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

Mark D. Zarella, Md. Alimoor Reza, Aladin Milutinovic, Robi Polikar, David E. Buen, Fernando U. Garcia
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

Adjuvant 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 variables for breast cancer management decision-making and patient survival. Axillary lymph node metastasis status remains one of the most critical prognostic dimensions were considered artifacts and discarded.

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.

Step 2: Feature extraction

The 100 cases were split into two sets of approximately equal size: 54 training cases and 46 test cases. The metrics were defined as:

- Shape-based metrics:
  - Area
  - Perimeter
  - Area/Perimeter
  - Aspect ratio

- Stochastic metrics:
  - Distance transform
  - Line sweep

- Color-based metrics:
  - Hue
  - Saturation
  - Value

- In order to quantitatively represent the properties of the nuclei, a set of information from each segmented nucleus was computed (2).

Stochastic metrics employ stochastic geometry to describe the anatomic nature of nuclear 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

A distribution of values was derived for each metric, where each value corresponded to the individual cell nucleus 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.

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).

Step 4: Classification

- The 13 predictors 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.

- The computed set of optimal weights indicated that shape-based, stochastic, and color-based information all contribute to the prediction.

- Of the 100 cases tested, 76 were correctly classified. This is significantly more than expected by chance (p < 0.001)

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 28 most confident scores.

Step 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 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 for future exploration.

We demonstrated that shape-based, stochastic, and color-based metrics contributed to the prediction, although to varying degrees. The most predictive nuclear features were those which described the irregularity of its contours (e.g., area/perimeter ratio, line sweeping), as well as the saturation and intensity values. These results will help guide future investigations into the predictive quality of image features.

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 < 0.001)

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 28 most confident scores.

References