JMI Journal of Medical Imaging

RESEARCH PAPER

Local texture statistics augment the power spectrum in modeling radiographic judgments of breast density

Craig K. Abbey,[a,*] Margarita L. Zuley,[b] and Jonathan D. Victor[c]

[a]University of California, Santa Barbara, Department of Psychological and Brain Sciences, Santa Barbara, California, United States

[b]University of Pittsburgh Medical Center, Department of Radiology, Pittsburgh, Pennsylvania, United States

[c]Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, United States

ABSTRACT. **Purpose:** Anatomical "noise" is an important limitation of full-field digital mammography. Understanding its impact on clinical judgments is made difficult by the complexity of breast parenchyma, which results in image texture not fully captured by the power spectrum. While the number of possible parameters for characterizing anatomical noise is quite large, a specific set of local texture statistics has been shown to be visually salient, and human sensitivity to these statistics corresponds to their informativeness in natural scenes.

Approach: We evaluate these local texture statistics in addition to standard power-spectral measures to determine whether they have additional explanatory value for radiologists' breast density judgments. We analyzed an image database consisting of 111 disease-free mammographic screening exams (4 views each) acquired at the University of Pittsburgh Medical Center. Each exam had a breast density score assigned by the examining radiologist. Power-spectral descriptors and local image statistics were extracted from images of breast parenchyma. Model-selection criteria and accuracy were used to assess the explanatory and predictive value of local image statistics for breast density judgments.

Results: The model selection criteria show that adding local texture statistics to descriptors of the power spectra produce better explanatory and predictive models of radiologists' judgments of breast density. Thus, local texture statistics capture, in some form, non-Gaussian aspects of texture that radiologists are using.

Conclusions: Since these local texture statistics are expected to be impacted by imaging factors like modality, dose, and image processing, they suggest avenues for understanding and optimizing observer performance.

Keywords: breast density; breast texture; power spectrum; local texture statistics

Paper 23126GR received May 23, 2023; revised Oct. 5, 2023; accepted Oct. 16, 2023.

1 Introduction

Texture in medical images can be an important source of patient information. In breast-cancer screening by x-ray mammography—the primary focus of this work—exams are routinely classified into one of four breast-density categories (A–D) defined by the Breast Imaging, Reporting, and Data System (BI-RADS), based on the appearance of breast tissue in the images.[1–4]

*Address all correspondence to Craig K. Abbey, ckabbey@ucsb.edu

2329-4302/2023/$28.00 © 2023 SPIE

Breast-density judgments have important consequences for patients, as breast density is associated with risk of developing breast cancer,[3,5,6] and with the effectiveness of standard mammographic screening.[7,8] For patients whose exams have been categorized with BI-RADS densities of C or D (C, heterogeneously dense; D, extremely dense)—which are considered to be mammographically dense—alternative imaging approaches for screening may be considered.[9–13] While there has been a substantial effort to develop automated breast-density assessment,[6,14–16] most density judgments are made by radiologists as they read screening images. How radiologists accomplish this has been studied but is still not completely understood.[1,2,8,17]

The most common quantitative description of visual texture—both in general and in mammograms—is the power spectrum, which includes variability arising from acquisition noise and so-called "anatomical" noise representing the distribution of tissue in the images.[18–23] Anatomical noise is typically characterized by a power-law (PL) profile for the power spectrum from ~0.1 to 1.0 cyc/mm, above which acquisition noise dominates.[20,24]

A main motivation for the use of the power spectrum to characterize visual texture is that, under the assumption that the image is a stationary Gaussian random field, the power spectrum is a complete description of the statistical properties of the images. However, this is a strong assumption and it does not hold for mammograms.[25] The non-Gaussian nature of mammograms is not just a technical fine point: there are obvious visual differences between real mammograms and images that are constructed to have the same power spectrum.[25] Thus, non-Gaussian textural properties of mammograms are available to human observers. There is also a wealth of evidence that human observers can detect aspects of texture that are not captured by the power spectrum in other contexts, including natural textures[26] and purely synthetic ones.[27–29] The neural basis of visual sensitivity to these statistics is likely visual area V2.[26,30]

The question addressed here is whether these additional texture properties are actually used in radiologists' judgments of breast density, an important real-world judgment. Our approach is to build explanatory models of breast density judgments using standard features that capture Gaussian aspects of texture (i.e., the power spectrum). Then we determine whether the model improves when we add features that capture non-Gaussian aspects of texture. For the latter, we use local texture statistics, a choice that is attractive because human sensitivity to them is well characterized[31] and has been shown to parallel their importance in natural scenes (NSs).[32] We use step-wise regression with model selection criteria [Akaike information criteria (AIC) and Bayesian information criteria (BIC)[33–35]] that penalize larger numbers of features to quantify the degree to which these additional features explain radiologist judgments of breast density. We evaluate the predictive accuracy of these models using cross-validation.

2 Methods

2.1 Image Data

A dataset of 111 screening exams from the University of Pittsburgh Medical Center were used as the basis for this study. The full-field digital mammography (DM) images were acquired on Hologic systems (Lorad Selenia and Selenia Dimensions systems; Marlborough, Massachusetts, United States) with an isotropic detector pixel size of 0.07 mm using standard clinical acquisition protocols and "for-display" processing. Each exam consisted of two standard bilateral mammographic views [mediolateral oblique view (MLO) and cranio-caudal (CC) views of each breast (e.g. LCC is left CC view), for a total of 4 images per exam and 444 images in total. All exams were assigned a BI-RADS interpretive rating of 1 (negative for malignancy) and had follow-up for at least 1 year to verify the true-negative status of the exam. This precaution was taken to ensure that the presence of disease did not somehow bias the evaluation of breast density. A BI-RADS breast density score was given for each case by an evaluating radiologist, and thus represents screening practice at University of Pittsburgh Medical Center. Figure 1 shows samples from each of the four breast density categories as well as examples of the four images produced in a screening exam. All data were collected under an IRB-approved human-subjects protocol and de-identified for use in this study.

The cases in this dataset represent the population frequencies of each of the breast density categories up to sampling variability. This means that there are relatively fewer of the extreme—

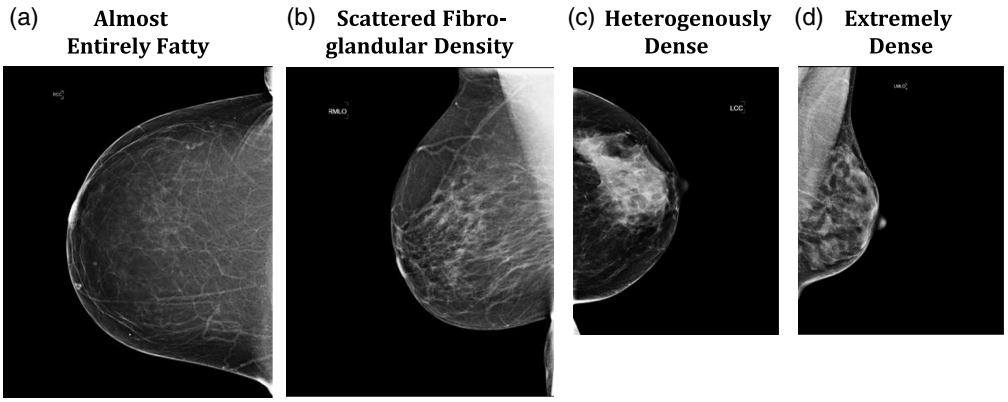

Fig. 1 Breast Density Categories. Images from each of the four breast density categories are shown with examples of the right (a, b) and left (c, d) breasts and the CC (a, c) and MLO (b, d) views. Note that the different sizes of the images shown reflect the field of view for images of smaller breasts.

A and D—density categories (13.5% and 15.3%, respectively) relative to the middle—B and C—categories (22.5% and 48.6%, respectively). Breast density is typically used to classify women as having mammographically dense breasts (C and D) or not (A and B), and this is the density judgment that we model in this work. In this dataset, 64% of the cases are considered mammographically dense.

2.2 Statistical Properties of Mammograms

The goal of this work is to use image-derived statistics as explanatory variables in models of breast density judgments made by the examining radiologists. As a first step in computing them, we define the breast interior using manual segmentation ~1 cm inside the skin line and nipple and excluding pectoralis muscle. Figure 2(a) shows an example of a CC view image with the segmentation boundary added as an overlay. Texture features are extracted from the interior of the segmented region.

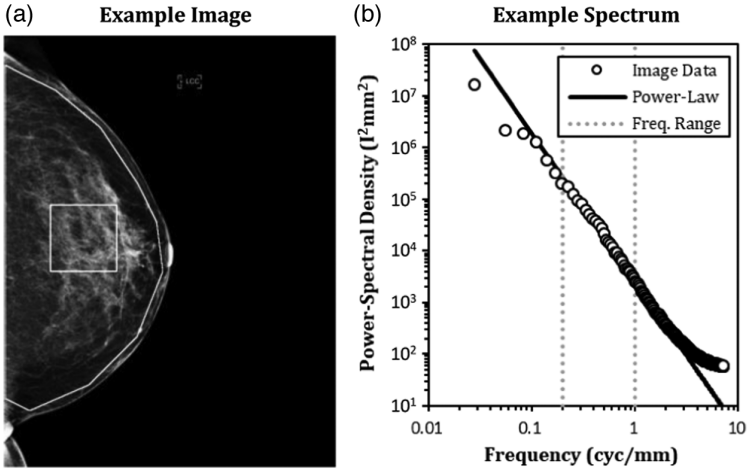

Fig. 2 Breast Segmentation and Power-Spectrum. The image (a) shows a standard LCC view with manually segmented interior region. For reference, a 512 × 512 ROI is shown in the middle of the segmented breast interi or. The power spectrum resulting from 50 randomly positioned ROIs is plotted (b) along with a power-law fit over the frequency range from 0.2 to 1.0 cyc/mm. Above 1.0 cyc/mm, the observed spectrum is seen to diverge from the power-law model due to acquisition noise.

2.2.1 *Power spectral features*

For each image, the power spectrum of the breast interior was computed in way that is similar to previous studies,[20,25,36,37] by randomly positioning square regions of interest (ROIs) entirely within the segmented breast interior. The size of the ROI was determined iteratively with an initial value of 512 pixels (35.8 mm on a side, 12.8 cm^2 total area), and a total of 50 ROIs were used for each image. The area of the ROI centers that allowed placement of the ROI so that the entire square was within the segmented breast interior was then calculated. If this area was at least as large as the area of the ROI, then the ROI size was kept, and if not, a smaller ROI size was evaluated. If the 512-pixel ROI size did not result in enough area, then the size of the ROI was reduced to 384 pixels. If this size was determined to be too large, then an ROI size of 256 pixels was used. For 71.8% of the images, the 512-pixel ROI was adequate, while the 384-pixel and the 256-pixel ROIs were used in 25.7% and 2.5% of the images, respectively. Because of the random positioning of the ROIs, there is a variable amount of overlap between ROIs. The average overlap fraction between two ROIs in this procedure was 18.3%.

The spectrum was computed from the 50 ROIs using standard power-spectral-density estimation,[38–40] including a circular Hanning window that extended the length of the ROI in each dimension. The two-dimensional power spectrum was collapsed to a single dimension by averaging over radial bins the width of a single frequency sample. Within a frequency range of 0.2 to 1.0 cyc/mm, a power-law was fit to the power spectrum by fitting a line to the spectrum in the log-frequency log-power-spectral-density domain, as described by Burgess et al.[20] The power spectrum and power-law fit for the image in Fig. 2(a) are plotted in Fig. 2(b).

We extracted three features from the spectral analysis of the mammograms. These are the slope and intercept of the power-law fit in log–log coordinates (i.e., the power-law exponent and amplitude at 1 cyc/mm), and the integrated power over the fitting range from 0.2 to 1.0 cyc/mm. These parameters characterize the spectral anatomical properties of the image.

2.2.2 *Local texture statistics*

Non-Gaussian aspects of texture were quantified by local texture statistics, which describe the prevalence of local patterns of bright and dark in an image, as shown in Fig. 3. To determine the local texture statistics, we used a procedure similar to that used by Hermundstad et al.[32] for natural images, with the main difference that we did not whiten the images. The whitening step

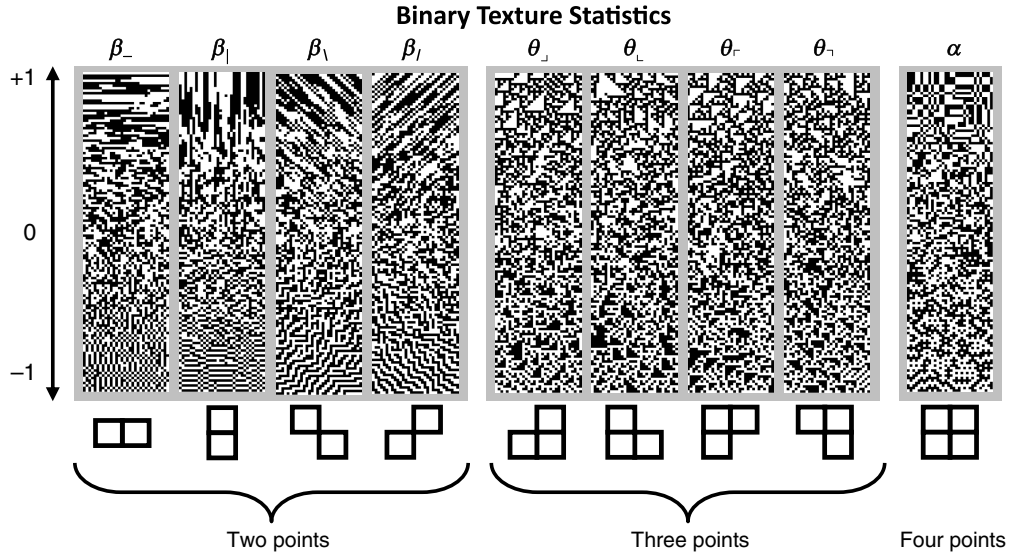

Fig. 3 Binary texture statistics. Each of the nine panels shows the appearance of a binary texture as the coefficient for each texture statistic changes from −1 to 1; a coefficient of 0 corresponds to local independence. The boxes below each panel show the pixel configurations used to define co-occurrence for that statistic. Adapted with permission from Victor et al.,[31] Figure 1.

in Hermundstad et al. consisted of a spatial filtering step, in which the amplitudes of each image's Fourier components were divided by the Fourier amplitude of the power spectrum of the entire image ensemble. Here, we omitted this step because the goals of the two studies were different: we wanted the statistics of each image to be computable from just that image. Correspondingly, for comparison of the statistics of mammograms to those of the natural image ensemble used in previously,[32] we recomputed the statistics of those natural images without the whitening step.

Local texture statistics were computed at scales ranging from 1 to 32 pixels (0.07 to 2.24 mm), following an initial step to limit the analysis to the interior of the breast. Figure 4 shows this process for an example mammogram. To ensure that the analysis at each scale used exactly the same image pixels, we constructed a coarsened segmentation region consisting of 32×32 pixel blocks, each of which was entirely contained within the full-resolution interior segmentation region described above. An example of an image and the coarsened segmentation region is shown in Fig. 4(a). Next, to remove spurious large-scale spatial variations that may arise during the imaging process (e.g., from non-parallel breast compression paddles), the image was then detrended within the interior segmented region of each breast image. This was done by least-squares fitting a quadratic polynomial to the image intensity within the segmented breast interior, and subtracting the fitted polynomial from the actual image intensities. In this way, we removed linear and quadratic trends across the breast interior. The result of detrending the image in Fig. 4(a) is shown in Fig. 4(b). Following detrending, the segmentation region was then down-sampled to each scale of interest using pixel averaging, and the median value of the down-sampled image was used to classify each pixel as positive ($\geq$ median, displayed as white) or negative ($<$ median, displayed as black). Figure 4(c) shows the binarized patterns obtained for the sample image at scales from 1 (no down-sampling) to 32.

Local texture statistics were then computed from these images. Since the previous step binarized the images at the median, they necessarily have an equal number of positive and negative checks. So the first group of meaningful binary image statistics are those that describe patterns of signs in pairs of neighboring checks, denoted $\beta_{\_}$, $\beta_{|}$, $\beta_{\backslash}$, and $\beta_{/}$. They are defined as follows. $\beta_{\_}$ measures the frequency of occurrence of horizontally adjacent checks with the same sign, compared to the frequency of occurrence of horizontally adjacent checks with the opposite sign. Its value can range from -1 to 1, where 1 means that every pair of horizontally adjacent checks have the same sign; -1 means that horizontally adjacent checks always differ in sign; and 0 means that there is no correlation between the signs in horizontally adjacent checks. The local image statistics $\beta_{|}$, $\beta_{\backslash}$, and $\beta_{/}$ are defined in the analogous way, for checks that are adjacent vertically ($\beta_{|}$) or share a corner along one diagonal ($\beta_{\backslash}$) or the other $\beta_{/}$). The first four strips in Fig. 3 illustrate the visual impact of these image statistics. The strips are synthetic textures in which the value of β changes gradually from -1 at the bottom of each strip, to $+1$, at the top; other than the pairwise correlations required by the value of β, these textures are as random as possible. Note that in our

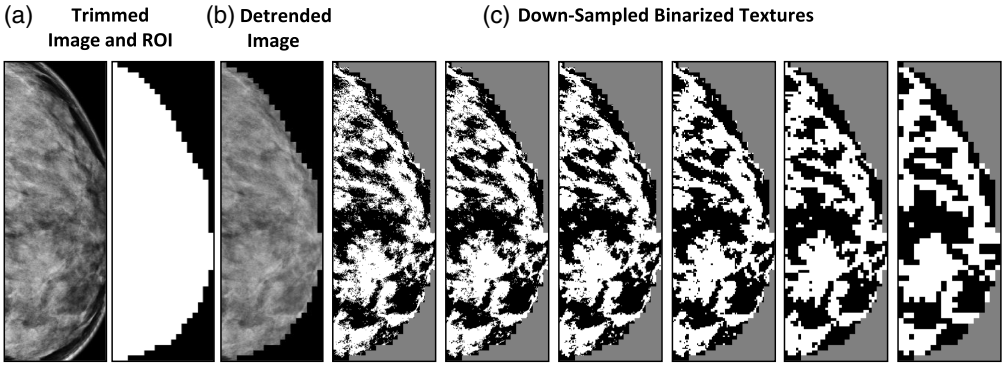

Fig. 4 Binarization procedure for co-occurrence statistics. The image and segmentation region (a) are cropped to fit the segmented region, and then the image intensities are detrended using a low-order polynomial fit (b). Binarization images are constructed by median-thresholding (c) at down-sampled scales from 1×1 to 32×32 pixels. The co-occurrence statistics are computed by counting specific 2×2 patterns.

analysis, the right facing views (RMLO and RCC) are mirrored so that the statistics from left-facing and right-facing views correspond to the same anatomical direction.

As detailed in Victor and Conte,[41] this strategy extends to patterns of triplets and quadruplets of checks. The local image statistics $\theta_{\lrcorner}$, $\theta_{\llcorner}$, $\theta_{\ulcorner}$, and $\theta_{\urcorner}$ describe patterns of triplets, by specifying the fraction of triangular neighborhoods of checks filled with an odd number of positive checks (3 or 1), versus an even number of positive checks (2 or 0). There are four such statistics because there are four possible orientations of a triangular cluster of checks. The corresponding strips in Fig. 3 show that large deviations of these statistics toward either $+1$ or -1 correspond to textures that have clusters of either positive or negative regions. Note that the θ-statistics—and the visual features that they capture—are independent of the β's, as there are no pairwise correlations in any of these strips. Similarly, α describes the patterns of quadruplets of checks, by specifying the fraction of 2×2 neighborhoods that contain an even number of positive checks, versus an odd number. The rightmost strip in Fig. 3 shows its visual impact: values close to 1 generate textures with equally prominent positive and negative blobs; values close to -1 generate textures with a weave-like appearance.

Finally, we note that at each scale, the nine image statistics described above ($\beta_{\_}$, $\beta_{|}$, $\beta_{\backslash}$, $\beta_{/}$, $\theta_{\lrcorner}$, $\theta_{\llcorner}$, $\theta_{\ulcorner}$, $\theta_{\urcorner}$, and α) are effectively independent "coordinates" in texture space (they can be manipulated independently in synthetic textures); they are all visually salient; and along with the fraction of positive and negative checks (fixed at 0.5 here because of median thresholding), they fully specify the distribution of patterns in 2×2 neighborhoods.[41] In natural images, these statistics are approximately scale-independent[32] after whitening; here, we do not assume that this is the case, so we analyze them at a range of scales.

2.3 Generalized Linear Models of Breast Density Judgments

The models used in this work consider breast density judgments for mammograms to be a binomial variable, Y, that has a value of 1 if a mammogram is considered dense (BI-RADS density score of C or D) and a value of 0 if it is considered not-dense (BI-RADS density score of A or B). We dichotomize the scale to recognize that the distinction between dense and not-dense is the main clinical distinction, and because it simplifies the analysis. We then model this judgment as a binomial random variable with a probability that depends on the value of image statistics (spectral parameters and possibly local texture statistics) through a logistic link function.

The models are formalized as follows. Let Y_n be the density judgment for exam n ($n = 1, \ldots, 111$), and let $p_{n,v}$ represent the model's binomial probability, calculated from the information available in image view v ($v = 1, \ldots, 4$ for LCC, LMLO, RCC, and RMLO), that this mammogram is judged as dense. An accurate model will tend to produce binomial probabilities that are closer to 1 for images judged as dense, and closer to 0 for images that are not judged to be dense. To implement the logistic link function in a generalized linear model, the binomial probability is related to an internal variable $u_{n,v}$ as

$$p_{n,v} = \frac{e^{u_{n,v}}}{1 + e^{u_{n,v}}}, \tag{1}$$

where $u_{n,v}$ is modeled as a linear combination of image statistics. That is, if $X_{n,v,j}$ represents the value of the j'th statistic of the image in view v of exam n, our model is that

$$u_{n,v} = c_0 + \sum_{j=1}^{J} c_j X_{n,v,j}, \tag{2}$$

where c_0 controls the overall propensity to label an image as dense, and $c_j (j \geq 1)$ is the weight associated with each statistic that controls the influence the statistic on the judgment for each image.

As is standard for logistic regression, the model parameters c_j were determined by maximizing the likelihood that the observed judgments Y_n are made, given that their probabilities are determined by $p_{n,v}$ via Eqs. (1) and (2). Specifically, a given set of parameter values (i.e., values of c_j) leads to values for $u_{n,v}$ and $p_{n,v}$ via Eqs. (2) and (1), and these lead to a log-likelihood for the observed judgments Y_n of

$$\lambda = \sum_{n=1}^{111} \sum_{v=1}^{4} Y_n \ln(p_{n,v}) + (1 - Y_n) \ln(1 - p_{n,v}). \quad (3)$$

Note that this quantity will be maximized when, for $Y_n = 1$ (BI-RADS density score C or D), $p_{n,v}$ is large, while for $Y_n = 0$ (BI-RADS density score A or B), $p_{n,v}$ is small. Maximum-likelihood estimates of the c_j terms are found by maximizing the log likelihood [Eq. (3)].

As an important component of this investigation is to determine whether incorporation of non-Gaussian statistics results in a better explanatory model for breast density judgments, there is a crucial potential confound to exclude. Generically, adding more parameters may appear to yield better models, because of over-fitting. We address this in three different ways: (1) likelihood-ratio tests for the significance of additional parameters related to local texture statistics, (2) stepwise regression with model-selection criteria that penalize the likelihood function according to the number of parameters, and (3) cross-validated measures of predictive accuracy (see Sec. 2.4).

Likelihood-ratio tests[42] are used for nested models to evaluate the significance of additional parameters. This approach is well suited to this work since we are interested the additional benefit of local texture statistics over spectral statistics, which are inherently nested models. The difference in log-likelihood values between the full model (i.e., spectral statistics and local texture statistics) and a nested model (only spectral statistics) is asymptotically distributed as chi-squared with degrees of freedom (DFs) equal to the difference in the number of parameters.

Stepwise regression[43,44] is a standard procedure for determining the subset of available parameters that are most relevant to a regression model. This procedure starts with the one-parameter model consisting of the parameter, c_0, and then selectively adds regressors (i.e., image statistics with an associated weighting parameter), in a process called forward-selection, before removing regressors that become redundant (backward-selection). Regressors are added or removed on the basis of reducing a model-fit criterion until convergence is achieved. The outcome of interest in all of these studies is the value of the model-fit criterion when different sets of image statistics are used as potential regressors.

For model-fit criteria, we use the AIC and BIC.[33–35,45] The AIC is defined as

$$\mathrm{AIC} = -2\lambda + 2(J + 1), \quad (4)$$

where $J + 1$ represents the number of parameters used in the model (J statistics and an intercept). Lower values of AIC imply a better model, and a model with a larger number of parameters needs to increase the log-likelihood considerably in order to offset the penalty of having more parameters. The BIC applies a stronger penalty for the number of parameters

$$\mathrm{BIC} = -2\lambda + \ln(N_D) \cdot (J + 1), \quad (5)$$

where N_D is the number of data points fit. Here $N_D = 444$ (111 exams with 4 views each). A discussion of the tradeoffs between these measures (and others) can be found in the statistical modeling literature (see Ref. 46).

We apply these criteria to a number of different models based on image statistics as regressors for breast density. For reference, we include an overly simplistic one-parameter model that only considers the overall tendency to judge an exam as dense—c_0 in Eq. (2). The spectral feature model (SFM) adds the three spectral features described in Sec. 2.2.1, for a total of four free parameters. Then we add local texture statistics to the SFM at each level of downsampling, yielding models referred to as SFM + D1 to SFM + D32. These are 6 different models, each with 13 parameters (c_0, the 3 spectral features, and the 9 local texture statistics). Finally, we consider the SFM with all the local texture statistics (9 statistics at each of 6 scales) for a total of 58 parameters, referred to as SFM + All. The SFM may be considered as a nested model within each of the models that includes local texture statistics.

2.4 Predictive Accuracy

In addition to evaluating the statistical models in terms of likelihood-ratio tests and model-selection criteria, we also evaluate the predictive accuracy of models as a way to understand the relevance of the image statistics they are based on. We do this in a cross-validated fashion: a leave-2-out approach in which we hold out one exam on a patient with mammographically dense breasts and one exam on a patient with non-dense breasts from the model fitting procedure.

After the model has been fit to the remaining data using stepwise regression techniques described above, it is applied to the data from the two withheld exams. This results in eight predictions of judgments: for each of the four views in each of the two exams. We then compute the area under the empirical ROC curve (AUC) for these eight responses for each held-out pair. The AUC is then averaged across all possible hold-2-out pairs of exams (2840) to form a final estimate of discrimination performance. Uncertainty in these averages is estimated by bootstrapping. In the bootstrap procedure, dense and non-dense exams are resampled independently.

We also use the bootstrap procedure to compare the accuracy of models that include local texture statistics to the power-spectrum statistics by themselves. The null hypothesis is that there is no improvement in accuracy. The significance of differences in accuracy values is determined by computing the fraction of bootstrapped differences in which the accuracy of the augmented model exceeds the accuracy of the power-spectrum model (a one-sided p-value).

2.5 Gaussian Surrogate Images

Since our models made use of only a parametric description of the power spectrum (i.e., average intensity, slope, amplitude, and integrated power) rather than the power spectrum itself, there is the possibility that the improvements in model performance that result from incorporation of local texture statistics might in fact reflect, in an indirect way, Gaussian statistics that were not captured by the chosen spectral parameters. To control for this possibility, we also apply the methods described above to simulated images that are constructed to have the same power spectrum as the mammograms, but in which deviations from Gaussian behavior are eliminated. These images were generated on an image-by-image basis from each mammogram using a stationary Gaussian random field, resulting in a total of 444 of these images. Each Gaussian surrogate (GS) image is also matched in terms of mean intensity and the power spectrum of the corresponding breast image and is generated on the same size pixel grid.

The image generation procedure for GS images is fairly standard[19,47,48] and begins with a sample of standardized Gaussian white noise. This initial image is transformed to the frequency domain and multiplied by the square root of the power spectrum for the corresponding breast image. This product is then inverse-transformed back to the spatial domain where the average intensity of the corresponding breast image is added to all pixels. Image statistics are then computed from the GS images using exactly the same segmented region as the corresponding breast image. Figure 5 shows an example of a segmented breast image along with a mean and power-spectrum matched GS.

2.6 Calculations

The above calculations were performed in IDL (NV5 Geospatial, Boulder CO, Secs. 2.2.1 and 2.5), or Matlab (The Mathworks, Natick, Massachusetts, United States), using custom software (Secs. 2.2.2, 2.3, and 2.4) and the stepwise generalized linear model function stepwiseglm (Sec. 2.3).

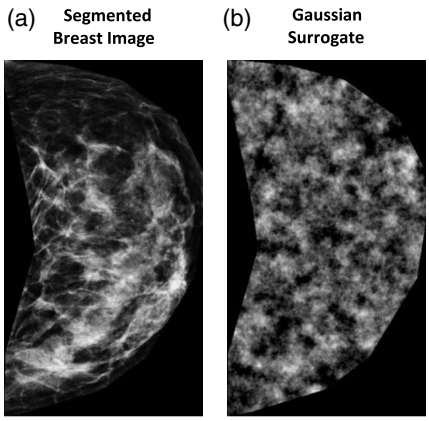

Fig. 5 Mammography and Gaussian Surrogate images. A segmented breast image is shown (a) along with a Gaussian surrogate image (b) that is generated with the same mean and power spectrum.

3 Results

Results of this study fall into three categories: (1) a description of the local texture statistics of mammograms, (2) an evaluation of goodness-of-fit for models that include local texture statistics as explanatory variables, and (3) a demonstration of predictive accuracy for these models.

3.1 Characteristics of Local Texture Statistics in Mammograms

Figure 6(a) shows the average values of the local texture statistics of mammogram images and GSs. For comparison, we also plot the local texture statistics determined from a standard database (the Penn Natural Image Database[32,49]) of natural scenes (NSs), computed in the same way as for mammograms and their surrogates. There are major differences between the local image statistics of mammograms and those of NSs: across scales, local texture statistics of NSs deviate further from zero, both in the positive direction (for the βs and α), or the negative direction (for the θs)—formalizing a sense in which NSs are more highly structured than mammograms. Importantly, the statistics of mammograms and GSs differ from each other as well, although these differences are not as large as the differences between mammograms and NSs. These differences, which are

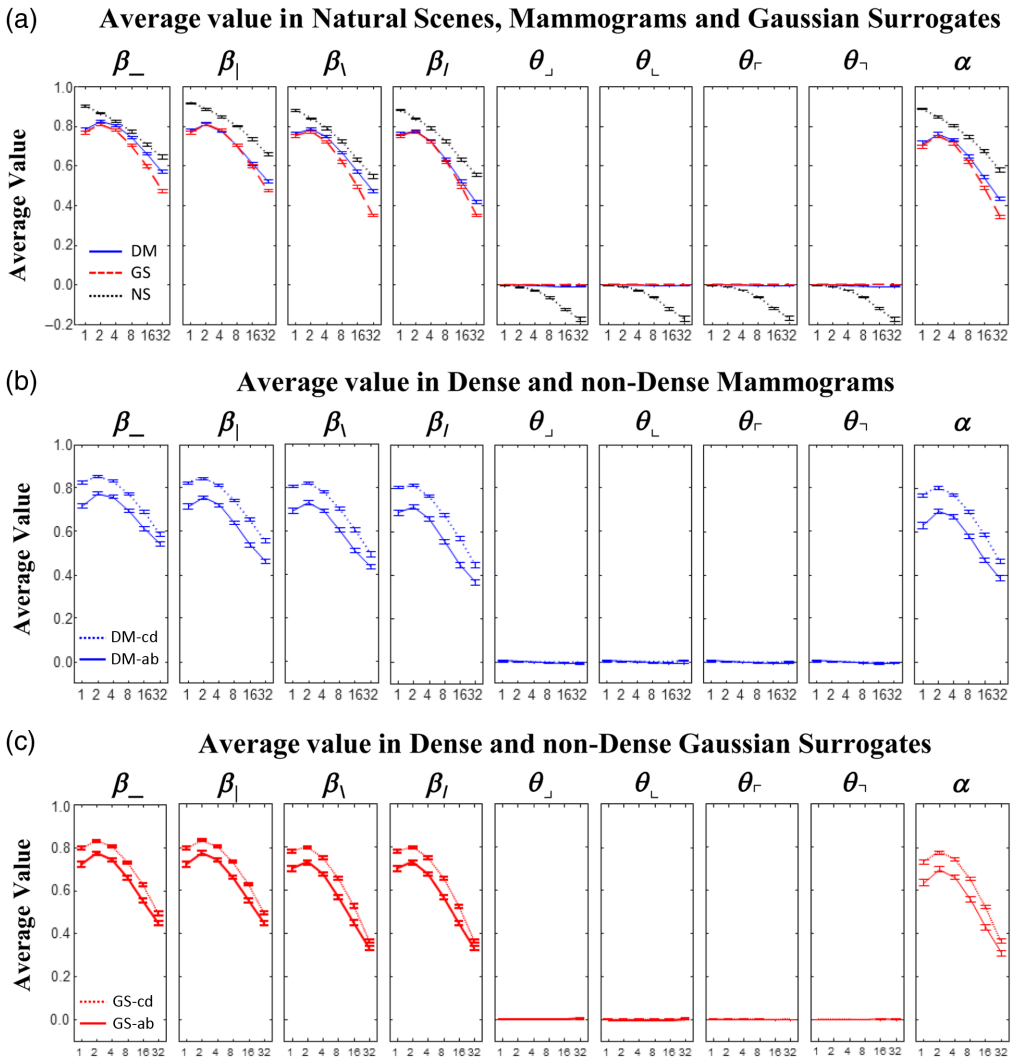

Fig. 6 Local texture coefficients across scales. Each plot shows average values of the nine binary-texture statistics (β = 2-point statistics; θ = 3-point statistics; α = 4-point statistics) across various levels of downsampling (1–32). A comparison of different image types (a) shows that average values for DMs, GSs, and NSs can differ across scales. The statistics are also sensitive to breast density, with differences between dense (BI-RADS C and D categories) and non-dense (BI-RADS A and B categories) mammograms (b) and GSs (c). Error bars represent ± 1 standard error.

most prominent for β's and α, express ways in which mammograms and their GSs are inequivalent. Note that these differences are primarily present at the coarser spatial scales. This is consistent with the expectation that at the finest scales, images are dominated by acquisition noise, which is approximately Gaussian, while at larger scales, anatomical noise plays a progressively more important role.

Figure 6(b) compares the local texture statistics for mammograms judged to have BI-RADS density scores of A or B (non-dense), versus C or D (dense). These demonstrate differences across all spatial scales, and the differences are substantial—comparable to the differences between mammograms and NSs. The statistics of GSs of mammograms for non-dense versus dense mammograms in Fig. 6(c) also show differences, but in comparison to the behavior of mammograms [Fig. 6(b)], these differences are attenuated, especially as spatial scale increases.

3.2 Modeling Breast Density Judgments

Generalized linear models were used as explanatory models for breast density judgments on the basis of image statistics, with likelihood ratio tests and model selection criteria (AIC and BIC) used to quantify the importance of including local texture statistics. The results of the likelihood-ratio tests are shown in Table 1. In all comparisons between the SFM model and models that add local texture statistics, the additional parameters are highly significant ($p < 10^{-6}$) and remain significant after multiple-comparisons corrections for the seven tests (six individual scales, SFM + D1 to SFM + D32, and the all-scale model, SFM + All).

Figure 7 shows the results of stepwise regression and model-selection analysis, focusing on the incremental benefit of adding local texture statistics to spectral measures. As a baseline, the "constant" model includes only the constant term (c_0 in Eq. (2) above); it does not make use of any image statistics and serves as a benchmark for comparison. The SFM model adds the spectral features (slope and intercept of the power-law fit, and the integrated power over the range from 0.2 to 1.0 cyc/mm). Not surprisingly, including spectral statistics always improves model-selection criteria relative to the constant model, by either AIC or BIC, a difference of more than 200 units in these measures. As expected, the improvement is nearly identical for mammograms and their GSs (since the latter has similar spectra).

We then consider models that add the local texture statistics. For mammograms, including local texture statistics at each scale improves the SFM further, yielding a reduction in the model selection criteria of 35 to 133 units. Across all models, the minimum AIC or BIC is achieved by the model that uses local texture statistics across all scales (SFM + All). The scale-by-scale analysis shows that finer scales contribute more than coarser scales, even though the fine scales likely contain significant imaging noise. Interestingly, the all-scale model is only minimally better than the best single-scale models (1 × 1 downsampling), suggesting that textural information is somewhat redundant across scales.

A supporting point is that for the GS images, the local texture statistics yield a much smaller improvement (19 to 49 units) over the SFM model. The gap between the improvement seen for

Table 1 Likelihood ratio tests for nested models. Various tests of the SFM with additional local texture statistics to the SFM itself. For each test, the DFs are given along with the test statistic value (χ^2) and significance (p-value), which have not been corrected for multiple comparisons.

Full model	Null model	DFs	χ^2	p-value
SFM + D1	SFM	9	225.6	$< 10^{-6}$
SFM + D2	SFM	9	147.5	$< 10^{-6}$
SFM + D4	SFM	9	124.0	$< 10^{-6}$
SFM + D8	SFM	9	113.7	$< 10^{-6}$
SFM + D16	SFM	9	113.0	$< 10^{-6}$
SFM + D32	SFM	9	116.9	$< 10^{-6}$
SFM + All	SFM	54	70.5	$< 10^{-6}$

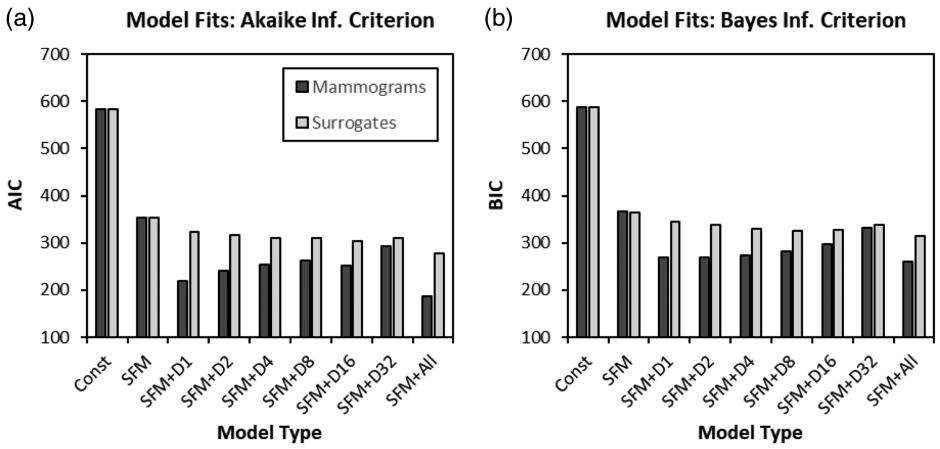

Fig. 7 Model selection criteria. The plots show AIC (a) and BIC (b) as a function of the breast density models used for mammograms and GS images. For both measures of fit, lower values indicate the better model. The constant model (on the left side of each plot) uses an intercept as its only parameter and serves as a reference for the benefit of additional parameters. The SFM model uses the three spectral features (along with an intercept) as parameters. The SFM + D# models add the nine binary texture statistics at each level of downsampling (1–32). The SFM + All model uses all 54 binary texture statistics (across all downsampling levels) in addition to the SFM. All models are fit by stepwise regression.

mammograms and for GS accounts for contributions of power-spectrum not captured by the specific spectral features used here. Independent of the AIC and BIC findings, this gap confirms that the benefit of local image statistics is not a spurious improvement resulting from overfitting.

3.3 Predictive Accuracy of the Models

To assess the impact of local texture statistics on predictive power of the model, we used a leave-2-out fitting and prediction procedure (see Sec. 2.5), comparing models constructed with just spectral features, or spectral features augmented by local texture statistics. As in the assessment of explanatory power, the analysis was carried out both for mammograms and GS data, and for models fit by optimizing AIC or BIC. Bootstrapping with 20,000 resampled data arrays was used to determine 95% confidence intervals on the AUC and to perform inference on these comparisons.

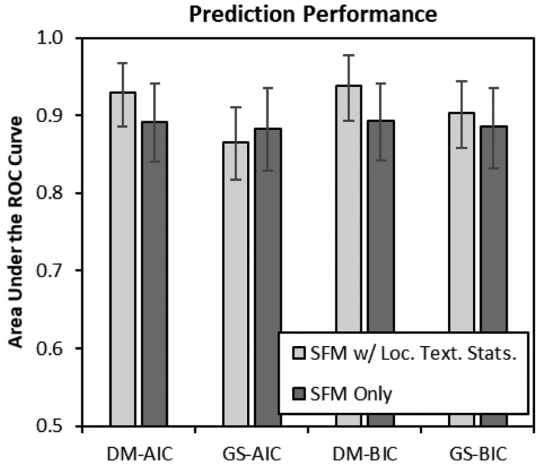

Fig. 8 Predictive accuracy. The plot shows AUC averaged across all held-out data, as a function of the figure of merit used in the stepwise regression procedure (AIC or BIC) and the type of data being evaluated, DM or GS. Results are plotted for the spectral features (SFM Only) and for the power-law features with local texture statistics. Error bars represent a 95% confidence interval over 20,000 bootstrap resamples.

Results are shown in Fig. 8. Overall discrimination accuracy is relatively high, with average AUC across conditions near 0.9 in all cases. For the mammograms (DM), there was an increase in predictive accuracy for models that included local texture statistics, compared to those that only included spectral features. For AIC, this difference is considered marginally significant ($p = 0.092$), and for BIC, the observed difference is significant ($p = 0.045$). For the GS data, the differences from adding local texture statistics are smaller (BIC) or nonexistent (AIC), and these differences are not significant (AIC: $p = 0.741$ and BIC: $p = 0.246$).

To further demonstrate that the value of local image statistics is specific to the structure of mammograms, we examined how the type of image (DMs or GSs) impacts discrimination performance of the models. For the spectral features with local-texture statistics, there is a significant increase in accuracy for statistics extracted from mammograms compared to the same statistics extracted from GSs for AIC ($p = 0.004$), and a marginal increase in accuracy for BIC ($p = 0.065$). The differences in accuracy for models based on spectral features alone, as expected, are not significant for either AIC or BIC ($p = 0.24$ and $p = 0.29$, respectively).

4 Discussion

Projection x-ray images, such as DMs, are formed by a process of transmission; this is in contrast with the processes of reflection and occlusion that form the images that make up the vast majority of our visual experience. This was recognized by Burgess and Judy[48] in a seminal paper that built on earlier works describing the power-law spectral model for breast anatomy in mammograms.[20,24,50] As Burgess and Judy showed, the transmissive nature of x-rays and the resulting superpositions of structure that occur in the image leads to a steeper "slope," i.e., a spatial frequency spectrum that is fit by power-law exponents near -3. This contrasts with the power spectrum of NSs, whose spatial frequency spectrum is well-fit by a power law with an exponent that is near -2.[51,52]

The findings in Fig. 6(a) show that, along with differences in the power spectrum, mammograms and NSs differ in their local texture statistics. These differences, which include pairwise, three-point, and four-point correlations, go beyond spectral characteristics: they are not captured by GSs whose power spectra match those of mammograms.

These observations, along with two other considerations, motivate our study of the relevance of these local statistics to radiologists' judgments of breast density. First, perceptual studies have shown that humans are sensitive to these statistical characteristics[31]—and interestingly, the pattern of sensitivity matches the importance of these statistics in NSs.[32] Second, as shown in Fig. 6(b), local image statistics differ in dense versus non-dense mammograms—and this difference is not fully accounted for by their spectral differences [Fig. 6(b) versus 6(c)]. However, while many studies in vision science have demonstrated that local image statistics can be detected,[26,28,31,53] there has not been a demonstration of the relevance of local image statistics to an important real-world task.

Here, we show that this is the case. Radiologist's breast-density judgments in DMs are substantially improved by augmenting the spectral feature set with local texture features. For the AIC, the improvement for adding the local texture statistics is nearly as large (77%) as the improvement found by adding the power-law features to a constant model that is entirely independent of image statistics. For the BIC, the advantage is smaller (48%) but still considerable. This stands in contrast to the GS images, where we would expect less information to be carried in the local texture statistics. In this case, there is a smaller benefit from adding the local texture statistics (33% for AIC and 23% for BIC).

The improvements in model fit are also reflected in the predictive accuracy of the stepwise regression models, with improvements in accuracy for including the local texture statistics that are marginally significant to significant. Given that the variables to be predicted (radiologist judgments) are themselves subject to within- and between-observer variability,[54] it is unlikely that any prediction approach could achieve an AUC of 1. For example, a previous study by Castella and colleagues[55] investigating breast density judgments found that rates of agreement of 82% to 91% for individual radiologists compared to a gold-standard consensus density score. Nevertheless, including local texture statistics closes almost half of the gap between the model based only on spectral features and perfect performance (AUC = 1).

The highest predictive accuracy is found using the more conservative BIC as the basis for training the model using stepwise regression. This results in a more parsimonious model (5 selected features) relative to using AIC as the basis for training (15 parameters). Thus, improved predictive accuracy with BIC may indicate some overfitting when using AIC.

4.1 Implications

The combination of model-fitting and predictive accuracy results suggests that radiologists access some information that is specifically contained in the local texture statistics and absent from the power-law spectral features, when making breast density judgments. This raises important questions for understanding radiologists' performance and for optimizing the images that they use to make diagnoses.

There are still many aspects of how radiologists perform complex tasks based on images that are unknown. This includes what features they use, how these features are combined into a clinical response, and how this process may vary between radiologists or even within a given radiologist at different times. A crucial step in better understanding radiologists, at a quantitative level, is identifying features that have explanatory and predictive value. This has been done fairly extensively for power-law spectral features, particularly the power-law exponent.[20–22,37,48] The results of this work demonstrate that there are other textural features that appear to capture the influence of visual texture on radiologists' breast density judgments specifically. The results here motivate a more thorough investigation of local texture statistics in this context.

The power-spectrum effects of image-processing steps that use frequency modulation to implement smoothing and/or contrast enhancements are well understood. However, it is not clear how such operations impact higher-order statistical properties of the images that are captured by the local texture statistics. These results serve to motivate characterization of these higher-order effects.

As the field increasingly turns to machine-learning approaches to automate image interpretation that is routinely performed by radiologists or other medical professionals, including breast density judgments, methods are needed to assure that the algorithms involved are making comparable findings. Demonstrating that these algorithms make similar use of features influencing breast density judgments can serve as a way to validate such algorithms.

4.2 Limitations

This study asks a specific question: whether a pre-determined class of local texture statistics, which capture visually salient non-Gaussian properties of breast images, can help explain and/or predict radiologists' breast density judgments in observational data from clinical practice. To this end, the analysis utilized a number of simplifications, such as compressing the four density categories into binary classification (dense or not-dense). More broadly, this study does not indicate how radiologists might access these local texture statistics, nor does it specify the extent to which these statistics will generalize to other kinds of image-based decisions and imaging modalities with different textural properties.

The results here serve as motivation for more focused investigations into these topics. There are many ways that the current investigation could be modified or extended to better understand how radiologists form judgments of breast density from screening mammograms. One promising direction would be understanding how texture statistics are combined across views or across the right and left breast. These investigations will likely require larger datasets than the 444 images used here. Nonetheless, we contend that the results of this study, which establish an association between radiologist breast-density judgments and non-Gaussian binary texture statistics, add relevance to such investigations.

5 Conclusion

Even in this relatively small sample of mammograms, it is apparent that local texture statistics derived from the parenchymal region play a role in judgments of mammographic density. Consistent with previous findings, parameters that characterize the power spectrum of breast anatomy (power-law slope, intercept, and total power) are found to have significant explanatory value in generalized linear models of breast density. However, the inclusion of local texture

statistics improves standard model-selection criteria (AIC and BIC). Furthermore, when logistic regression GLMs are used as predictive models in a leave-one-of-each-class out cross-validation, the addition of local texture statistics results in significantly more accurate prediction of breast-density judgments, as defined by the aggregate BI-RADS density scores.

Since local texture statistics capture aspects of images that are not captured by the power spectrum (a second-order moment), they are by definition a higher-order statistic. When the full-field DMs are replaced by Gaussian textures matched to the mean and power-spectrum, the local texture coefficients have substantially less explanatory/predictive value, demonstrating that higher-order statistical properties of the images are critical in the visual assessment of breast density. Taken in sum, the results of this study provide quantitative evidence that radiologists access higher-order images statistics when making judgments of breast density.

Disclosures

M.L.Z. acts as the PI for institutional funding from Hologic, Inc. C.K.A. is an advisor, consultant, and holds stock options for Izotropic, LLC and acts as a consultant to Canon Medical Systems USA.

Code and Data Availability

The data that support the findings of this article are not publicly available due to privacy concerns. The data and code used to analyze it can be requested from the author at ckabbey@ucsb.edu.

Acknowledgments

This work was supported by the US National Institutes of Health (Grant Nos. R21-CA241705, R01-EY07977, and R01-CA237827).

References

1. S. Ciatto et al., "Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories," *The Breast* **14**(4), 269–275 (2005).
2. A. Redondo et al., "Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms," *Br. J. Radiol.* **85**(1019), 1465–1470 (2012).
3. P. E. Freer, "Mammographic breast density: impact on breast cancer risk and implications for screening," *Radiographics* **35**(2), 302–315 (2015).
4. E. U. Ekpo et al., "Assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the BI-RADS atlas," *Am. J. Roentgenol.* **206**(5), 1119–1123 (2016).
5. K. Kerlikowske et al., "Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer risk," *J. Natl. Cancer Inst.* **99**(5), 386–395 (2007).
6. K. R. Brandt et al., "Comparison of clinical and automated breast density measurements: implications for risk prediction and supplemental screening," *Radiology* **279**(3), 710–719 (2016).
7. J. T. Schousboe et al., "Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness," *Ann. Internal Med.* **155**(1), 10–20 (2011).
8. E. F. Conant, B. L. Sprague, and D. Kontos, "Beyond BI-RADS density: a call for quantification in the breast imaging clinic," *Radiology* **286**(2), 401–404 (2018).
9. V. P. Jackson et al., "Imaging of the radiographically dense breast," *Radiology* **188**(2), 297–301 (1993).
10. V. Corsetti et al., "Breast screening with ultrasound in women with mammography-negative dense breasts: evidence on incremental cancer detection and false positives, and associated cost," *Eur. J. Cancer* **44**(4), 539–544 (2008).
11. C. B. Hruska, "Molecular breast imaging for screening in dense breasts: state of the art and future directions," *Am. J. Roentgenol.* **208**(2), 275–283 (2017).
12. M. F. Bakker et al., "Supplemental MRI screening for women with extremely dense breast tissue," *N. Engl. J. Med.* **381**(22), 2091–2102 (2019).
13. A. Vourtsis and W. A. Berg, "Breast density implications and supplemental screening," *Eur. Radiol.* **29**, 1762–1777 (2019).
14. O. Alonzo-Proulx et al., "Reliability of automated breast density measurements," *Radiology* **275**(2), 366–376 (2015).
15. A. Arieno, A. Chan, and S. V. Destounis, "A review of the role of augmented intelligence in breast imaging: from automated breast density assessment to risk stratification," *Am. J. Roentgenol.* **212**(2), 259–270 (2019).
16. C. Zhou et al., "Computerized image analysis: estimation of breast density on mammograms," *Med. Phys.* **28**(6), 1056–1069 (2001).

17. D. van der Waal et al., "Comparing visually assessed BI-RADS breast density and automated volumetric breast density software: a cross-sectional study in a breast cancer screening setting," *PLoS One* **10**(9), e0136667 (2015).
18. F. O. Bochud et al., "Importance of anatomical noise in mammography," *Proc. SPIE* **3036**, 74–81 (1997).
19. F. O. Bochud et al., "Estimation of the noisy component of anatomical backgrounds," *Med. Phys.* **26**(7), 1365–1370 (1999).
20. A. E. Burgess, F. L. Jacobson, and P. F. Judy, "Human observer detection experiments with mammograms and power-law noise," *Med. Phys.* **28**(4), 419–437 (2001).
21. L. Chen et al., "Anatomical complexity in breast parenchyma and its implications for optimal breast imaging strategies," *Med. Phys.* **39**(3), 1435–1441 (2012).
22. L. Chen, C. K. Abbey, and J. M. Boone, "Association between power law coefficients of the anatomical noise power spectrum and lesion detectability in breast imaging modalities," *Phys. Med. Biol.* **58**(6), 1663–1681 (2013).
23. J. J. Heine et al., "On the statistical nature of mammograms," *Med. Phys.* **26**(11), 2254–2265 (1999).
24. J. J. Heine et al., "Multiresolution statistical analysis of high-resolution digital mammograms," *IEEE Trans. Med. Imaging* **16**(5), 503–515 (1997).
25. C. K. Abbey et al., "Non-Gaussian statistical properties of breast images," *Med. Phys.* **39**(11), 7121–7130 (2012).
26. J. Freeman et al., "A functional and perceptual signature of the second visual area in primates," *Nat. Neurosci.* **16**(7), 974–981 (2013).
27. T. Caelli and B. Julesz, "On perceptual analyzers underlying visual texture discrimination: Part I," *Biol. Cybern.* **28**(3), 167–175 (1978).
28. B. Julész, E. Gilbert, and J. D. Victor, "Visual discrimination of textures with identical third-order statistics," *Biol. Cybern.* **31**(3), 137–140 (1978).
29. J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vision* **40**, 49–70 (2000).
30. Y. Yu, A. M. Schmid, and J. D. Victor, "Visual processing of informative multipoint correlations arises primarily in V2," *Elife* **4**, e06604 (2015).
31. J. D. Victor et al., "A perceptual space of local image statistics," *Vision Res.* **117**, 117–135 (2015).
32. A. M. Hermundstad et al., "Variance predicts salience in central sensory processing," *Elife* **3**, e03722 (2014).
33. H. Akaike, "A new look at statistical model identification," *IEEE Trans. Autom. Control* **19**, 716–723 (1974).
34. G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.* **6**, 461–464 (1978).
35. P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process Mag.* **21**(4), 36–47 (2004).
36. A. E. Burgess, "Mammographic structure: data preparation and spatial statistics analysis," *Proc. SPIE* **3661**, 642–654 (1999).
37. K. G. Metheany et al., "Characterizing anatomical variability in breast CT images," *Med. Phys.* **35**(10), 4685–4694 (2008).
38. I. Sechopoulos, K. Bliznakova, and B. Fei, "Power spectrum analysis of the x-ray scatter signal in mammography and breast tomosynthesis projections," *Med. Phys.* **40**(10), 101905 (2013).
39. J. Siewerdsen, I. Cunningham, and D. Jaffray, "A framework for noise-power spectrum analysis of multidimensional images," *Med. Phys.* **29**(11), 2655–2671 (2002).
40. N. J. Packard et al., "Effect of slice thickness on detectability in breast CT using a prewhitened matched filter and simulated mass lesions," *Med. Phys.* **39**(4), 1818–1830 (2012).
41. J. D. Victor and M. M. Conte, "Local image statistics: maximum-entropy constructions and perceptual salience," *J. Opt. Soc. Am. A* **29**(7), 1313–1345 (2012).
42. C. E. McCulloch, *Generalized Linear Models*, p. 452, Journal of the American Statistical Association (2000).
43. N. R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons (1998).
44. M. A. Efroymson, "Multiple regression analysis," in *Mathematical Methods for Digital Computers*, A. Ralston and H. S. Wilf, Eds., pp. 191–203, John Wiley and Sons, New York (1960).
45. J. Kuha, "AIC and BIC: comparisons of assumptions and performance," *Sociol. Methods Res.* **33**(2), 188–229 (2004).
46. S. Konishi and G. Kitagawa, Information Criteria and Statistical Modeling, Springer Series in Statistics, Springer (2008).
47. C. K. Abbey et al., "Evaluation of non-Gaussian statistical properties in virtual breast phantoms," *J. Med. Imaging* **6**(2), 025502 (2019).
48. A. E. Burgess and P. F. Judy, "Signal detection in power-law noise: effect of spectrum exponents," *J. Opt. Soc. Am. A* **24**(12), B52–B60 (2007).
49. G. Tkačik et al., "Natural images from the birthplace of the human eye," *PLoS One* **6**(6), e20409 (2011).
50. F. O. Bochud et al., "Detectability of radiological images: the influence of anatomical noise," *Proc. SPIE* **2436**, 156–164 (1995).

51. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A* **4**(12), 2379–2394 (1987).
52. B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Network Comput. Neural Syst.* **7**(2), 333–339 (1996).
53. J. D. Victor, M. M. Conte, and C. F. Chubb, "Textures as probes of visual processing," *Annu. Rev. Vision Sci.* **3**, 275–296 (2017).
54. J. M. Boone et al., "A breast density index for digital mammograms based on radiologists' ranking," *J. Digit. Imaging* **11**(3), 101–115 (1998).
55. C. Castella et al., "Semiautomatic mammographic parenchymal patterns classification using multiple statistical features," *Acad. Radiol.* **14**(12), 1486–1499 (2007).

Craig K. Abbey studies the transfer of diagnostic information to observers in medical imaging using analysis of image statistics as well as visual psychophysics. He received his PhD in applied mathematics from the University of Arizona in 1998 and was a postdoctoral fellow in medical physics at Cedars-Sinai Medical Center and UCLA from 1998 to 2001. He is currently a researcher in the Department of Psychological and Brain Sciences at UC Santa Barbara.

Margarita L. Zuley, MD, FACR, FSBI is a diagnostic radiologist specializing in breast imaging for early detection of cancer and a Professor of radiology at the University of Pittsburgh. She received her BS degree from the University of Notre Dame in 1987 and MD degree from the University of Pittsburgh School of Medicine in 1991. She currently serves as vice chair of quality and chief of breast imaging in the Department of Radiology at the University of Pittsburgh Medical Center. Her research utilizes both qualitative and quantitative methods to investigate improving reader performance in breast imaging though the application of new technologies and practice interventions.

Jonathan D. Victor received his BA degree in mathematics from Harvard College in 1973, his PhD from The Rockefeller University in 1979, and his MD degree from Cornell University Medical College in 1980. He is currently professor of neuroscience and Fred Plum Professor of Neurology in the Feil Family Brain and Mind Research Institute at Weill Cornell Medicine. His research interests include active sensation, the statistical properties of sensory signals, and development and application of novel signal processing methods.