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

David E. Breen1, Rian Hu1, Md. Almoor Reza2, Aladin Milutinovic1, Robi Polikar3, and Fernando U. Garcia2
1Drexel University, Philadelphia, PA
2Drexel University College of Medicine, Philadelphia, PA
3Rowan University, Glassboro, NJ

Goal / Motivation
Predict metastasis status of breast carcinoma via image analysis of primary tumor

Justification / Background
- Pathologists assess histologic grade by studying the variations in cell and tumor morphology and chromasia
- Histologic grade is correlated with axillary lymph node metastasis
- Stochastic geometry, morphometry and texture analysis are able to capture differences in shape and color
- Can we find shape and color features in the primary tumor which indicate that axillary metastasis has occurred?

Approach
- Computational Pipeline that includes
  - Selection and scanning of stained carcinoma histology slides
  - Automated segmentation of cancer cells and tumor structures
  - Geometric measures computation
  - Intensity distributions computation
  - Dimension reduction to produce feature vectors
  - Classification via stacked Relevance Vector Machines (RVMs)
  - 2-stage cross validation

Computational Pipeline

Selection & Scanning
- 100 stained samples are chosen by pathologist
- Samples with highest percentage of nuclear expression (as indicated by image analysis, Spectrum, Aperio)
  - ER, PR, MIB-1 and p53
- Full slide scan at 50,000 x 40,000
  - 0.5 μm / pixel
**ROI Extraction**

- Region-of-Interest is identified by a pathologist
- Area with highest density of stained carcinoma cells
- 6,000 x 6,000 resolution — 3 mm x 3 mm region

**Segmentation**

- ROI Image separated into Hue, Saturation and Value channels
- Brown-stained nuclei segmented in HV space
  - 320 ≤ H ≤ 340, 0 ≤ S ≤ 90, V ≤ 0.8
  - Red through yellow-green in Hue space
- Produces a binary image that identifies original brown pixels

**Filtering and Processing**

- Area-based filtering is applied
  - Removes white regions smaller than 17 pixels
  - Fills in black regions smaller than 17 pixels
- Morphological closing (structuring element radius = 8) applied to binary image
- Produces two binary images per sample
  - Malignant cell nuclei
  - Large-scale groupings of cells

**Shape Distribution Generation**

- Geometric measures are applied to binary images
  - Cell-based (per blob)
    - Area, perimeter, area/perimeter, aspect ratio
  - Stochastic geometry-based (ROI image)
  - Radial contact and line contact
- Calculations produce distributions

**Color Intensity Distribution Generation**

- Binary image is used as a mask
- Histograms that capture the pixel intensity distributions within each channel (HSV) are computed for color pixels corresponding to the white pixels in the binary image
- 3 Distributions are produced per sample
  - Hue, Saturation and Value

**Distribution Generation**

- 11 distributions are produced per sample at this stage
  - Area, perimeter, area/perimeter, aspect ratio, hue, saturation and value (cell-level)
    - radial contact, line contact (cell-level & large-scale)
- Distributions widths vary greatly
  - 320 to 2000s at bins
- Individual bin values vary greatly
  - 1 to millions
- Natural log is applied to radial/line contact distributions
- Total number of distributions : 13
**Study Results**

- 100 stained breast carcinomas (with known metastasis status) were scanned and processed with our pipeline:
  - 47 metastatic (N1) samples
  - 53 non-metastatic (NO) samples
- Correctly classified:
  - All 53 NO samples
  - 37 of the 47 N1 samples
- Specificity: 100%
- Sensitivity: 79%
- Positive Predictive Value: 100%
- Negative Predictive Value: 84%
- Overall Accuracy: 90%

**Study Results (cont.)**

- Area under the curve: 0.840

---

**Supervised Machine Learning**

- 2-stage “stacked” Relevance Vector Machines (RVMs) perform supervised machine learning / classification of the feature vectors.

---

**Dimension Reduction**

- Each bin of the distributions represents a dimension.
- Need feature vectors with a smaller dimension.
- Compute 10 values from each distribution:
  - Mean, median, mode, standard deviation, skew, kurtosis, distribution width, max bin value, mean and standard deviation of bin values.
- For distributions with a fixed width (aspect ratio and HSV), average distributions down to 10 bins.
- Produces 17 10-dimensional feature vectors for each sample.

**Cross Validation**

- Leave-One-Out cross validation is used for each stage of the stacked classifier.
- Each stage of the classifier is validated separately.

---

**Supervised Machine Learning**

- First stage consists of 17 RVMs, one for each feature vector.
- Second stage combines intermediate prediction (a 17-D feature vector) to produce final classification of a sample.
Conclusions & Future Work

• Shape and color analysis of primary breast tumors shows promise as an effective means to determine the absence of nodal metastasis
• May obviate the need for axillary dissection in some cases
• More study is needed
  – Will process more cases
  – Improve cross-validation
  – Identify most relevant features

Acknowledgments

• Funding provided by
  – Commonwealth of Pennsylvania Tobacco Fund
  – U.S. Army Medical Research Acquisition Activity; Cooperative Agreement W81XWH 04–1–0419
• Carcinoma samples provided by the Drexel College of Medicine Breast Cancer Databank