Predicting lymph node metastasis status in primary breast carcinoma via image analysis of tumor histology

Mark D. Zarella[1], Md. Alimoor Reza[2], Aladin Milutinovic[1], Robi Polikar[3], David E. Breen[2], Fernando U. Garcia[1]

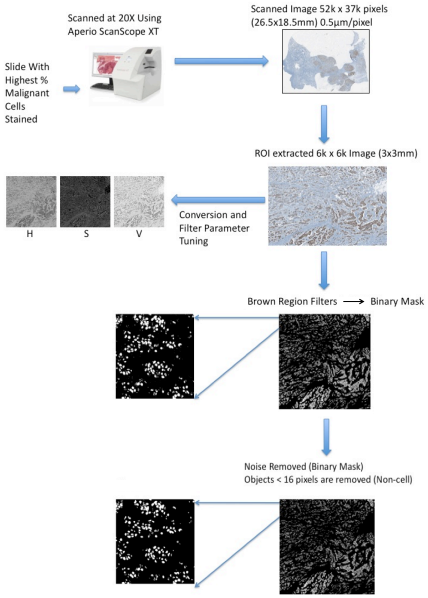
1. Pathology Department, Drexel University College of Medicine, Philadelphia, PA, USA. 2. Department of Computer Science, College of Engineering, Drexel University, Philadelphia, PA, USA.
3. Department of Electrical and Computer Engineering, Rowan University

Motivation

Axillary lymph node metastasis status remains one of the most critical prognostic variables for breast cancer management decision-making and patient survival. Methods for determining metastasis status of breast carcinoma need improvement in order to avoid unnecessary surgeries and complications. The objective of our study is to demonstrate that lymph node metastasis status may be predicted via computerized image analysis of primary breast tumor histology. The procedure described here is comprised of four steps.

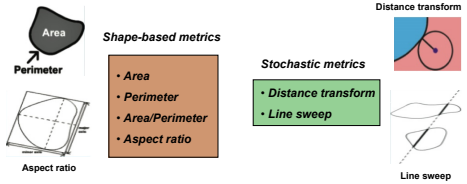
- Step 1: Apply a segmentation algorithm to identify individual cell nuclei.
- Step 2: Characterize nucleus structure with a set of 13 metrics that describes the shape and color information of the sample.
- Step 3: Represent the distributions in a reduced form.
- Step 4: Implement a two-stage classifier to predict metastasis status based on image features.

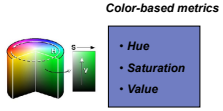
Step 1: Image acquisition and segmentation



- Whole-slide high resolution (0.5 μm/pixel) RGB images of hematoxylin and eosin stained tissue were acquired from 100 patients diagnosed with Invasive Mammary Carcinoma. The 100 specimens consisted of 47 N1 (positive for lymph node metastasis) cases and 53 N0 (negative for lymph node metastasis) cases.
- To extract information about individual cell nuclei, we implemented a segmentation procedure in which a threshold was applied to the image after first being passed through a hue-intensity filter previously optimized for segmentation (1).
- Groups of pixels that were too small to be consistent with typical nucleus dimensions were considered artifacts and discarded.

Step 2: Feature extraction





- In order to quantitatively represent the properties of the nuclei, a set of information from each segmented nucleus was computed (2).
- *Shape-based* metrics describe the basic geometric properties of the nucleus.
- *Stochastic* metrics employ stochastic geometry to describe the anisotropic nature of nucleus shapes and their spatial distribution in a probabilistic manner.
- *Color-based* metrics transform the RGB properties of each nucleus into HSV space, allowing us to separately evaluate the hue, saturation, and pixel intensity information.
- A total of 13 metrics were used, consisting of the normalized metrics listed above as well as the log-transformed and large-scale structure of the stochastic measures.

Step 3: Dimensionality reduction

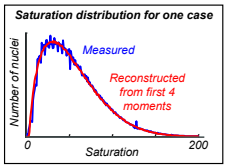
First four moments

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

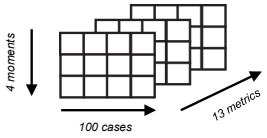
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$Skew = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^3}{\sigma^3}$$

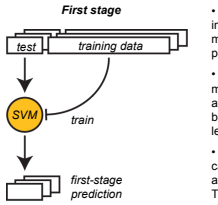
$$Kurtosis = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^4}{\sigma^4}$$



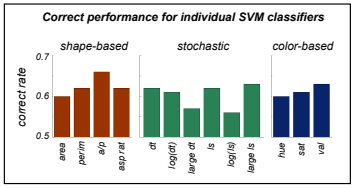
- A distribution of values was derived for each metric, where each value corresponded to the individual cell nuclei in the image.
- Distributions were replaced by the first four moments (mean, variance, skew, and kurtosis) which described most of the information within the distribution.
- This procedure produced a data set with a dimensionality appropriate for the classification stage. For the 100 cases analyzed, 4 moments were generated for each of the 13 metrics.



Step 4: Classification



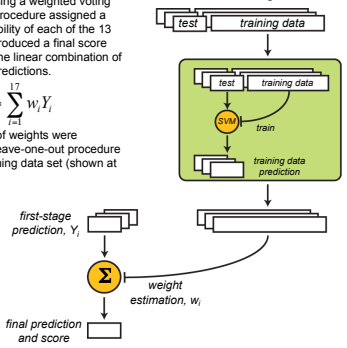
- The 4 moments from each metric were input individually to a support vector machine (SVM) classifier to form a set of 13 predictions.
- SVM creates a decision boundary in a multidimensional data space, produced from a set of training data. Subsequent data can be classified as N0 or N1 based on the learned decision boundary.
- 13 predictions were produced for all 100 cases by training the classifier with 99 cases and evaluating the remaining "test" case. This process was repeated 100 times, choosing a different test case each time.

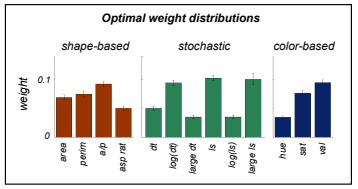


- The 13 predictions from the first stage were combined using a weighted voting procedure. This procedure assigned a weight to the reliability of each of the 13 predictions, and produced a final score that consisted of the linear combination of the weighted 13 predictions.

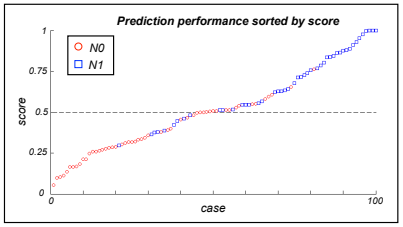
$$Score = \sum_{i=1}^{17} w_i Y_i$$

- The optimal set of weights were determined by a leave-one-out procedure applied to the training data set (shown at the right in green).

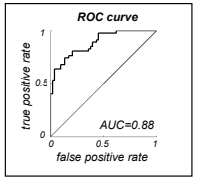


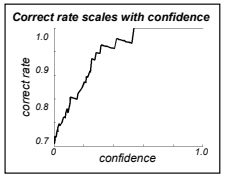


Performance



- The computed set of optimal weights indicate that shape-based, stochastic, and color-based information all contribute to the prediction.
- Of the 100 cases tested, 76 were classified correctly. This is significantly more than expected by chance ($p<10^{-7}$).
- The area under the ROC curve (AUC) was 0.885. Overall detection performance of >95% can be achieved if a false positive rate of 45% is accepted.





- The classification score provided valuable information about the confidence of the prediction. The correct rate rose to over 95% for the 54 most confident scores, and achieved 100% success for the 39 most confident scores.

$$Confidence = 2\times|Score-0.5|$$

Conclusion and future direction

- We have shown that lymph node metastasis status can be predicted from whole-slide histology images by implementing an automated segmentation and feature extraction routine, and adopting a machine learning approach.
- Using a two-stage cross-validation scheme, we quantified the capabilities of this algorithm. Moreover, we revealed that the scalar output of the system (the "classification score") could provide a confidence measurement to accompany the system's prediction, producing high quality results in the most confident cases. Characterizing the attributes of the most confident cases is a potential avenue of future exploration.
- We demonstrated that shape-based, stochastic, and color-based metrics contributed to the prediction, although to varying degrees. The most predictive nucleus features were those which described the irregularity of its contours (e.g. area/perimeter ratio, line sweep), as well as the saturation and intensity values. These results will help guide future investigations into the predictive quality of image features.

References

1) S. Petushi, J. Zhang, A. Milutinovic, D.E. Breen, F.U. Garcia, "Image-Based Histologic Grade Estimation Using Stochastic Geometry Analysis," Medical Imaging 2011: Computer-Aided Diagnosis, SPIE Proceedings, Vol. 7963, paper 79633E, March 2011.

2) J.Z. Zhang, S. Petushi, W.C. Regli, F.U. Garcia and D.E. Breen, "A Study of Shape Distributions for Estimating Histologic Grade," Proc. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, August 2008, pp. 1200-1205.