliable color region merging inspired by decision tree pruning. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2001;1:271-6.

50. Nock R, Nielsen F. Statistical region merging. IEEE Trans Pattern Anal and Mach Intell. 2004 Nov;26(11):1452-8.

51. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. Proc IEEE Symp Visual Languages. 1996:336-43.

52. The cancer genome atlas data portal. Available from: https://tcga-data.nci.nih.gov/tcga/.

53. Tissue imaging wiki site.; Available from: http://tissuewiki.bme.gatech.edu/index.php/TCGA-04-1338.