Investigating the link between radiologists' gaze, diagnostic decision, and image content

Georgia Tourassi,[1] Sophie Voisin,[1] Vincent Paquit,[1] Elizabeth Krupinski[2]

[1]Oak Ridge National Laboratory, Biomedical Science and Engineering Center, Oak Ridge, Tennessee, USA
[2]Department of Medical Imaging, University of Arizona, Tucson, Arizona, USA

Correspondence to
Dr Georgia D Tourassi, Oak Ridge National Laboratory, Biomedical Science & Engineering Center, 1 Bethel Valley Road, P.O. Box 2008, Oak Ridge, TN 37831-6085, USA; tourassig@ornl.gov

Received 16 November 2012
Revised 25 April 2013
Accepted 25 May 2013
Published Online First 20 June 2013

ABSTRACT

Objective To investigate machine learning for linking image content, human perception, cognition, and error in the diagnostic interpretation of mammograms.

Methods Gaze data and diagnostic decisions were collected from three breast imaging radiologists and three radiology residents who reviewed 20 screening mammograms while wearing a head-mounted eye-tracker. Image analysis was performed in mammographic regions that attracted radiologists' attention and in all abnormal regions. Machine learning algorithms were investigated to develop predictive models that link: (i) image content with gaze, (ii) image content and gaze with cognition, and (iii) image content, gaze, and cognition with diagnostic error. Both group-based and individualized models were explored.

Results By pooling the data from all readers, machine learning produced highly accurate predictive models linking image content, gaze, and cognition. Potential linking of those with diagnostic error was also supported to some extent. Merging readers' gaze metrics and cognitive opinions with computer-extracted image features identified 59% of the readers' diagnostic errors while confirming 97.3% of their correct diagnoses. The readers' individual perceptual and cognitive behaviors could be adequately predicted by modeling the behavior of others. However, personalized tuning was in many cases beneficial for capturing more accurately individual behavior.

Conclusions There is clearly an interaction between radiologists' gaze, diagnostic decision, and image content which can be modeled with machine learning algorithms.

BACKGROUND AND SIGNIFICANCE

Cognitive science is an important driver for innovation in the field of biomedical informatics, providing the framework for designing, developing, and properly assessing medical information technologies that meet the evolving needs of health professionals.[1 2] For example, advances in human perception understanding have been influencing medical image display technologies,[3–6] while studies of human reasoning and problem solving strategies have been guiding the development of medical decision support technology.[2] The importance of merging cognitive and information sciences has also been recognized by the Board of the American Medical Informatics Association (AMIA) in a recent white paper drafting the core competencies for graduate education in biomedical informatics.[7]

In medical imaging informatics specifically, there have been great technological advances in both image data visualization and interpretation.[8] Yet, these advances have not translated into substantial reduction of medical error which has puzzled the medical community. In 1949, Garland estimated an average radiologic error rate of 30% in his landmark work[9 10] and since then error rates have remained virtually unchanged.[11] Actually, diagnostic error and inconsistencies among physicians in the interpretation of medical images remain two of the biggest challenges in radiology practice.[11 12] Furthermore, the increasing volume and complexity of medical imaging data generated today exacerbate radiologists' visual strain and cognitive fatigue, thus increasing the risk of medical error.[13 14]

Imaging-based screening programs such as breast cancer screening are particularly vulnerable to error and variability due to the low disease prevalence and the large volumes of imaging studies generated daily. For example, approximately 37 million mammograms are performed annually in the USA and an estimated 1.2 million additional women become eligible for screening each year.[15] With breast cancer prevalence at 0.5% in a typical screening population, searching for signs of cancer in mammograms is one of the most challenging tasks in radiology.[16] Reportedly, up to 30% of breast lesions go unreported in screening mammograms and as many as 65% of those missed lesions are deemed visible on retrospective review.[17–19] Studies repeatedly show that specialized training and experience make a big difference. Breast imaging experts and radiologists who read regularly high volumes of mammograms are significantly more accurate than general radiologists for whom mammography is a small part of their clinical workload.[20–22]

There is a significant body of work aiming to understand the nature of diagnostic error in radiology. Eye-tracking researchers have attempted to understand better the visual search process[23–29] and the sources of diagnostic error with respect to the clinical task at hand.[30 31] These studies clearly show that eye movements correlate with radiologists' diagnostic decisions. Furthermore, the same studies verified that the three types of diagnostic errors found in chest imaging[32] occur in mammography as well: search errors (ie, the radiologist fails to fixate on the cancer), recognition errors (eg, the radiologist fixates on the cancer but for short duration and fails to recognize it), and decision making or cognitive errors (eg, the radiologist fixates on the cancer for long duration but decides not to report it). Search and recognition errors are also known as perceptual errors.[33] Reportedly, perceptual errors and cognitive errors are equally prevalent in radiology, including breast imaging.[33 34] These findings inspired a whole new research area targeting the development of decision support

To cite: Tourassi G, Voisin S, Paquit V, *et al*. *J Am Med Inform Assoc* 2013;**20**:1067–1075.

systems for mammography. For example, one study suggested that the radiologists' accuracy could be improved if they are given feedback about areas of prolonged fixation,[35] an idea proposed and supported in chest radiography as well.[36–39] Other studies proposed the development of tailored image processing algorithms focusing on spectral features of breast abnormalities that are typically missed by radiologists.[40–42] Although these studies are a step towards the right direction, they have a fundamental limitation: they are based on the premise that radiologists make similar errors. This assumption is inconsistent with clinical experience. Radiologists have diverse perceptual and cognitive patterns when reading medical images. But even when radiologists reach the same diagnosis, it may be due to different reasons. For example, two radiologists may miss the same breast cancer, but one due to errors of search while the other due to errors of reasoning. Therefore, these two radiologists would benefit from different types of training and decision support to improve their accuracy. Understanding individual differences in human perception and cognition of medical imaging data can provide important insights in the development and successful use of clinical decision support and education support information systems that meet the personal needs of clinicians involved in the interpretation of medical images.

OBJECTIVE

Our overarching goal in this study is to leverage eye-tracking, user-modeling, and machine learning (ML) to elucidate the potential link between medical image content and the perceptual and cognitive behavior a radiologist displays when viewing an image. We approach the problem in a systematic way with a series of research questions that aim to: (i) associate localized image characteristics with radiologists' perceptual and cognitive behaviors; and (ii) investigate whether such errors can be predicted reliably by integrating radiologists' gaze metrics, medical decisions, and image content. Furthermore, our study explores whether leveraging common perceptual and cognitive patterns observed in a group of radiologists is sufficient to predict an individual radiologist's pattern or whether personalized modeling is a more effective approach.

MATERIALS AND METHODS

Image database

This study investigated the potential link between radiologists' gaze behavior, diagnostic decisions, and image content for the specific task of mass detection in screening mammograms. Masses, microcalcifications, architectural distortions, and focal asymmetries are the most common manifestations of breast cancer. Masses comprise the overwhelming majority of missed breast cancers due to their diverse range of shapes, sizes, and contrast.[18 19] Moreover, mammographic masses can be extremely subtle, often obscured by normal breast parenchyma.

To perform this study 20 screen-film mammograms were selected from the Digital Database of Screening Mammography (DDSM).[43] The DDSM contains 2500 screening mammograms along with associated patient and image information, including ground truth established via biopsy, additional imaging, or 2-year follow-up. DDSM cases from the Lumisys volumes were randomly selected. Higher selection priority was given to challenging cases containing subtle masses and denser breasts, according to the mass subtlety and parenchymal density ratings included in the DDSM truth files. Of the 20 study cases, 10 included 14 biopsy-proven malignant masses, five cases included five biopsy-proven benign masses, and the remaining five cases were normal as determined during a 2-year cancer-free follow-up patient evaluation. The benign cases included challenging masses requiring further follow-up according to the American College of Radiology (ACR) guidelines. Thus, the dataset did not include any easy, 'benign–without-callback' cases. Table 1 provides the list of the selected DDSM cases along with a brief description for each one.

All mammograms were first preprocessed with standard image processing techniques for optimized softcopy display. First, the grayscale histogram of each mammogram was analyzed to identify the gray level distribution of the breast region and the background. Then, a window and level function was applied to the mammograms using a sigmoid curve to balance the need for contrast at the breast center with visualization of the breast skin line. An experienced radiologist with specialized training in breast imaging and 11 years of experience in mammographic interpretation visually assessed and approved the quality of the processed images. This radiologist did not participate any further in the study. The images were displayed in single view mode (ie, only one view of the breast was shown) on a Totoku 5 M-pixel LCD monitor. The monitor was calibrated to the DICOM Gray Scale Display Function Standard. Ambient room lights were turned off for viewing.

Eye-tracking data collection

Institutional review board approval was obtained prior to the study. Six readers were asked to view the cases and report the location of any suspicious masses as they typically do in clinical practice. Three readers were experienced MQSA-certified breast imagers while the other three were fourth year radiology residents with at least one rotation in mammography. During the reading sessions the readers wore an H6 head-mounted eye tracker, with 60 Hz sampling speed, and eye-head integration from Applied Science Laboratories (ASL, Bedford, Massachusetts, USA). The eye-tracker recorded each reader's eye position data within 1° of accuracy. The readers were instructed to view each case until they were satisfied with the viewing phase. When the readers were ready to give their diagnostic opinion, the eye-tracking recording phase was halted until the readers recorded their findings and they were ready to proceed with viewing the next case. Prior to the study, each reader was carefully calibrated using the 9-point calibration protocol provided by ASL.

The raw eye-position data were analyzed using a spatio-temporal clustering algorithm that is well established and commonly used in radiology.[45] First, the x, y coordinates of the eye-position data points were grouped sequentially according to a running mean distance calculation having an average of 0.5° radius threshold to determine fixations.[46] These fixations were then grouped into clusters; that is, circular areas with 2.5° radius centered at the mean x, y location of the group of fixations contributing sequentially to a cluster. This radius corresponds to the useful human visual field.[46 47] Finally, cumulative clusters were calculated by combining fixation clusters generated when the radiologist re-fixated the same image area at any point in time. A re-fixation cluster contributes to a cumulative cluster if they overlap by at least 50%. The new centroid was defined by averaging the contributing clusters. If the centroid of a true mass was within a cumulative fixation cluster, then the fixation cluster was attributed to the corresponding mass lesion.

Three gaze metrics were extracted per cumulative fixation cluster: (i) total dwell time per cumulative fixation cluster (*dwell*), (ii) time from beginning of the case reading until the reader fixated on the reported image location for the first time (*initial*), and (iii) number of times the reader returned to the

Table 1 Description of study cases

Case	Ground truth	Patient age	Breast density	Mass description
B3021	Malignant	59	Heterogeneous	(1) Shape: lobulated, Margin: ill-defined, Subtlety: 3
	Malignant			(2) Shape: lobulated, Margin: ill-defined, Subtlety: 3
	Malignant			(3) Shape: lobulated, Margin: ill-defined, Subtlety: 3
B3054	Malignant	71	Heterogeneous	Shape: irregular, Margin: speculated, Subtlety: 4
B3070	Malignant	56	Heterogeneous	Shape: irregular—architectural distortion, Margin: spiculated, Subtlety: 2
B3072	Malignant	38	Heterogeneous	Shape: irregular, Margin: ill-defined, Subtlety: 3
B3376	Malignant	65	Heterogeneous	Shape: irregular- architectural distortion, Margin: obscure-spiculated, Subtlety: 3
C0142	Malignant	60	Dense	Shape: architectural distortion, Margin: ill-defined, Subtlety: 3
C0149	Malignant	57	Fibroglandular	Shape: oval, Margin: obscure, Subtlety: 1
C0157	Malignant	62	Fatty	(1) Shape: oval, Margin: microlobulated, Subtlety: 5
	Malignant			(2) Shape: oval, Margin: microlobulated, Subtlety: 5
	Malignant			(3) Shape: oval, Margin: microlobulated, Subtlety: 5
C0162	Malignant	40	Heterogeneous	Shape: irregular, Margin: ill-defined, with associated calcifications, Subtlety: 4
C0339	Malignant	44	Heterogeneous	Shape: round, Margin: spiculated, Subtlety: 5
B3122	Benign	65	Fibroglandular	Shape: round, Margin: microlobulated, Subtlety: 3
B3151	Benign	60	Heterogeneous	Shape: lobulated, Margin: circumscribed, Subtlety: 3
C0274	Benign	54	Fibroglandular	Shape: oval, Margin: circumscribed, Subtlety: 5
C0286	Benign	38	Fibroglandular	Shape: round, Margin: obscured, Subtlety: 4
C0311	Benign	46	Fibroglandular	Shape: round, Margin: obscured, Subtlety: 3
B3616	Normal	50	Fibroglandular	N/A
B3633	Normal	62	Fatty	N/A
B3660	Normal	41	Fibroglandular	N/A
B3673	Normal	39	Heterogeneous	N/A
B3677	Normal	75	Fibroglandular	N/A

The breast density, mass shape, and mass margin descriptors are according to the BI-RADS (Breast Imaging Reporting and Data System) lexicon established by the American College of Radiology.[44] The subtlety rating ranges from 1 (subtle lesion) to 5 (obvious lesion).

particular image location for additional viewing (*returns*). Locations that attracted prolonged viewing (>1 s) were considered locations of long dwell. This cut-off threshold has been suggested before as appropriate for mammographic studies since it has been shown to identify 80% of true positive (TP) decisions and 65% of false negative (FN) decisions.[48] Locations that attracted some viewing (<1 s) were considered short dwells. All other locations that did not attract any viewing were treated as locations of no dwell.

Local image analysis

Fixed size (512×512 pixels) regions of interest (ROIs) were automatically extracted around the centroid of each cumulative fixation cluster. In addition, ROIs were extracted around masses that attracted no visual dwell. Six texture image signatures were calculated for each ROI: two based on first order statistics (*entropy*, *SD*) and four based on second order statistics, also known as Haralick features[49 50] (*contrast*, *correlation*, *energy*, *homogeneity*). The Haralick features were calculated from the ROI's gray-level co-occurrence matrix for a single distance (d=1 pixel) and four angular directions (θ=0°, 45°, 90°, 135°). The MATLAB Image Processing Toolbox (The MathWorks, Natick, Massachusetts, USA) and standard MATLAB functions were used to extract the above signatures.

The ROIs were also analyzed using the Gabor wavelet analysis method due to its perceptual relevance.[51] Reportedly, Gabor filters model the spatial frequency and orientation responses of simple cells in the primary visual cortex.[52 53] Specifically, we implemented the Gabor analysis method introduced by Manjunath *et al*.[54] The principle of the specific method is as follows. First, the Gabor filter bank is generated by selecting (1) its size equal to the ROI size, (2) the spatial bandwidth, (3) the number of levels for the multiresolution decomposition, and (4) the filter orientations. The number of filters per bank is equal to the number of decomposition levels times the number of dimensions. In our case we had 24 filters as we relied on 4 decomposition levels and 6 orientations, which is the same configuration employed in Manjunath and Ma.[54] The spatial bandwidth was also fixed according to Manjunath and Ma,[54] under the assumption that those parameters are optimal for many scenarios. There was only a single scale investigated, equal to the ROI size. Each ROI was processed to generate two texture features: the mean μ_{ij} and the SD σ_{ij}, for i=4 decomposition levels and j=6 orientations (μ_{11}, σ_{11}, ...,μ_{46}, σ_{46}), both aimed at describing locally homogeneous textures. Thus, this process resulted in 48 Gabor signatures extracted in total (2 features × 4 decomposition levels × 6 orientations). In total, each ROI_i was described by a feature vector f ($ROI_{i,j}$) which was constructed using 54 local image texture signatures (2 first-order+4 Haralick+48 Gabor) and 3 gaze metrics for the specific reader j who viewed the mammographic case containing the specific ROI_i.

Coding diagnostic errors

Each $ROI_{i,j}$ was associated with a diagnostic decision (*yes* or *no*), depending on if the specific reader j decided to report the image location as suspicious of containing a mass lesion (*yes*) or not (*no*). Depending on the reader's decision and the ground truth associated with the ROI, there were four possible types of decisions collected: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

Experimental design

The collected feature vectors were analyzed using various ML algorithms with and without feature selection to elucidate the

link between image content, human perception, human cognition, and human error within the context of the specific clinical task. Specifically, three study hypotheses were pursued in a hierarchical way to test progressively the link between (i) image content and gaze, (ii) image content, gaze, and cognition, and (iii) image content, gaze, cognition, and medical error.

- *Hypothesis 1*: Local image content of a mass-containing mammographic region can be used to predict whether a breast mass will attract prolonged dwell from the readers (thus, image content is predictor of gaze behavior).
- *Hypothesis 2*: Local image content and gaze metrics can be used to predict readers' diagnostic decisions (thus, image content and gaze behavior are predictors of cognitive behavior).
- *Hypothesis 3*: Local image content, gaze metrics, and cognitive opinions can be used to predict readers' erroneous decisions (thus, image content, gaze behavior, and diagnostic opinion are predictors of human error).

For each study hypothesis, different ML algorithms were developed and tested using the leave-one-case-out and/or the leave-one-reader-out cross-validation sampling schemes, depending on what was appropriate and feasible. Initially, the data collected from all readers were grouped and analyzed as one dataset to determine whether group-based predictive modeling is indeed possible. If yes, then a radiologist's perceptual and cognitive behavior could be predicted reliably by observing and modeling the behavior of many other radiologists and using such a model to predict the behavior of a new (unobserved) individual. However, as discussed earlier, inconsistencies among radiologists in the diagnostic interpretation of medical images are well documented. Therefore, user-specific modeling was also explored to determine whether personalized models capture more effectively individual behavior than a group-based model. All models were developed and validated using the WEKA data mining package (University of Waikato, New Zealand).[55] Predictive performance was assessed in terms of receiver operating characteristics (ROC) analysis[56] using the JROCKIT software.[57] ROC analysis is preferred when a performance metric independent of decision threshold and class prevalence is necessary for meaningful comparison among predictive models. The area under the ROC curve (AUC) estimated by JROCKIT was the selected performance index.

RESULTS

Readers' diagnostic accuracy

Table 2 summarizes the readers' decisions. Note that readers 1–3 are the radiology residents and readers 4–6 are the breast imaging experts. The table groups the readers' decisions according to their total dwell classification (*long*, *short*, *no*). For example, reader 3 missed five masses in total, three malignant and two benign. Thus, this reader made three FN decisions related to malignant masses and two FN decisions related to benign masses. Of the three malignant masses missed, two were visually scrutinized for a long time, suggesting a cognitive error, while the third one was viewed for a short time, suggesting recognition error. Similarly, one of the benign masses missed by reader 3 did not attract any dwell at all, suggesting a search error. On the other hand, reader 2 marked three normal locations as masses (ie, two FPs). Of those two overcalls, one FP corresponded to a location that attracted long dwell, which suggests cognitive error. The second FP corresponded to a location of short dwell, suggesting a recognition error. A possible explanation is that the breast parenchyma in the particular location resembled an obvious mass to reader 2, who immediately called it without prolonged viewing and deliberation. In addition, there were no TP or FP decisions made for locations that attracted no dwell, as expected since a reader does not mark a location that has not gazed at all.

Table 2 Correct (TP, TN) and erroneous (FP, FN) decisions grouped according to the dwell time of the reader who made the corresponding decision

Reader	Dwell	Malignant	Benign	Normal
1	Long	9 TPs+2 FNs	2 TPs+0 FNs	12 TNs+1 FP
	Short	3 TPs+0 FNs	3 TPs+0 FNs	*+1 FP
	No	0 TPs+0 FNs	0 TPs+0 FNs	*+0 FPs
2	Long	6 TPs+0 FNs	4 TPs+0 FNs	15 TNs+1 FP
	Short	5 TPs+2 FNs	1 TPs+0 FNs	*+1 FP
	No	0 TP+1 FN	0 TPs+0 FNs	*+0 FP
3	Long	7 TPs+2 FNs	2 TPs+1 FN	16 TNs+0 FPs
	Short	4 TPs+1 FN	1 TP+0 FNs	*+0 FPs
	No	0 TP+0 FNs	0 TPs+1 FN	*+0 FPs
4	Long	10 TPs+1 FN	3 TPs+0 FNs	16 TNs+2 FPs
	Short	2 TPs+0 FNs	2 TPs+0 FNs	*+0 FPs
	No	0 TPs+1 FN	0 TPs+0 FNs	*+0 FPs
5	Long	8 TPs+0 FNs	3 TPs+0 FNs	14 TNs+1 FP
	Short	6 TPs+0 FNs	2 TPs+0 FNs	*+0 FPs
	No	0 TPs+0 FNs	0 TPs+0 FNs	*+0 FPs
6	Long	9 TPs+0 FNs	4 TPs+0 FNs	10 TNs+1 FP
	Short	3 TPs+2 FNs	1 TPs+0 FNs	*+0 FPs
	No	0 TPs+0 FNs	0 TPs+0 FNs	*+0 FPs

*Indicates all normal breast parenchyma which the radiologist spent <1 s viewing and correctly decided not to report.
FN, false negative; FP, false positive; TN, true negative; TP, true positive.

Overall the readers' diagnostic accuracy varied between 78.6% and 100% for malignant masses (mean=85.7%, 95% CI 79.5% to 91.9%) and was 80–100% for benign masses (mean=96.7%, 95% CI 90.1% to 100%). Their case recall rate varied between 0% and 10% (mean=6.7%, 95% CI 3.5% to 9.9%). The observed performance is consistent with what is reported in the literature. Although the residents' detection accuracy was on average lower than that of the experts, their recall rate was on average the same.

Experiment 1: predicting perceptual behavior from image content

Hypothesis 1 focused on the 19 breast masses present in our dataset. The purpose of this hypothesis was to determine the group-based (ie, global) and user-based (ie, individual) links between the image signatures of breast masses and the readers' tendency to visually dwell on the masses for a prolonged period.

To assess the global component, we explored various ML algorithms for predicting whether a particular mass attracted prolonged dwell from the majority of the study participants (ie, at least four readers). This experiment was done using a leave-one-case-out sampling scheme. In other words, starting with the set of 19 masses, each mass was excluded once to serve as a test case. A predictive model was developed using the remaining 18 masses and then tested on the one left out. The same process was repeated 19 times until each mass in the dataset served as a test case. The test results were aggregated to derive the predictive accuracy of the developed model. The predictive power of each group of textural signatures (statistical, Gabor) was evaluated separately. To assess the individual component, ML predictive models were explored for each reader's data separately.

Table 3 provides details on the architecture and performance of the best performing global and personalized models for this experiment. The best performing predictive models varied in terms of selected features and ML algorithms. Overall, the results suggest that readers' general gaze behavior for breast masses can be predicted to a good extent using image features. Stepwise forward feature selection combined with an Adaboost classifier and a Radial Basis Function (RBF) network as weak learner produced the best results; 73.5% overall classification accuracy (83.3% accuracy for long dwells vs 57.1% accuracy for short dwells) and ROC area of 0.820±0.112. Energy was the single most useful feature in predicting the perceptual behavior of the participating readers when analyzed collectively as a group. Personalized modeling was also effective, producing user models with AUC ranging from 0.667±0.127 (reader 5) to 0.867±0.108 (reader 3). Even though there were notable differences among the predictive accuracy of the individual models, these differences were not statistically significant. The lack of significant difference could be attributed to the small sample size of mammographic cases. Furthermore, the individual reader models developed for the breast imaging radiologists had lower AUC metrics than those developed for the radiology residents. These differences were not statistically significant either, but the general trend suggests that the gaze behavior of radiologists may be more complex than that of the residents, and the image signatures used in our study were not sufficient to capture such complexity as well as they did for the radiology residents. Finally, the most predictive image features varied among the individual models, suggesting that different image features are important for capturing effectively the perceptual behavior of individuals.

Experiment 2: predicting cognitive behavior from image content and perceptual behavior

Hypothesis 2 focused on the readers' diagnostic decisions (ie, to report a particular image location or not) and their potential link to image content and visual gaze. As with the previous experiment, different predictive models were explored in the WEKA environment to determine a group model as well as personalized user models. The group model was evaluated both with a leave-one-case-out as well as a leave-one-reader-out cross-validation sampling scheme. The group model would be relevant in a specific clinical setting built using data collected from all practicing radiologists in that setting. Then the developed model could be deployed for use but it would be applicable to the specific radiologists who provided the training data. On the other hand, the leave-one-reader-out sampling scheme was implemented to investigate whether an individual's performance could be predicted from a group-based understanding of the user community. The leave-one-reader-out sampling scheme works as follows. A separate model is developed using data from five readers. The model is then tested on data collected from the sixth reader who was left out during training. The process is repeated six times so that each reader serves once for testing. If effective, such a model could be far more useful in clinical practice since it assumes that an individual's behavior

Table 3 Predictive accuracy of WEKA-generated classifiers for predicting readers' gaze behavior with respect to breast masses

Model	Features	Classifier	ROC area
All (*group model*)	Gabor	RotationForest	0.766±0.121
	Statistical	RandomForest	0.703±0.123
	All	MultilayerPerceptron	0.644±0.154
	Best: energy	Adaboost w/RBF	0.820±0.112
Reader 1	Gabor	Logistic	0.612±0.130
	Statistical	DecisionStump	0.731±0.178
	All	Logistic	0.612±0.130
	Best: μ_{23}	Logistic	0.821±0.102
Reader 2	Gabor	Bagging	0.589±0.139
	Statistical	MultilayerPerceptron	0.684±0.121
	All	Logistic	0.751±0.116
	Best: μ_{22}, μ_{23}	AdaBoost w/RBF	0.865±0.097
Reader 3	Gabor	RotationForest	0.791±0.119
	Statistical	NaiveBayes	0.561±0.169
	All	BayesNet	0.747±0.141
	Best: μ_{15}, μ_{25}	Adaboost w/RBF	0.876±0.108
Reader 4	Gabor	Logistic	0.523±0.128
	Statistical	Bagging	0.481±0.182
	All	Logistic	0.518±0.127
	Best: μ_{11}, μ_{15}, μ_{16}, energy	Bagging w/RBF	0.741±0.141
Reader 5	Gabor	DMNBtext	0.227±0.112
	S	DMNBtext	0.351±0.129
	A	DMNBtext	0.312±0.125
	Best: energy	Adaboost w/RBF	0.667±0.127
Reader 6	Gabor	MultilayerPerceptron	0.746±0.124
	Statistical	NaiveBayes	0.712±0.119
	All	RotationForest	0.778±0.132
	Best: μ_{15}, μ_{26}, entropy	Adaboost w/ BLR	0.767±0.116

BLR, Bayesian logistic regression; DMNBtext, discriminative multinomial naive Bayes classifier; RBF, radial basis function network; ROC, receiver operating characteristics.

can be predicted from observing past behavior of other radiologists, without explicit knowledge of the individual's past behavior. Finally, individualized models were also developed and cross-validated using a leave-one-case-out sampling scheme as in experiment 1, using data collected from the specific reader in question. The results are shown in table 4. Stepwise forward feature selection was also included in experiment 2 to determine the relative contribution of image content and human gaze features.

All models achieved their highest ROC performance using gaze features and a mix of image signatures that were common among many models. The high ROC performance observed consistently with leave-one-reader-out cross-validation suggests that there are similarities among readers' cognitive behavior. Therefore, pooling data together from multiple readers for group-based modeling is reasonable and it provides several advantages due to the increased sample size available for model development. For example, this was beneficial in the case of Reader 4 whose cognitive behavior was captured more effectively when leveraging the larger amount of data collected from other readers (AUC=0.919±0.051) rather than relying only on his data (AUC=0.799±0.073). However, there are some notable differences as well. Reader 6 was modeled better by using personal data (AUC=0.891±0.067) rather than data collected from other readers (AUC=0.808±0.081). Even though the difference was not statistically significant, the difference is substantial, suggesting that there are some unique characteristics in this reader's decision-making process which could not be captured as adequately by leveraging data from the other five readers only.

Experiment 3: predicting diagnostic errors from image content, perceptual behavior, and cognitive behavior

The last experiment focused on linking image content and readers' perceptual and cognitive characteristics with erroneous diagnoses (ie, either an FN one when failing to report a true breast mass or an FP one when overcalling a normal breast finding as mass). The small number of erroneous decisions made by the readers (1 to 5 per person, 22 in total) posed two limitations with this experiment. It was not feasible to explore individual models or feature selection. Only group models were explored with both leave-one-case-out and leave-one-reader-out cross-validation. The results are reported in terms of percentage accuracy. Due to the very small sample size of the erroneous decisions per reader, ROC analysis was not feasible for the leave-one-reader-out cross-validation. Table 5 summarizes the results of this experiment.

The best ML algorithm for all models was the J48 decision tree. Overall, the leave-one-case-out group model demonstrated high predictive accuracy with an ROC area index of 0.929 ±0.038. The group predictive model was able to classify correctly 59% of the diagnostic errors and 97.3% of the correct diagnoses made by the readers. The leave-one-reader-out evaluation further confirmed the ability to predict an individual reader's correct and erroneous decisions by observing and modeling the error making patterns of other readers.

DISCUSSION

This study addressed the general problem of clinical error in the diagnostic interpretation of medical images and it proposed an ML approach for studying the interaction between the case under review and the clinician assigned to review the case. A series of experiments were performed to understand the possible link between image content, perception, cognition, and error in the radiology domain, and specifically in the context of cancer detection in screening mammography. The extent and nature of diagnostic error in mammography has been the focus of much research in the past. Therefore, mammography was an excellent application domain for the study we presented here.

In an earlier work we demonstrated that radiologists' individual error making patterns could be captured to a good extent by analyzing image characteristics that are visually extracted and recorded by the radiologists.[58 59] Those earlier studies were based only on radiology residents who were asked to decide the malignancy status of breast masses. In this study, we expanded the scope including (i) perceptual, cognitive, and error-making aspects of human behavior, (ii) a broader community of radiologists, (iii) automated image content analysis, and (iv) a different clinical task (detection rather than characterization). The underlying hypothesis of the study was that by monitoring the radiologist's gaze pattern for the specific case and integrating human gaze characteristics with image content is a promising way to infer radiologists' perceptual and cognitive behavior and whether the radiologist is at risk of making a diagnostic error.

Table 4 Predictive accuracy of WEKA-generated classifiers for predicting readers' decisions to report a mammographic location as suspicious based on image and gaze features

Model	Selected features	Classifier	ROC area
Group Model_LoCo	Gaze+Contrast+Correlation+Energy+μ_{22}+μ_{31}	Bayesian Network	0.900±0.022
Group Model_LoRo_R1	Gaze+Contrast+Entropy+Correlation+μ_{12}+μ_{22}	RBF	0.929±0.050
Group Model_LoRo_R2	Gaze+Contrast+Entropy+Energy+μ_{22}+μ_{31}	Bayesian Network	0.888±0.056
Group Model_LoRo_R3	Gaze+Correlation+Contrast+Energy+μ_{12}+μ_{22}+σ_{44}	RBF	0.872±0.056
Group Model_LoRo_R4	Gaze+Correlation+Energy+ μ_{22}+μ_{31}	Bayesian Network	0.919±0.051
Group Model_LoRo_R5	Gaze+Contrast+Correlation+Energy+μ_{12}+μ_{22}	Adaboost w/ MLP	0.907±0.051
Group Model_LoRo_R6	Gaze+Contrast+Correlation+Energy+μ_{12}+μ_{22}+σ_{32}	MLP	0.808±0.081
Individual_Model_R1	Gaze	Adaboost w/ MLP	0.927±0.050
Individual_Model_R2	All features	MLP	0.864±0.061
Individual_Model_R3	Dwell+Returns	NaiveBayes	0.766±0.082
Individual_Model_R4	All features	MLP	0.799±0.073
Individual_Model_R5	All features	MLP	0.919±0.046
Individual_Model_R6	All features	MLP	0.891±0.067

Gaze, *dwell+initial+returns*; LoCo, leave-one-case-out; LoRo, leave-one-reader-out; RBF, radial basis function; ROC, receiver operating characteristics.

Table 5 Accuracy of WEKA-generated classifiers for predicting readers' diagnostic errors

Model	% Accuracy	
	Correct DX	Wrong DX
Group Model_LoCo	97.3	59 (13/22)
Group Model_LoRo_R1	100	50 (2/4)
Group Model_LoRo_R2	100	60 (3/5)
Group Model_LoRo_R3	93.9	40 (2/5)
Group Model_LoRo_R4	97	100 (1/4)
Group Model_LoRo_R5	92.6	66.7 (0/1)
Group Model_LoRo_R6	85.2	66.7 (2/3)

LoCo, leave-one-case-out; LoRo, leave-one-reader-out.

We tested the hypothesis by performing a series of experiments for the problem of mass detection in mammograms, a well-known challenge in breast cancer screening.

The first experiment drew the link between image content and radiologists' perception. We found that indeed radiologists' perceptual behavior can be predicted by local image content. Furthermore, there appear to be both a global and an individual component when modeling such behavior. The global component was adequately captured with a single texture feature (ie, energy), which was able to predict radiologists' general tendency to dwell or not on a particular image location. Individual modeling appeared to offer advantages for some radiologists. However this finding was not consistent. To some extent this inconsistency could be attributed to sample size; some radiologists may require a larger number of cases to be properly modeled than others.

The second experiment furthered the link from image content and radiologists' perception to radiologists' cognition. The experimental results showed that cognitive behavior can be predicted to a good extent by observing gaze behavior. In many cases combining image content offered a clear advantage. This experiment also suggested that there is a global as well as an individual component when modeling radiologists' decisions. This observation makes intuitive sense. Even though certain cases are considered straightforward for the vast majority of radiologists, there are always differences of opinions on a per case basis, depending on how the individual radiologists apply the ACR guidelines in their daily practice.

The third and final experiment completed the exploration by drawing the final link from image content, perception, and cognition to human error. Although the paucity of diagnostic errors in our dataset does not allow us to make conclusive statements, the experimental results were quite promising, suggesting that contextual information collected from the image and the radiologist can be leveraged to predict diagnostic error. The small number of errors made by the study participants limited our investigation even further since we could not properly delineate the global and individual components. Still, our results supported the presence of the global component consistently for all readers. Therefore, leveraging a group-based understanding of radiologists' error-making patterns appears to be useful. However, given our findings with experiments 1 and 2, further fine-tuning using 'user-specific' data should be considered to develop more reliable predictive models of diagnostic error in radiology. It should be noted that the need for individual profiling of human error in mammography was discussed in an earlier study by Mello-Thoms *et al*.[60] Our observations certainly support such an approach. Overall, the promising results of the third experiment encourage a larger scale study for a deeper investigation of the above issues.

There are several limitations with our study. The biggest limitation is the relatively small number of cases and readers. Collecting carefully calibrated eye-tracking data in a clinical setting is a time consuming process. We recruited a mix of more and less experienced readers and we used an enriched dataset with very high prevalence of breast cancer (50%). Enriching artificially the clinical cases with more challenging abnormal cases is common for laboratory studies in radiology. Simply relying on consecutive screening cases would require an unrealistically high number of screening mammograms to be collected to reach an adequate number of cancer cases. A prior large-scale mammography study on the 'laboratory prevalence effect' topic concluded that such an effect is very small in magnitude and does not measurably alter the results of a study.[61] Another possible limitation of the study due to its retrospective, laboratory nature is that the diagnostic accuracy and inter-reader variability observed among the readers may not reflect their true clinical performance. Indeed, an earlier study suggested that experienced radiologists performed significantly better in a laboratory study involving diagnostic interpretation of mammograms than they did in the clinical environment.[62] Specifically, the radiologists demonstrated a very conservative approach when deciding to call a finding 'suspicious'. This may be true in our case as well, given that the three experienced radiologists in our study had an average detection rate of 90%, certainly higher than the performance of the average radiologist in practice. Finally, our user modeling relied on standard image texture features which may not be the best choice for optimal user modeling. However, they were a good starting point given our promising results. A follow-up study investigating a wide range of image features is certainly justified.

Regardless of the above limitations, our study produced some highly consistent trends in terms of applying effectively ML techniques for modeling effectively group and individual behaviors. Although our study was based on screen-film mammograms, there is no particular reason to suspect that our findings would not translate to full field digital mammograms. Reader studies have shown that the perceptual and cognitive behaviors of radiologists are similar in digital and screen-film mammograms.[63 64]

CONCLUSION

The results of our pilot study clearly demonstrated a link among image content, human perception, human cognition, and human error for mammographic breast cancer detection. Furthermore, our study delineated differences between group-based versus individualized understanding of the discovered links. An important finding is that monitoring clinicians' case-specific gaze metrics and cognitive behavior and merging those with computer-extracted information from the specific image could form the foundation for a system to predict diagnostic error in medical imaging. A second and equally important finding is that error-making patterns are not one-size-fits-all. Although studying collectively the broad community of clinicians involved in medical image interpretation can capture the common error-making aspects, personalized tuning appears to be beneficial in many cases. We believe that these findings encourage a paradigm shift in the way we think and develop computerized decision support systems and computerized education support systems for medical image interpretation.[65] A personalized approach is a promising way to improve existing

systems that are driven only by population-based understanding of the user community.

Acknowledgements This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05 00OR22725 with the US Department of Energy. The US Government retains and the publisher, by accepting the article for publication, acknowledges that the USA Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for USA Government purposes.

Contributors GT contributed to the conception, design, and interpretation of data, and drafted and revised the manuscript. She is guarantor. SV and VP contributed to the image analytics. EK contributed to the eye-tracking data collection and analysis. All authors contributed to critical revisions of the manuscript and approved the final version to be published.

Competing interests None.

Ethics approval Ethics approval was granted by the University of Arizona Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The study is based on a set of mammograms from the publicly available Digital Database of Screening Mammography (DDSM). Table 1 contains the complete list of cases used in the study. The collected gaze metrics will be available upon request.

REFERENCES

1 Patel VL, Kaufman DR. Medical informatics and the science of cognition. *J Am Med Inform Assoc* 1998;5:493–502.
2 Patel VL, Arocha JF, Kaufman DR. A primer on aspects of cognition for medical informatics. *J Am Med Inform Assoc* 2001;8:324–43.
3 Krupinski EA. The importance of perception research in medical imaging. *Radiat Med* 2000;18:329–34.
4 Andriole KP. Addressing the coming radiology crisis—the Society for Computer Applications in Radiology Transforming the Radiological Interpretation Process (TRIP™) initiative. *J Digit Imaging* 2004;17:235–43.
5 Krupinski EA. Current perspectives in medical image perception. *Atten Percept Psychophys* 2010;72:1205–17.
6 Krupinski EA, Berbaum KS. The medical image perception society update on key issues for image perception research. *Radiology* 2009;253:230–3.
7 Kulikowski CA, Shortliffe EH, Currie LM, *et al*. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *J Am Med Inform Assoc* 2012;19:931–38.
8 Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM. *Med Phys* 2008;35:5799–820.
9 Garland LH. On the scientific evaluation of diagnostic procedures. *Radiology* 1949;52:309–28.
10 Garland LH. Studies on accuracy of diagnostic procedures. *Am J Roentgenol Radium Ther Nucl Med* 1959;82:25–38.
11 Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. *Br J Radiol* 1997;70:1085–98.
12 Berlin L. Accuracy of diagnostic procedures: has it improved over the past five decades? *Am J Roentg (AJR)* 2007;188:1173–8.
13 Reiner B, Krupinski EA. The insidious problem of fatigue in medical imaging practice. *J Dig Img* 2012;25:3–6.
14 Krupinski EA, Berbaum KS, Caldwell RT, *et al*. Long radiology workdays reduce detection and accommodation accuracy. *J Am Coll Radiol* 2010;7:698–704.
15 Meyer J. *Age: Census 2000 brief*. Washington, DC: U.S. Census Bureau, 2001: C2KBR/01–12.
16 Beam CA, Conant EF, Sickles EA. Factors affecting radiologist inconsistency in screening mammography. *Acad Radiol* 2002;9:531–40.
17 Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology* 1992;184:613–17.
18 Birdwell RL, Ikeda DM, O'Shaughnessy KF, *et al*. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology* 2001;219:192–202.
19 Yankaskas BC, Schell MJ, Bird RE, *et al*. Reassessment of breast cancers missed during routine screening mammography: a community-based study. *Am J Roentg (AJR)* 2001;177:535–41.
20 Atlas SW. Embracing subspecialization: the key to the survival of radiology. *Am J Roentg (AJR)* 2007;4:752–3.
21 Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology* 2002;224:861–9.
22 Jensen A, Vejborg I, Severinsen N, *et al*. Performance of clinical mammography: a nationwide study from Denmark. *Int J Cancer* 2006;119:183–91.
23 Krupinski EA. Influence of experience on scanning strategies in mammography. *Proc SPIE Med Imag* 1996;2712:95–101.
24 Mugglestone MD, Gale AG, Cowley HC. Defining the perceptual processes involved with mammographic diagnostic errors. *Proc SPIE Med Imag* 1996;2712:71–7.
25 Krupinski EA. Visual scanning patterns of radiologists searching mammograms. *Acad Radiol* 1996;3:137–44.
26 Nodine CF, Mello-Thoms C, Weinstein SP, *et al*. Blinded review of retrospectively visible unreported breast cancers: an eye-position analysis. *Radiology* 2001;221:122–9.
27 Mello-Thoms C, Hardesty L, Sumkin J, *et al*. Effects of lesion conspicuity on visual search in mammogram reading. *Acad Radiol* 2005;12:830–40.
28 Mello-Thoms C. How does the perception of a lesion influence visual search strategy in mammogram reading? *Acad Radiol* 2006;13:275–88.
29 Mello-Thoms C, Britton C, Abrams G, *et al*. Head-mounted versus remote eye tracking of radiologists searching for breast cancer A comparison. *Acad Radiol* 2006;13:203–9.
30 Samuel S, Kundel HL, Nodine CF, *et al*. Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. *Radiology* 1995;194:895–902.
31 Manning DJ, Ethell SC, Donovan T. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *BJR* 2004;77:231–5.
32 Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol* 1978;13:175–81.
33 Kundel HL. How to minimize perceptual error and maximize expertise in medical imaging. *Proc SPIE* 2007;6515:651508-1–11.
34 Nodine CF, Kundell HL, Lauver SC, *et al*. Nature of expertise in searching mammograms for breast masses. *Acad Radiol* 1996;3:1000–6.
35 Krupinski EA, Kundel HL, Nodine CF. Enhancing recognition of lesions in radiographic images using perceptual feedback. *Opt Engin* 1998;37:813–18.
36 Kundel HL, Nodine CF, Krupinski EA. Computer-displayed eye-position as a visual aid to pulmonary nodule interpretation. *Invest Radiol* 1990;25:890–6.
37 Carmody DP, Nodine CF, Kundel HL. Finding lung nodules with and without comparative visual scanning. *Percept Psychophys* 1981;29:594–8.
38 Nodine CF, Kundel HL. A visual dwell algorithm can aid search and recognition of missed lung nodules in chest radiographs. *Visual Search* 1990;2:399–406.
39 Litchfield D, Ball LJ, Donovan T, *et al*. Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *J Exp Psychol Appl* 2010;16:251–62.
40 Mello-Thoms C, Dunn SM, Nodine CF, *et al*. The perception of breast cancers—a spatial frequency analysis of what differentiates missed from reported cancers. *IEEE Trans Med Imaging* 2003;22:1297–306.
41 Mello-Thoms C, Dunn SM, Nodine CF, *et al*. An analysis of perceptual errors in reading mammograms using quasi-local spatial frequency spectra. *J Digit Imaging* 2001;14:117–23.
42 Pietrzyk MW, Brennan PC, Donovan T, *et al*. Classification of radiological errors in chest radiographs, using support vector machine on the spatial frequency features of false-negative and false-positive regions. *Proc SPIE* 2011;7966:79660A.
43 Heath M, Bowyer K, Kopans D, *et al*. Current status of the digital database for screening mammography. In: *Digital Mammography*. Kluwer Academic Publishers, 1998. http://marathon.csee.usf.edu/Mammography/Database.html5
44 American College of Radiology. *Breast Imaging Reporting and Data System (BIRADS)*. 4th edn. VA: American College of Radiology, 2003.
45 Nodine CF, Kundel L, Toto C, *et al*. Recording and analyzing eye-position data using a microcomputer workstation. *Behavior Res Methods Instrum Comput* 1992;24:475–85.
46 Kundel HL, LaFolette PS. Visual search patterns and experience with radiological images. *Radiology* 1972;103:523–8.
47 Mackworth NH. Stimulus density limits the useful field of view. In: Monty RA, Senders JW. eds. *Eye movements and psychological processes*. Hillsdale, NJ: Erlbaum, 1976:307–22.
48 Krupinski EA, Nishikawa R. Comparison of eye position versus computer identified microcalcification clusters on mammograms. *Med Phys* 1997;24:17–23.
49 Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern* 1973;Smc-3:610–21.
50 Haralick RM, Shapiro LG. *Computer and robot vision*. Addison-Wesley Publishing Co, 1992.
51 Rubner Y, Tomasi C. *Perceptual metrics for image database navigation*. Springer, 2001.
52 Hubel DH, Weisel TN. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J Neurophysiol* 1965;28:229–89.
53 Campbell FW, Robson JG. Application of Fourier analysis to the visibility of gratings. *J Physiol* 1968;197:551–66.
54 Manjunath BS, Ma WY. Texture features for browsing and retrieval of image data. *Pattern Anal Mach Intell* 1996;18:837–42.
55 Hall M, Frank E, Holmes G, *et al*. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 2009;vol. 11:10–18.
56 Tourassi GD. *Receiver operating characteristics analysis: basic concepts and practical application. Handbook of medical image perception and techniques*. Cambridge, UK: Cambridge University Press, 2010.

57 Eng J. *ROC analysis: web-based calculator for ROC curves*. Baltimore: Johns Hopkins University [updated 17 May 2006]. http://www.jrocfit.org
58 Mazurowski MA, Baker JA, Barnhart HA, *et al*. Individualized computer-aided education in mammography based on user modeling: concept and preliminary experiments. *Med Phys* 2010;37:1152–60.
59 Mazurowski MA, Barnhart HA, Baker JA, *et al*. Identifying error-making patterns in assessment of mammographic BI-RADS descriptors among radiology residents using statistical pattern recognition. *Acad Radiol* 2012;19:865–71.
60 Mello-Thoms C, Dunn S, Nodine CF, *et al*. The perception of breast cancer: what differentiates missed from reported cancers in mammography? *Acad Radiol* 2002;9:1004–12.
61 Gur D, Rockette HE, Armfield DR, *et al*. Prevalence effect in a laboratory environment. *Radiology* 2003;228:10–14.
62 Gur D, Bandos AJ, Cohen CS, *et al*. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography Interpretations. *Radiology* 2008;249:47–53.
63 Krupinski EA. *Softcopy reading. Digital mammography*. Springer, 2010;107–119.
64 Skaane P, Diekmann F, Balleyguier C, *et al*. Observer variability in screen-film mammography versus full-field digital mammography with soft-copy reading. *Eur Radiol* 2008;18:1134–43.
65 Tourassi GD, Mazurowski MA, Harrawood BP, *et al*. Exploring the potential of context-sensitive CADe in screening mammography. *Med Phys* 2010;37:5728–36.