Breast Cancer Research

RESEARCH

Open Access

Image analysis-based identification of high risk ER-positive, HER2-negative breast cancers

Dong Neuck Lee[1], Yao Li[2], Linnea T. Olsson[3], Alina M. Hamilton[4], Benjamin C. Calhoun[6], Katherine A. Hoadley[5], J. S. Marron[2*] and Melissa A. Troester[3*]

Abstract

Background Breast cancer subtypes Luminal A and Luminal B are classified by the expression of PAM50 genes and may benefit from different treatment strategies. Machine learning models based on H&E images may contain features associated with subtype, allowing early identification of tumors with higher risk of recurrence.

Methods H&E images (n = 630 ER+/HER2- breast cancers) were pixel-level segmented into epithelium and stroma. Convolutional neural network and multiple instance learning were used to extract image features from original and segmented images. Patient-level classification models were trained to discriminate Luminal A versus B image features in tenfold cross-validation, with or without grade adjustment. The best-performing visual classifier was incorporated into envisioned diagnostic protocols as an alternative to genomic testing (PAM50). The protocols were then compared in time-to-recurrence models.

Results Among ER+/HER2- tumors, the image-based protocol differentiated recurrence times with a hazard ratio (HR) of 2.81 (95% CI: 1.73–4.56), which was similar to the HR for PAM50 (2.66, 95% CI: 1.65–4.28). Grade adjustment did not improve subtype prediction accuracy, but did help balance sensitivity and specificity. Among high grade participants, sensitivity and specificity (0.734 and 0.474, respectively) became more similar (0.732 and 0.624, respectively) in grade-adjusted models. The original and epithelium-specific images had similar performance and highest accuracy, followed by stroma or binarized images showing only the epithelial-stromal interface.

Conclusions Given low rates of genomic testing uptake nationally, image-based methods may help identify ER+/HER2- patients who could benefit from testing.

Keywords Breast cancer, CBCS3, Histology, Multiple instance learning, Image segmentation, Distance weighted learning

*Correspondence:
J. S. Marron
marron@unc.edu
Melissa A. Troester
troester@unc.edu
[1] Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA
[2] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC, USA
[3] Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA
[4] Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA
[5] Department of Genetics, University of North Carolina, Chapel Hill, NC, USA
[6] Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC, USA

Background

Breast cancer heterogeneity motivates a precision medicine approach. PAM50 subtypes and risk of recurrence (ROR) scores identify intrinsic molecular subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like, or Normal-like) or recurrence risk, respectively, and can provide prognostic information to aid chemotherapy decisions [1–3]. Among ER-positive (ER+), HER2-negative (HER2-) tumors, high ROR tumors are more aggressive, have higher relapse rates, and are more likely to benefit from chemotherapy compared to low ROR/Luminal A tumors[2–4]. However, genomic testing takes time and can be costly and may not be universally available to all patients. Our recent findings suggest that among eligible ER+/HER2-breast cancer patients, a minority (~ 40%) received prognostic or predictive genomic testing [5]. In contrast, hematoxylin and eosin (H&E)-stained biopsy slides are routinely collected in every patient during diagnostic workup [6]. We hypothesized that such slides could be used to predict genomic risk of recurrence scores among ER+/HER2-tumors, thereby enabling the identification of patients who may benefit from genomic testing.

We envisioned three risk-stratification protocols among non-metastatic, ER+/HER2-breast cancers, based on stage, histological and/or genomic factors. In a reference protocol (Fig. 1. **A**) most similar to current guidelines, patients with advanced stage (stage 3) were considered higher-risk and patients with lower stage disease (1, 2) were recommended to receive genomic testing. Current guidelines recommend such tests for node negative or positive (1–3 nodes), early stage patients. For comparison with this approach, we envisioned an image-based approach that stratified risk only on AI-models trained on histology (**B**). Finally, we envisioned a hybrid image-plus-genomics testing where genomic testing was applied for lower-stage tumors with histologically-predicted high risk (**C**). Comparisons of these approaches offer insights about efficient application of histologic and genomic methods in resource scarce contexts.

Methods

Study population

The primary data for this study were from the Carolina Breast Cancer Study Phase 3 (CBCS3, 2008–2013), a population-based study of breast cancer in 44 counties of North Carolina. The study oversampled younger women (<50 years old) and Black women, resulting in a diverse study population. This analysis utilized hematoxylin and eosin (H&E)-stained TMA images, excluding participants who did not have a scanned H&E stained image (n = 993), who did not have research-based genomic

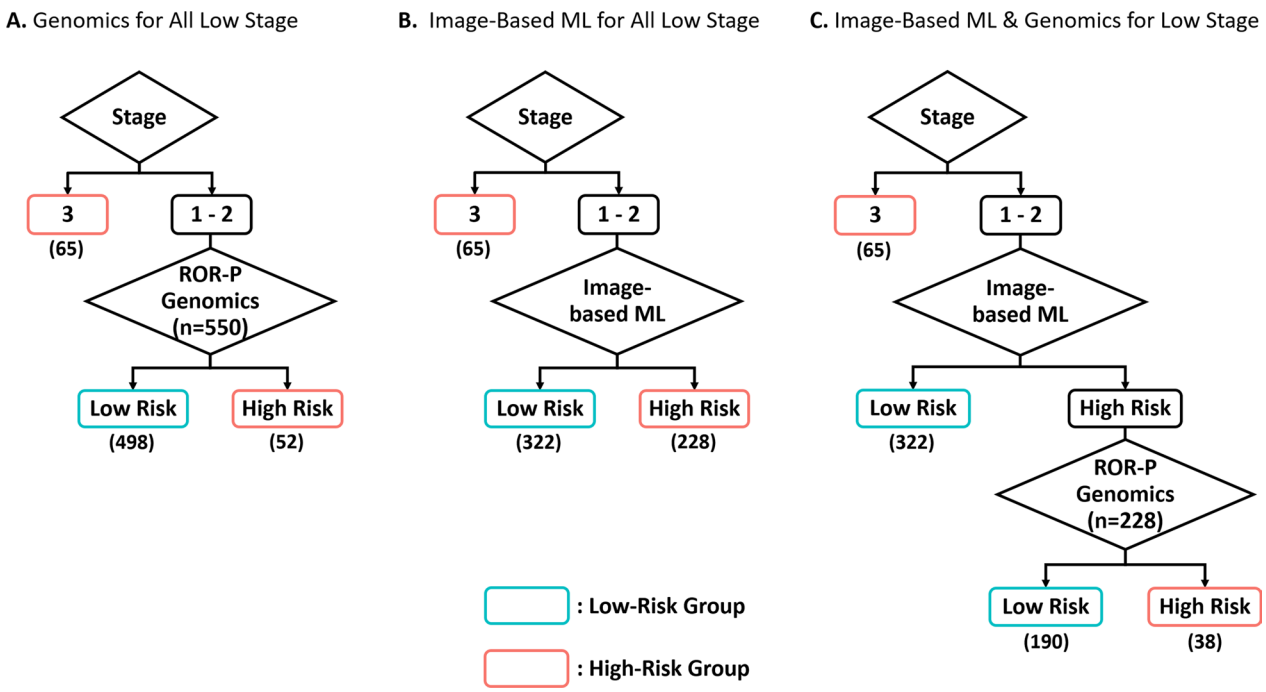


Fig. 1 Scenarios for Screening Low Stage ER+/HER2-cancers. **A** A conceptualized reference breast cancer diagnosis protocol containing genomic testing. **B** An alternative protocol which can be implemented by replacing genomic testing with a machine learning model. **C** A hybrid protocol which recommends genomic testing only for potential high risk breast cancer patients as predicted by the machine learning model. The numbers below boxes represent the corresponding patient count from the CBCS cohort

test results (n = 708), and who were missing grade (n = 47). We also excluded those who were ER- (n = 239) or who were HER2+ (n = 136) as these patients are not the target population for existing prognostic and predictive genomic tests. The final study population included 630 ER+/HER2-participants. The distribution of the participants is shown in Table 1.

To evaluate the generalizability of our findings in CBCS3 data, we employed the Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) dataset as an external validation group [7, 8]. This dataset includes FFPE diagnositic H&E-stained whole slide images (WSIs) and corresponding clinical data. For this analysis, we focused on a subset of 635 ER+/HER2-patients.

We aimed to distinguish between Luminal A (n = 408) and Luminal B (n = 156) among lower stage, genomic-testing eligible tumors. Therefore, we excluded participants with other breast cancer subtypes: Basal-like (n = 42), HER2-enriched (n = 7), and Normal-like (n = 17) in training. After developing the best-performing machine learning model for discriminating between Luminal A and Luminal B subtypes, we applied the model to all 630 ER+/HER2-participants, regardless of subtype, to classify them into either the low (Luminal A) or high-risk (Luminal B) group.

Image preprocessing and patch-level feature extraction

We used 20× scanned images from formalin-fixed paraffin embedded (FFPE) histologic tissue microarrays (CBCS3) and whole slide diagnostic images (TCGA-BRCA). The total number of core images from 630 CBCS participants was 2260, and the numbers of pixels varied. The median dimension is 2600 × 2600 pixels. To ensure consistent image processing across datasets, we utilized 4053 pre-selected 2000 × 2000 pixel regions from the 635 TCGA WSIs, with one WSI per patient [9]. Figure 2 illustrates the overall workflow of constructing subject-level feature vectors from core images. First, to reduce stain intensity variations by slides, every core is stain-normalized as described [10, 11] (Fig. 2. A). Next, we segmented the core images into epithelium and collagenous stroma regions. To investigate shapes of the two tissue types, we constructed binary images where epithelium and collagenous stroma regions are colored red and green, respectively (Fig. 2B). This process is discussed in detail in Supplementary Information S1. We divided

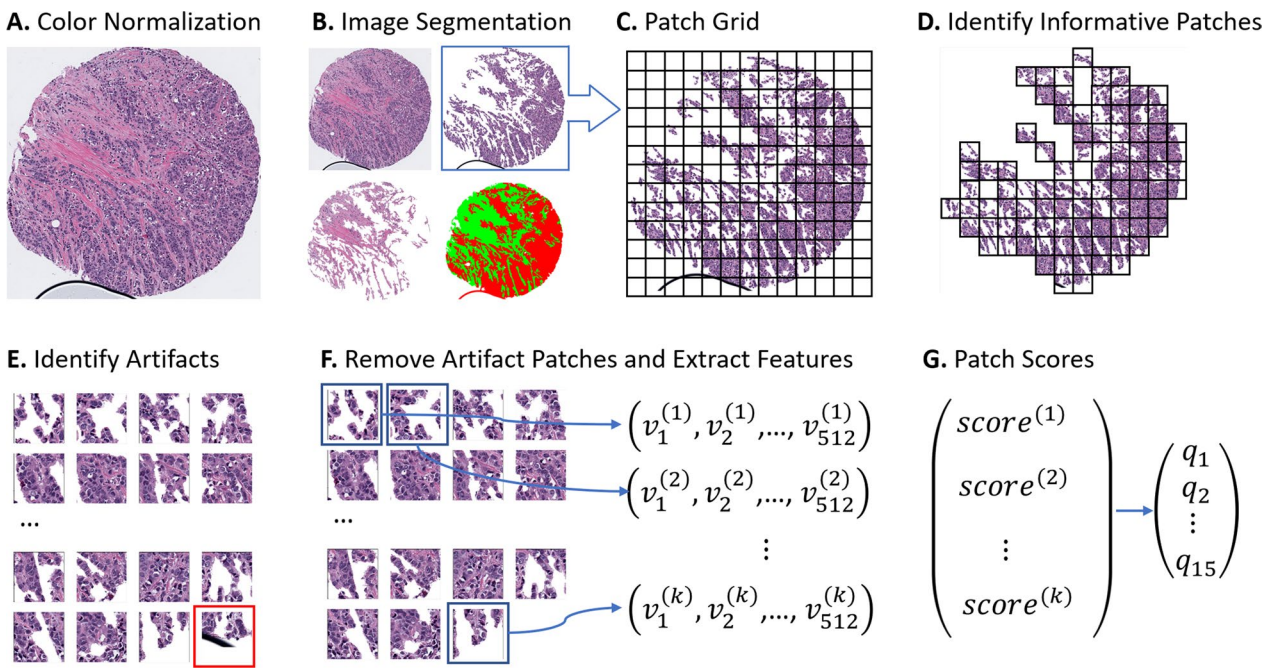


Fig. 2 Pipeline for extracting a 15-dimensional feature vector from a core image. This figure illustrates the process for an epithelial image among four types of segmented core images shown in **B**. **A** Every core was stain normalized to reduce stain intensity variations by slides. **B** The color-normalized H&E core image was separated into two tissue types, epithelium and collagenous stroma, using pixel-level image segmentation. Additionally, we constructed binary images to investigate the regional shape of the epithelium and collagenous stroma. **C** We divided each core into *k* patches with a size of 200 × 200 pixels. **D** Non-informative patches with background pixels above a patch-specific threshold were excluded. **E** Patches with artifacts were excluded by a trained artifact detector. **F** Image features were extracted from each informative patch using the convolutional layers of the pre-trained VGG16 architecture. **G** A one-dimensional patch score was calculated by projecting the patch features in the estimated direction that discriminates between Luminal A and Luminal B subtypes. To construct the core-level image feature vector, we summarize the *k* patch scores into 15 equally spaced quantiles

Table 1 Characteristics of ER-positive/HER2-negative breast cancer in the carolina breast cancer study phase 3, 2008–2013

	Total (*N* = 630)	Luminal A (*N* = 408)	Luminal B (*N* = 156)	HER2-enriched (*N* = 7)	Basal-like (*N* = 42)	Normal-like (*N* = 17)
Age of diagnosis						
Median Age (range), years	50 (24–74)	52 (24–74)	48 (28–74)	56 (36–72)	47 (30–74)	56 (42–74)
Age < 50, *n* (%)	309 (49.0)	185 (45.3)	88 (56.4)	3 (42.9)	29 (69.0)	4 (23.5)
Age ≥ 50, *n* (%)	321 (51.0)	223 (54.7)	68 (43.6)	4 (57.1)	13 (31.0)	13 (76.5)
Self race, *n* (%)						
African American	280 (44.5)	169 (41.4)	80 (51.3)	1 (14.3)	24 (58.5)	6 (35.3)
White	329 (52.3)	225 (55.1)	70 (44.9)	6 (85.7)	17 (41.5)	11 (64.7)
Other	20 (3.2)	14 (3.4)	6 (3.8)	–	–	–
Missing	1	–	–	–	1	–
Tumor grade, *n* (%)						
High	187 (29.7)	63 (15.4)	82 (52.6)	4 (57.1)	37 (88.1)	1 (5.9)
Intermediate	285 (45.2)	207 (50.7)	63 (40.4)	3 (42.9)	3 (7.1)	9 (52.9)
Low	158 (25.1)	138 (33.8)	11 (7.1)	–	2 (4.8)	7 (41.2)
Stage, *n* (%)						
I	265 (42.1)	189 (46.3)	49 (31.4)	5 (71.4)	13 (31.0)	9 (52.9)
II	285 (45.2)	170 (41.7)	86 (55.1)	1 (14.3)	21 (50.0)	7 (41.2)
III	65 (10.3)	42 (10.3)	16 (10.3)	–	7 (16.7)	–
IV	15 (2.4)	7 (1.7)	5 (3.2)	1 (14.3)	1 (2.4)	1 (5.9)
Node status, *n* (%)						
Negative	374 (59.5)	251 (61.7)	84 (53.8)	5 (71.4)	24 (57.1)	10 (58.8)
Positive	255 (40.5)	156 (38.3)	72 (46.2)	2 (28.6)	18 (42.9)	7 (41.2)
Missing	1	1				
Size, *n* (%)						
≤ 2cm	350 (55.7)	245 (60.2)	70 (45.2)	6 (85.7)	18 (42.9)	11 (64.7)
> 2cm	278 (44.3)	162 (39.8)	85 (54.8)	1 (14.3)	24 (57.1)	6 (35.3)
Missing	2	1	1			
ROR-P, *n* (%)						
High	65 (10.3)	–	35 (22.4)	–	30 (71.4)	–
Intermediate	347 (55.1)	205 (50.2)	121 (77.6)	7 (100.0)	12 (28.6)	2 (11.8)
Low	218 (34.6)	203 (49.8)	–	–	–	15 (88.2)
Chemo therapy, *n* (%)						
Yes	339 (53.8)	183 (44.9)	107 (68.6)	5 (71.4)	37 (88.1)	7 (41.2)
No	291 (46.2)	225 (55.1)	49 (31.4)	2 (28.6)	5 (11.9)	10 (58.8)

each core into k patches with a size of 200 × 200 pixels (Fig. 2C). Non-informative patches whose background pixels were above a proportional threshold were excluded (90% for original and binarized images, and 70% for epithelium and stroma images; see Fig. 2D). In addition, patches with artifacts were excluded by an artifact detector trained to identify folds, occlusions, and other image defects (Fig. 2E). Supplementary Information S2 provides additional details.

CNN features were extracted from each informative patch by transfer learning using the convolutional layers of the pre-trained VGG16 architecture [12, 13] (Fig. 2F). This process outputs a 512-dimensional vector for each patch.

Grade-adjusted patch-level classification

Figure 2**G** illustrates that patch-level *scores* were calculated for each patch from the 512-dimensional feature vector(additional details in Supplementary Information S3 and Fig. S3). Briefly, to efficiently summarize the information in the distribution of patch scores in each *core*, we used quantiles of the patch score distributions. We chose 15 equally spaced quantiles, which results in a 15-dimensional vector of quantiles summarizing each core. In the

last step, we generated a patient-level feature vector by taking the element-wise average of core-level features.

We adopted the Multiple Instance Learning paradigm (MIL, [14, 15]) in conjunction with weighted distance-weighted discrimination (wDWD, [16, 17]). By estimating patch-level wDWD direction discriminating Luminal A versus Luminal B, patch-level scores were calculated such that a more negative score means that the patch is more likely to be from a Luminal A (while Luminal Bs have positives scores). However, classification rules for Luminal A versus Luminal B tend to be confounded by tumor grade, which precludes capturing subtle features associated with Luminal subtypes, especially among lower grade participants. To address this, we performed grade adjustment using another trained classifier that discriminates tumor grade.

Patient-level classification with image features

To conduct patient-level classification between Luminal A and Luminal B subtypes, we employed ER-positive and HER2-negative samples for training, excluding ER-borderline and HER2-borderline tumors. We included the borderline samples for model validation. Following this approach, we employed the 15-dimensional features of participants with Luminal A (n = 404) or Luminal B (n = 150) subtypes for training. For the validation set, Luminal A (n = 408) and Luminal B (n = 156) samples were considered, encompassing ER-borderline and HER2-borderline cases. The dataset was divided using a stratified tenfold cross-validation technique, ensuring that each of the tenfolds maintained similar proportions in terms of tumor grade and subtype. The classification model employed in this study utilized weighted distance-weighted discrimination.

Patient-level classification with image features and clinical variables

The approaches above emphasized visualization of image features, which could be selected with respect to 15 dimensional quantiles. However, classification accuracy is a vital goal for clinical applications. Hence, we developed an additional model with a high level of accuracy and assessed its effectiveness in diagnostic protocols, which we will herein call the *stratified model*. The stratified model used the wDWD method with tenfold cross-validation to train and differs from the prior models in two ways: it directly utilizes the tumor grade (categorized as low-to-medium or high) to determine the classifier threshold by grade, and it incorporates clinical variables as additional predictors.

Test data: TCGA-BRCA whole slide images

Our models described so far are trained on the CBCS core images. To assess their generalizability, we applied to the Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) dataset [7, 8], using a virtual TMA constructed from whole slide diagnostic images [9].

While our grade-adjusted stratified model demonstrated superior performances (detailed in Results), we applied an unadjusted model in TCGA-BRCA with strong results. The visual features in the model were selected based on the CBCS core images and applied to the TCGA-BRCA without further training. A detailed comparison of our proposed diagnosis protocols on TCGA-BRCA is provided in Diagnosis Protocol Comparison in TCGA-BRCA.

Results

Clinical characteristics of the 630 participants are presented in Table 1. The characteristics of the ER+/HER2- breast cancers included in the study population are similar to the expected distribution in clinical and surveillance datasets, with higher grade and larger tumor size associated with Luminal B breast tumors.

Subtype classifiers with image features

Figure 3 presents the average sensitivity, specificity, and AUC scores of the cross-validated samples along with their standard errors. Consistent with confounding by grade, the specificity of the unadjusted original model (Fig. 3A) is 0.790 in low or intermediate grade participants but drops to 0.474 in high grade participants. Furthermore, this model provides unbalanced results in terms of sensitivity and specificity. Specifically, in low or intermediate grade participants, sensitivity is 0.555, while specificity is 0.790. The grade-adjustment process led to more balanced sensitivity and specificity, with sensitivity increasing to 0.664, and specificity being 0.760 among low or medium grade participants. The stratified model in the third row shows the best performance in terms of balance and accuracy as well as AUC.

To evaluate the role of various tissue components in prediction, segmented results were evaluated in Panel B. The model based on features of the epithelium image (first row of Fig. 3B) had the best AUC. Its performance parallels the model based on unsegmented original images (second row of Fig. 3A). Results were less accurate for stromal compartment and the binarized model, however still has reasonably high accuracy, suggesting the importance of the tumor-stroma interface in risk assessment.

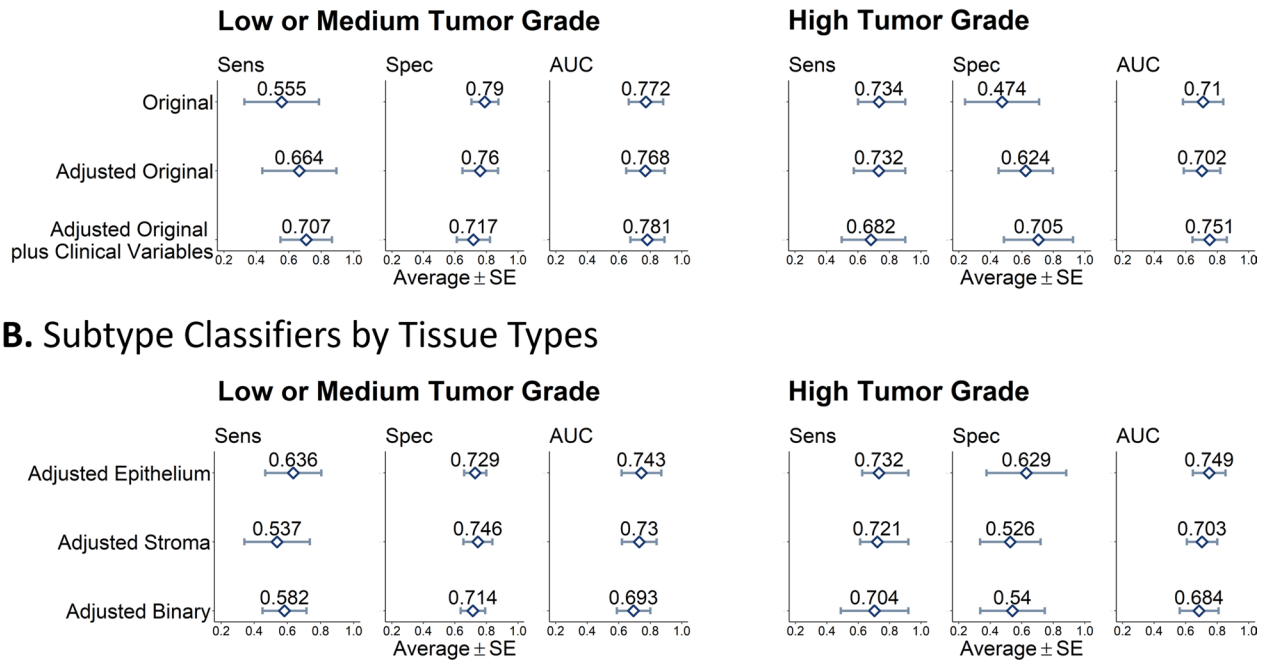


Fig. 3 Performance of subtype classifiers in the CBCS validation set. Average sensitivities, specificities, and AUC scores from tenfold cross-validation, along with their standard errors are provided for both low/intermediate and high grade. **A** Models trained on image features extracted from original core images for unadjusted, grade-adjusted, and stratified models. **B** Grade-adjusted models by image type (epithelium, stroma, or binary)

Visualization

A visual understanding of the Luminal A versus B image-based difference is shown in Fig. 4 using the cases of the 1st, 10th, 90th, and 99th percentiles of the distribution of DWD core-level scores. Representative patches are shown for each percentile, with the images on the left side representing the adjusted original images, with the images on the right side representing the adjusted binary model. In the original images, the Luminal A cores and patches had dense collagenous stroma with wavy collagen fibers. Most of the Luminal B cores, on the other hand, were occupied by high cellularity invasive carcinoma (i.e., malignant tumor cells) with a small amount of collagenous stroma and a few adipocytes. Luminal B tumor cell groups were bounded by thin bands of collagenous stroma. While all cores contained stroma and epithelium, patches with high stromal content or high epithelial cellularity represented the extremes of the DWD classifier for Luminal A and B, respectively. In binarized images, a more jagged shape of the interface between epithelium and collagenous stroma was characteristic of Luminal B tumors.

Proposed diagnosis protocols

Using risk stratification protocols as described in Methods and Fig. 1, we classified patients to perform time-to-event analyses for recurrence. Figure 5 displays Kaplan–Meier curves by risk groups for each protocol and reports the hazard ratio (HR) and p-values characterizing the differences between the risk groups. The proposed protocol **A** applied genomic testing to all low stage patients, **B** was solely reliant on histological information (Image-Based ML for All Low Stage), and **C** utilized both genomic testing and machine learning model

(See figure on next page.)

Fig. 4 Representative images for special CBCS cores located at the 1st, 10th, 90th, and 99th percentiles of the Luminal A and Luminal B DWD distribution, along with their representative patches. Each panel includes two sets of representative images: those generated from the grade-adjusted model using original images (left) and those from the grade-adjusted model of binary images (right). 1st and 10th percentile Luminal A cores and patches (upper row) exhibit dense collagenous stroma with a wavy collagen fiber pattern in original images. Conversely, 90th and 99th Luminal B cores (lower row) predominantly display high cellularity invasive carcinoma surrounded by thin collagenous stroma bands. Binarized images highlight a more irregular interface between epithelial (red) and collagenous stroma (green) regions in Luminal B tumors

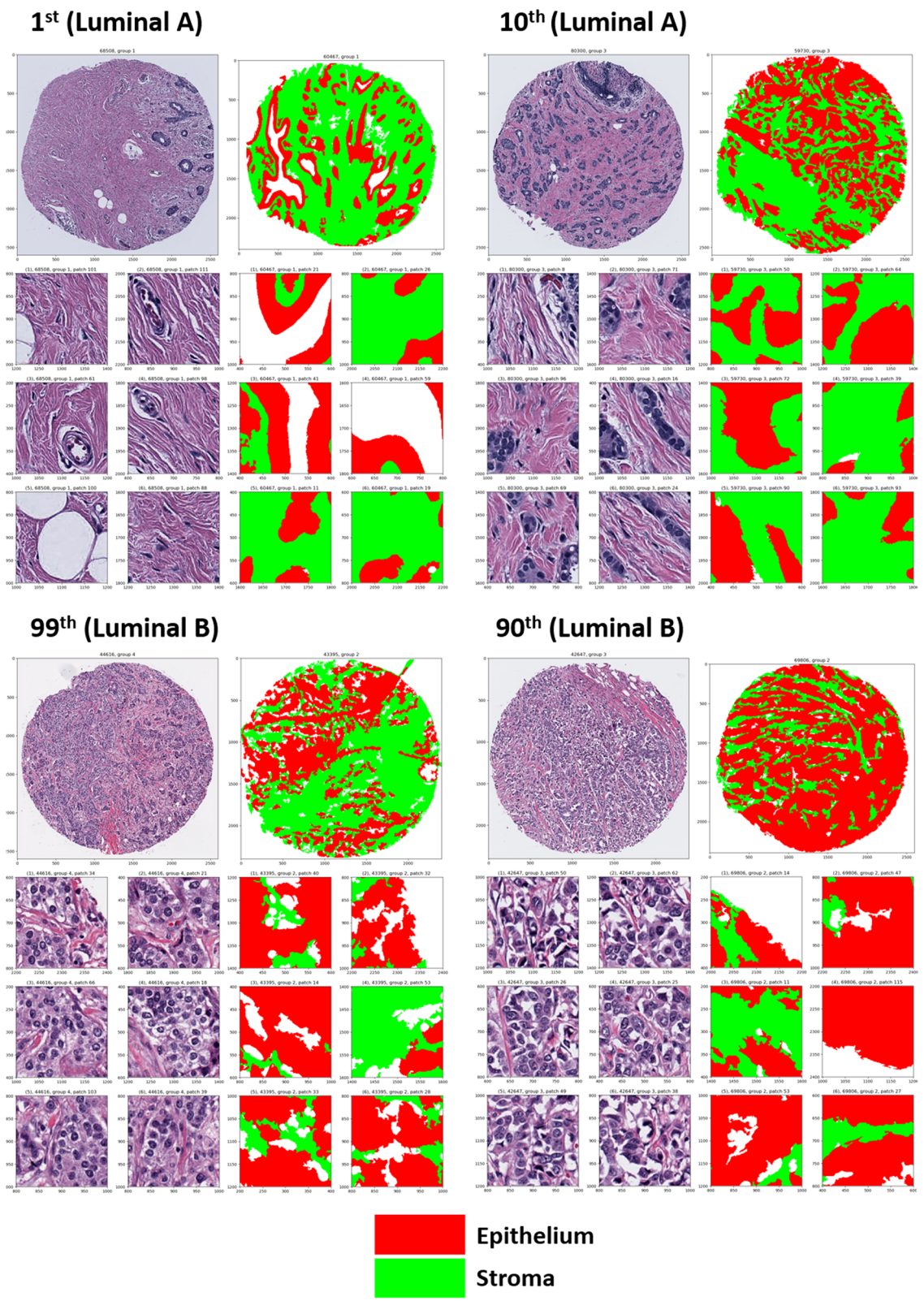
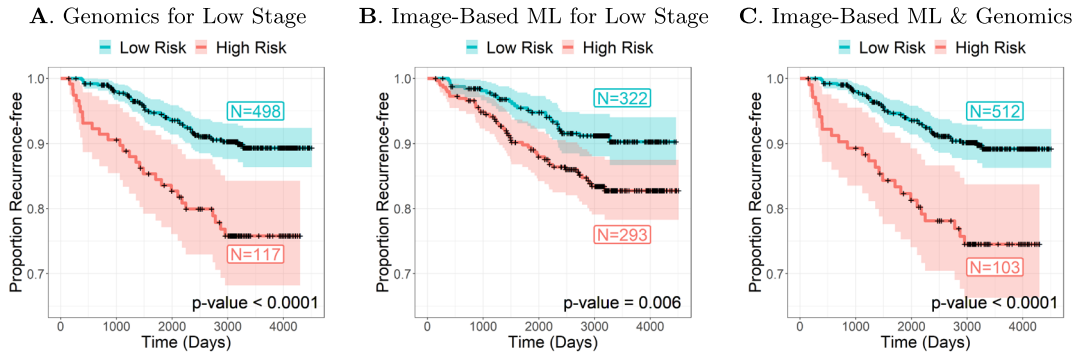


Fig. 4 (See legend on previous page.)



	A Genomics		B Image-Based ML		C Image-Based ML & Genomics	
Models	HR (95% CI)	*p*-value	HR (95% CI)	*p*-value	HR (95% CI)	*p*-value
Univariate Cox Model[1]	2.66 (1.65 - 4.28)	<0.0001	1.95 (1.21 - 3.13)	0.006	2.81 (1.73 - 4.56)	<0.0001
Multivariate Cox Model[1 2]	1.48 (0.85 - 2.58)	0.164	1.24 (0.75 - 2.07)	0.398	1.51 (0.85 - 2.67)	0.160

[1] Patients with stage 4 were excluded.
[2] Tumor size(<2 cm, ≥2 cm) + Node status(positive, negative) + Tumor grade(low-intermediate, high)

Fig. 5 Survival analysis results for the three protocols in **CBCS** data. Each panel displays Kaplan-Meier curves for the low-risk and high-risk groups classified by the corresponding protocol: **A** Genomics-based risk groups, **B** Image-based risk groups, or **C** Hybrid Image and Genomics-based risk groups. Results of univariate and multivariate Cox models for each protocol, including hazard ratios and *p*-values for differences between the risk groups are provided below the Kaplan-Meier plots

(Image-Based ML & Genomics for Low Stage). The image only protocol had lower univariate Cox HR (1.95, 95% CI: 1.21–3.13) than the other two models, but the hybrid model had similar HR to the fully genomically-tested protocol (HR 2.81, 95% CI: 1.73–4.56 for hybrid vs. HR 2.66, 95% CI: 1.65–4.28).

In addition to the univariate Cox model, we also employed the multivariate Cox model to assess recurrence time by risk groups, while appropriately controlling for tumor size, node status, and tumor grade. Addition of covariates decreased the precision of the effect estimates but did not substantially change the magnitude of the hazard ratios. We note that these decreases in precision may reflect collinearity among covariates. Nonetheless, in all cases the hybrid protocol **C** consistently performs well relative to the genomic protocol **A**, suggesting its potential clinical utility.

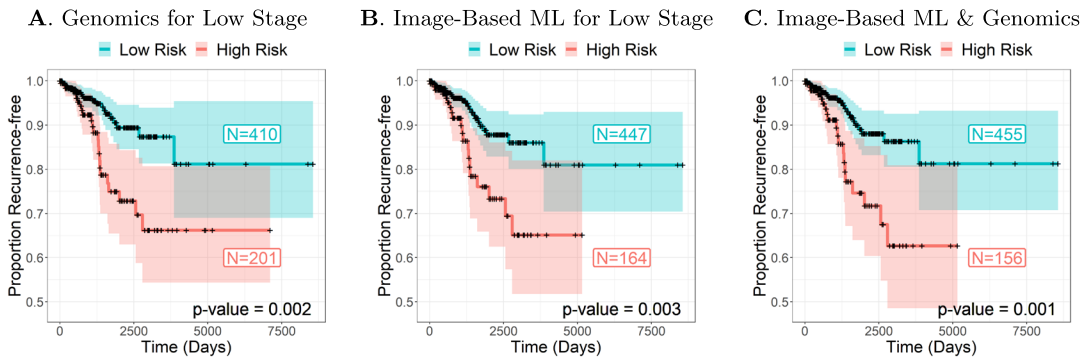
Additional multivariate Cox models incorporating PR status and Ki-67 percentage as covariates are presented in Table S3 in Supplementary Information. Sensitivity analysis was also performed stratifying on chemotherapy (Fig. S4). In all analyses, the hybrid protocols demonstrated comparable performance to the genomic-based protocol.

Diagnosis protocol comparison in TCGA-BRCA

To assess the generalizability of our findings, we conducted survival analyses on ER+ and HER2-tumors within the TCGA-BRCA dataset. We applied the unadjusted model (trained on CBCS core images) to TCGA-BRCA dataset. Given the limited availability of clinical covariates in TCGA-BRCA, only node status was incorporated into the multivariate Cox models. This would be expected to diminish model performance as grade was an important contributor to the CBCS models. However, application of our model to the TCGA-BRCA data demonstrated comparable prognostic performance (Fig. 6). Specifically, the hybrid protocol **C** (HR = 2.66, 95% CI: 1.49–4.78) exhibited similar hazard ratios to the genomic-based protocol **A** (HR = 2.53, 95% CI: 1.42–4.52), the magnitudes of which were similar to those observed in CBCS.

Discussion

Our results suggest that image features of low stage ER+/HER2-breast cancers can be used to identify patients with higher probability of high genomic risk, and that a hybrid protocol that uses histology on all patients and flags patients with higher risk tumors may have value in ensuring greater equity of access to genomic tests. Genomic tests are less frequently applied in uninsured women [5] and in some clinical settings that vary by geography, however if histologic images could be evaluated to identify tumors that are likely to be at higher risk, this may help encourage greater access to testing. At this stage, we do not advocate for replacing genomic testing with image-based model, but we posit that our model serves as proof of principle that image data could be a



	A Genomics		B Image-Based ML		C Image-Based ML & Genomics	
Models	HR (95% CI)	*p*-value	HR (95% CI)	*p*-value	HR (95% CI)	*p*-value
Univariate Cox Model[1]	2.53 (1.42 - 4.52)	0.002	2.44 (1.36 - 4.37)	0.003	2.66 (1.49 - 4.78)	0.001
Multivariate Cox Model[1 2]	2.11 (1.01 - 4.4)	0.046	1.92 (0.82 - 4.46)	0.132	2.32 (0.98 - 5.47)	0.055

[1] Patients with stage 4 were excluded.
[2] Node status(positive, negative)

Fig. 6 Survival analysis results in **TCGA** data. Each panel displays Kaplan-Meier curves for the low-risk and high-risk groups classified by the corresponding protocol: **A** Genomics-based risk groups, **B** Image-based risk groups, or **C** Hybrid Image and Genomics-based risk groups. Results of univariate and multivariate Cox models for each protocol, including hazard ratios and *p*-values for differences between the risk groups are provided below the Kaplan-Meier plots

valuable tool for identifying individuals who may benefit from genomic testing. Our results showed that the classification model predicted by image features worked best on grade-adjusted unsegmented images. However, the interesting finding that jagged borders suggest higher risk may be useful for pathologists to further consider in evaluating histologic images. These results imply that tumor images, including both epithelial features like grade and the relative regional shapes of epithelium and collagenous stroma contain information that has value in discriminating higher and lower risk tumors.

In breast cancer, numerous studies have previously utilized machine learning techniques to predict genomic subtype [18–20]. Carmichael et al. explored the relationship between PAM50 gene expression and H&E stained images, identifying interesting variations shared by both data types [11, 21]. Other studies have employed machine learning models to predict PAM50 subtypes using tumor images [22, 23]. Phan et al. (2021) employed diverse CNN architectures for subtype classification and conducted a performance comparison between these architectures [22]. Couture et al. (2018) utilized a multiple instance learning scheme to assign labels to small regions, achieving robust accuracies in various tasks, including PAM50 subtype (Basal-like vs. non Basal-like), tumor grade, and histology type classification [23].

Our results show similar accuracy to prior models, and consistent with Carmichael et al., suggest that image features can be identified for distinguishing Luminal A and B tumors. One advantage of our approach was that we considered sensitivity and specificity, indicating the value of grade adjustment. For a test that is being used to exclude low risk patients from a potentially harmful therapy on the basis of limited benefit, sensitivity to detect risk is important. Optimizing the balance of sensitivity and specificity may also be of greater importance than accuracy in some applications of visual tools, and thus considering all three metrics is important.

Further, our use of pixel-level tissue segmentation provided novel biological insights, Inspired by the work of Kilmov et al. (2019) [24]. While Kilmov et al. employed a slide annotation classifier for identifying histologic features at the patch level to construct a recurrence risk classifier, our innovative contribution lies in the development of a pixel-level segmentation model, enabling a more comprehensive discrimination between epithelium and stroma. This approach used herein to visualize image features directly related to the subtype differences may be useful to other researchers to identify features of predictive value.

Our model was developed and evaluated using data from a population-based cohort of North Carolina breast cancer patients. The robust performance of our proposed protocol in predicting recurrence within this diverse, real-world population suggests that our protocol has strong discriminating performance in predicting recurrence, similar to genomic tests themselves. We further validated the generalizability of the protocol by

testing our model on the external TCGA-BRCA dataset. While some clinical variables with predictive value (notably grade) were missing in the TCGA data, our protocol still demonstrated prognostic performance comparable to the genomics-based protocol. Furthermore, while our model was trained on TMA core images, these results suggest the potential utility of this approach for standard diagnostic images, although further research is needed to optimize whole slide image analysis. Despite dataset limitations, our protocol exhibited strong, externally validated prognostic value.

Our results suggest promising practical utility for machine learning-based protocols in clinical settings. To operationalize this approach broadly, data science methods for normalizing images, sampling patches, and adjusting for grade based on established parameters are needed. Our next step is to develop a user-friendly tool incorporating the machine learning pipeline to facilitate application in routinely collected diagnostic images. Additionally, our study implies the value in expanding histologic image analysis to more tumor types, with the potential to offer more equitable prognostic information at diagnosis to a larger number of cancer patients. In particular, physicians are in need of low cost tools that can help prioritize and improve the equity of precision medicine tools, and image-based analysis may provide one avenue toward this goal.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-024-01915-5.

Additional file 1

Funding

This research was supported by a grant from UNC Lineberger Comprehensive Cancer Center, which is funded by the University Cancer Research Fund of North Carolina, the Susan G Komen Foundation (OGUNC1202, OG22873776, SAC210102, TREND21686258), National Cancer Institute (R01CA253450), the National Cancer Institute Specialized Program of Research Excellence (SPORE) in Breast Cancer (NIH/NCI P50-CA058223), NIH Advanced Research Projects Agency for Health (140D042490008), and the US Department of Defense (HT94252310235). This research recruited participants &/or obtained data with the assistance of Rapid Case Ascertainment, a collaboration between the North Carolina Central Cancer Registry and UNC Lineberger. RCA is supported by a grant from the National Cancer Institute of the National Institutes of Health (P30CA016086). The authors would like to acknowledge the University of North Carolina BioSpecimen Processing Facility for sample processing, storage, and sample disbursements (http://bsp.web.unc.edu/) and the Breast Cancer Research Foundation HEI-23-003. We are grateful to CBCS participants and study staff.

Availability of data and materials

Carolina Breast Cancer Study is actively following patients and under and IRB-approved protocol that does not permit data sharing on public websites. However, we share data through an IRB-approved data use agreement system as described on our website (https://unclineberger.org/cbcs/for-researchers/).

Declarations

Ethics approval

The study was approved by the University of North Carolina Institutional Review Board in accordance with U.S. Common Rule. All study participants provided written informed consent prior to study entry. This study complied with relevant ethical regulations, including the Declaration of Helsinki.

Code availability

The pre-trained model and code used in this study are publicly available on GitHub (https://github.com/eastk90/cbcs-lumAB/).

Competing interests

The University of North Carolina, Chapel Hill has a license of intellectual property interest in GeneCentric Diagnostics and BioClassifier, LLC, which may be used in this study. The University of North Carolina, Chapel Hill may benefit from this interest that is/are related to this research. The terms of this arrangement have been reviewed and approved by the University of North Carolina, Chapel Hill Conflict of Interest Program in accordance with its conflict of interest policies.

Received: 16 February 2024 Accepted: 12 September 2024
Published online: 04 December 2024

References

1. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–52.
2. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7. https://doi.org/10.1200/JCO.2008.18.1370.
3. Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thürlimann B, Senn H-J. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol. 2011;22(8):1736–47. https://doi.org/10.1093/annonc/mdr304.
4. Tran B, Bedard PL. Luminal-B breast cancer and novel therapeutic targets. Breast Cancer Res. 2011;13(6):221. https://doi.org/10.1186/bcr2904.
5. Van Alsten SC, Dunn MR, Hamilton AM, Ivory JM, Gao X, Kirk EL, Nsonwu-Farley JS, Carey LA, Abdou Y, Reeder-Hayes KE, et al. Disparities in oncotypedx testing and subsequent chemotherapy receipt by geography and socioeconomic status. Cancer Epidemiol Biomark Prevent. 2024;33(5):654–61.
6. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng. 2009;2:147–71. https://doi.org/10.1109/RBME.2009.2034865.
7. 13, B..W.H..H.M.S.C.L...P.P.J..K.R., Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, G., Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, I., et al: Comprehensive molecular portraits of human breast tumours. Nature 2012;490(7418):61–70.
8. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015;163(2):506–19.
9. Li Y, Alsten SCV, Lee DN, Kim T, Calhoun BC, Perou CM, Wobker SE, Marron J, Hoadley KA, Troester MA. Visual intratumor heterogeneity and breast tumor progression. Cancers. 2024;16(13):2294.
10. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan Xiaojun, Schmitt C, Thomas NE. A method for normalizing histology slides for quantitative analysis. In: Woosley JT, editor. 2009 IEEE international symposium on biomedical imaging: from nano to macro, 2009. Boston: IEEE; 2009. p. 1107–10. https://doi.org/10.1109/ISBI.2009.5193250.
11. Carmichael I, Calhoun BC, Hoadley KA, Troester MA, Geradts J, Couture HD, Olsson L, Perou CM, Niethammer M, Hannig J, Marron JS. Joint and

individual analysis of breast cancer histologic images and genomic covariates. Ann Appl Stat. 2021. https://doi.org/10.1214/20-AOAS1433.
12. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015. arXiv. arXiv:1409.1556 [cs] .
13. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in pytorch 2017.
14. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell. 1997;89(1–2):31–71. https://doi.org/10.1016/S0004-3702(96)00034-3.
15. Maron O, Lozano-Pérez T. A framework for multiple-instance learning. Adv Neural Inf Process Syst. 1997;10:570–6.
16. Marron JS, Todd MJ, Ahn J. Distance-weighted discrimination. J Am Stat Assoc. 2007;102(480):1267–71. https://doi.org/10.1198/016214507000001120.
17. Qiao X, Zhang HH, Liu Y, Todd MJ, Marron JS. Weighted distance weighted discrimination and its asymptotic properties. J Am Stat Assoc. 2010;105(489):401–14. https://doi.org/10.1198/jasa.2010.tm08487.
18. Rakhlin A, Shvets A, Iglovikov V, Kalinin AA. Deep convolutional neural networks for breast cancer histology image analysis. In: Image analysis and recognition: 15th international conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, proceedings 15, 2018; pp. 737–744. Springer.
19. Ektefaie Y, Yuan W, Dillon DA, Lin NU, Golden JA, Kohane IS, Yu K-H. Integrative multiomics-histopathology analysis for breast cancer classification. NPJ Breast Cancer. 2021;7(1):147.
20. Chen RJ, Chen C, Li Y, Chen TY, Trister AD, Krishnan RG, Mahmood F. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2022; pp. 16144–16155.
21. Feng Q, Jiang M, Hannig J, Marron JS. Angle-based joint and individual variation explained. J Multivar Anal. 2018;166:241–65. https://doi.org/10.1016/j.jmva.2018.03.008.
22. Phan NN, Huang C-C, Tseng L-M, Chuang EY. Predicting breast cancer gene expression signature by applying deep convolutional neural networks from unannotated pathological images. Front Oncol. 2021;11:769447. https://doi.org/10.3389/fonc.2021.769447.
23. Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron J, Perou CM, Troester MA, Niethammer M. Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer. 2018;4(1):30.
24. Klimov S, Miligy IM, Gertych A, Jiang Y, Toss MS, Rida P, Ellis IO, Green A, Krishnamurti U, Rakha EA, et al. A whole slide image-based machine learning approach to predict ductal carcinoma in situ (DCIS) recurrence risk. Breast Cancer Res. 2019;21:1–19.

Publisher's Note