Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Modeling visual search behavior of breast radiologists using a deep convolution neural network

Suneeta Mall
Patrick C. Brennan
Claudia Mello-Thoms

SPIE.

Suneeta Mall, Patrick C. Brennan, Claudia Mello-Thoms, "Modeling visual search behavior of breast radiologists using a deep convolution neural network," *J. Med. Imag.* **5**(3), 035502 (2018), doi: 10.1117/1.JMI.5.3.035502.

Modeling visual search behavior of breast radiologists using a deep convolution neural network

Suneeta Mall,* Patrick C. Brennan, and Claudia Mello-Thoms
University of Sydney, Faculty of Health Sciences, Medical Image Optimisation and Perception Research Group (MIOPeG), Lidcombe, New South Wales, Australia

Abstract. Visual search, the process of detecting and identifying objects using eye movements (saccades) and foveal vision, has been studied for identification of root causes of errors in the interpretation of mammograms. The aim of this study is to model visual search behavior of radiologists and their interpretation of mammograms using deep machine learning approaches. Our model is based on a deep convolutional neural network, a biologically inspired multilayer perceptron that simulates the visual cortex and is reinforced with transfer learning techniques. Eye-tracking data were obtained from eight radiologists (of varying experience levels in reading mammograms) reviewing 120 two-view digital mammography cases (59 cancers), and it has been used to train the model, which was pretrained with the ImageNet dataset for transfer learning. Areas of the mammogram that received direct (foveally fixated), indirect (peripherally fixated), or no (never fixated) visual attention were extracted from radiologists' visual search maps (obtained by a head mounted eye-tracking device). These areas along with the radiologists' assessment (including confidence in the assessment) of the presence of suspected malignancy were used to model: (1) radiologists' decision, (2) radiologists' confidence in such decisions, and (3) the attentional level (i.e., foveal, peripheral, or none) in an area of the mammogram. Our results indicate high accuracy and low misclassification in modeling such behaviors. [DOI: 10.1117/1.JMI.5.3.035502]

Keywords: visual search; breast cancer; deep learning; behavior modeling; mammography; eye tracking; machine learning.

Paper 18082RR received Apr. 14, 2018; accepted for publication Jul. 24, 2018; published online Aug. 11, 2018.

1 Introduction

The human eyes are more attracted toward areas of higher information content than areas of higher redundancy.[1] The capability to process information using human eye is highest at the fovea centralis (fovea) and decreases rapidly as one moves toward parafovea (periphery).[2] This decrease in the capability of processing information (e.g., detail, shape, and color) is known to be the same as a probability density function[2] with highest capability concentrated around 2.5-deg radial angle (foveal vision) of the center of gaze. Saccades, high-speed rapid eye movements, together with foveal (direct, overt, and detailed) and peripheral (indirect, covert, and less detailed) visual attention, allow an efficient search of specific targets to be carried out. Some of these aspects of visual search behavior, specifically foveal vision, have been studied for errors in interpretation of mammograms.[3–6] However, factors that influence the level of attentional deployment (foveal/peripheral/none) largely remain unexplored. Better understanding of the aspects of attentional deployment is vital in building strategies to reduce the number of cancers that are missed, thereby increasing the accuracy of mammographic interpretation.

Accuracy of mammographic interpretation is not ideal, as it has been shown that about 7 to 12%[7] lesions are falsely reported (false positives) and that 4% to 34%[8] cancers are missed (false negatives). Using visual search behavior analysis, it has been shown that 70% of cancers that are not reported in fact attracted radiologists' attention.[9–11] Cancers that are correctly reported (true positives) are known to differ in their energy profile from cancers that attract foveal attention but are not reported (false negatives).[3–5] Evidently, features of an area containing the lesion affect, to some extent, the radiologist's decision outcome and confidence in that decision.[6]

These features are critical and could potentially be used to improve the accuracy of radiologists' decisions and the patients' health care experience. This can be achieved by improving training programs to better understand the radiologists' search behaviors and understand what lesions are going to be missed and where an erroneous decision, such as a false positive (FP), is likely to be made. Building a system that models visual search behavior of radiologists is essential in achieving this. Computer-aided detection (CADe) has been used to address some of these aspects, such as providing information about where cancer is likely to be present,[12] however except for one study,[12] these algorithms, to our knowledge, have not utilized visual search behaviors. Due to high false-positives, successful adaptation of CADe in clinical settings has been limited.[13]

In this study, we model radiologists' visual search behavior to understand whether deep machine learning techniques can assist in improving radiologists' diagnostic assessment. A few studies[14–16] have previously used machine learning techniques to model or to predict radiologists' decision outcomes; however, these models were based on neural-nets or support vector machines (SVM). One of the limitations with these models is that they are trained with handpicked features. Deep machine learning learns the features by itself[17] as it trains the model through the input data—suggesting that training is not biased by the preselected features that the model was provided with.

*Address all correspondence to: Suneeta Mall, E-mail: smal5514@uni.sydney.edu.au

2329-4302/2018/$25.00

The accuracy of deep learning algorithms is also known to be better than other contemporary machine learning techniques.[18] Deep convolutional neural networks (hereby referred as ConvNet) are a specialized class of deep learning algorithms that are biologically inspired, a multilayer perceptron simulating the visual cortex. With a ConvNet model (with transfer learning[18,19]), as used in this study, we aim to ascertain (with reasonably high probability) the radiologists' decisions (and confidence in such decisions). We also aim to determine which areas of mammograms are likely to receive visual attention or to be disregarded.

2 Materials and Methods

Eight mammography quality standards act-certified radiologists participated in this fully crossed multireader multicase visual search study of digital mammography involving 120 two-view [craniocaudal (CC) and mediolateral oblique (MLO)] cases (59 cancers). The cases were obtained from a routine screening program using a Selenia full-field digital mammography system (Hologic Inc., Marlborough, Massachusetts).

Ground truth was established by a separate Mammography Quality Standards Act-certified breast radiologist, who did not participate as an observer in this study, using pathology reports and additional imaging. All cancer cases were biopsied, and all normal cases had a follow up of 2 years.

2.1 *Study Protocol*

The radiologists were seated 60 cm from a workstation that contained two calibrated medical-grade 5 megapixel flat-panel portrait-mode displays (model C5i, Planar Systems Inc., Beaverton, Oregon), with a resolution of 2048×2560 pixels, typical brightness of 146 ftL, and 3061 unique shades of gray. The radiologists wore a head-mounted eye-position tracking (ET) system (ASL Model H6, Applied Sciences Laboratory, Bedford, Massachusetts) that used an infrared beam (at temporal resolution of 60 Hz) to calculate line of gaze by monitoring the pupil and the first corneal reflection. A magnetic head tracker was used to monitor head position, and this allowed the radiologists to freely move their heads from side to side as well as toward the displays, up to 20 cm, at which point they were outside the range of the head tracker. The ET integrates eye position and head position to calculate the intersection of the line of gaze and the display plane. The system has an accuracy (measured as the difference between true eye position and computed eye position) of <1 deg of visual angle, and it covers a visual range of 50 deg horizontally and 40 deg vertically.

Prior to the beginning of each reading session, a calibration of ET was performed wherein a 3×3 grid was shown on both the displays. After every five cases, the ET system was rechecked and if necessary, it was recalibrated, but this was only required twice at most during each reading session.

After the calibration, the first (or next) case appeared on the displays wherein the left- and right-hand side monitors would, respectively, display CC and MLO views of the case. The eye tracker captured the X- and Y-co-ordinates of fixation location on the ASL plane, dwell time, view, radiologists' distance to the monitor, and other details. Radiologists were advised to mark the location of malignant lesions on the screen using a mouse-controlled cursor, along with providing a confidence score on the likelihood of malignancy at the location. The software had the capability to capture both these pieces of information on screen with the help of pop-up dialog boxes. Upon termination of search for a given case, the radiologists used a mouse-controlled cursor to click on a button in the display to select the next case of their reading sequence and were not allowed to come back to previously assessed cases.

Visual search maps of the radiologists were obtained as they assessed the cases and identified potential malignancies. Radiologists were asked to provide their decisions (i.e., locations of suspected malignancy) alongside a five-point scale confidence score on likelihood of malignancy in the location of these decisions (five being most confident 81% to 100%, one being least confident 1% to 20%).

From these mammographic images, three types of areas, namely foveal clusters (FCs), peripheral clusters (PC), and never fixated clusters (NFC), were extracted.

1. FCs: FCs are defined as the breast areas measuring 2.5-deg radial angle (about 160 pixels × 160 pixels square area) consisting of at least three temporally sequential fixations (Fig. 1). FC cluster extraction algorithm involves performing: (1) A fixed radius nearest neighbor algorithm using K-dimensional (KD)-tree and bounded deformation (BD)-tree[20] to obtain all clusters containing fixation points that are within 2.5-deg radial angle to each other, followed by (2) removal of the redundant clusters, and (3) selection of clusters that contained at least three temporally sequential fixation points.

2. PCs: PCs are defined as the breast areas within 2.5-deg radial angle, from the location of a lesion where a decision was made by radiologists, consisting of <3 temporally sequential fixations (Fig. 2). To extract PC clusters: a square of 160 pixels around the location where radiologists made a decision but had <3 temporally sequential fixation points was automatically extracted from the image. These clusters were retrospectively checked to ensure that they contained at least one fixation point.

3. NFCs: NFCs are defined as the breast areas that did not receive any fixation by any of the eight radiologists (Fig. 3). NFC were extracted by: (1) overlaying all eight radiologists' visual search maps on the cases, (2) identifying 2.5-deg radial angle areas per view per case that did not receive any fixation by any of the radiologists and extracting center co-ordinate for these areas, and (3) automatically extracting these areas from the image. Only one such cluster per view per case was obtained bringing the total to about 240 NFCs.

A more in-depth description of the extraction algorithm to obtain these clusters can be found elsewhere.[6] FC, PC, and NFC were then classified into four categories of decision outcome [true positive (TP), FP, true negative (TN), and false negative (FN)] based on the accuracy of decision. NFC (all TN) clusters, however, were not used in modeling either decision outcome or confidence in such decision, but they were used in modeling attentional level.

These labeled datasets were used to model radiologists' visual search behavior and decisions. The following three models of search behavior of radiologists and their decisions were trained separately using a deep ConvNet, specifically "Inception-ResNet (V2)"[21] with workflow shown in detail in Fig. 4

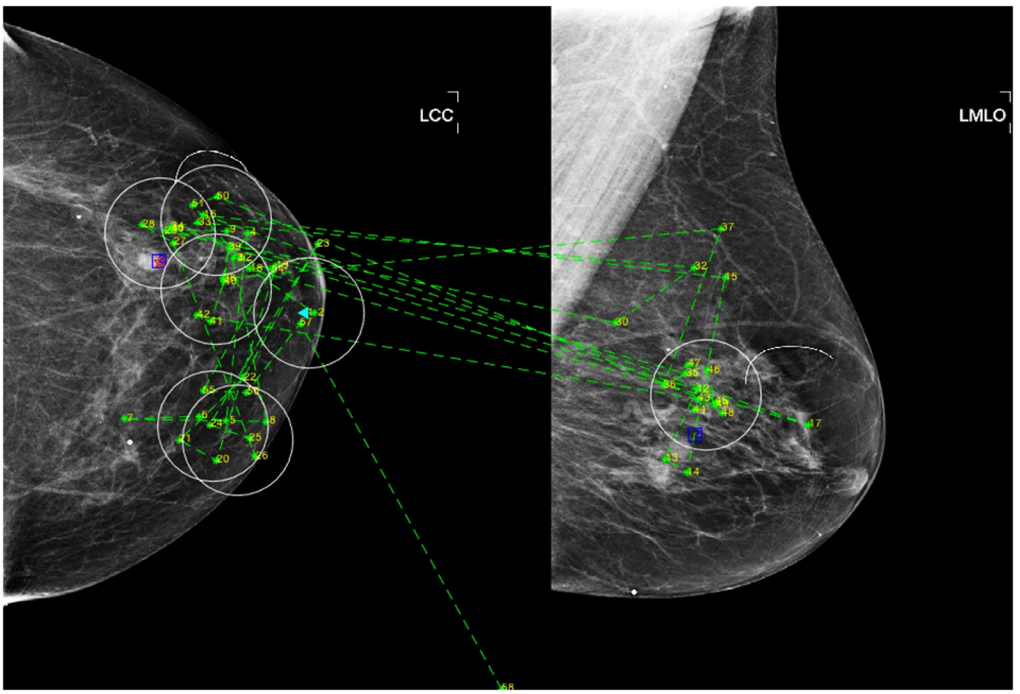

Fig. 1 FCs are the breast areas measuring 2.5-deg radial angle consisting of at least three temporally sequential fixations. These are highlighted with white circles. Red star indicates true malignancy and blue square marking indicates location where a radiologist reported a malignant finding. Green points and dotted lines represent the temporal visual search behavior (fixation points and the temporal sequencing amid these points). The FC containing blue star in this figure on left view has been classified as TP as true cancer lies within the FC area, whereas the FC containing blue star in right view has been classified as FP because no true malignancy was present within 2.5-deg radial angle area.

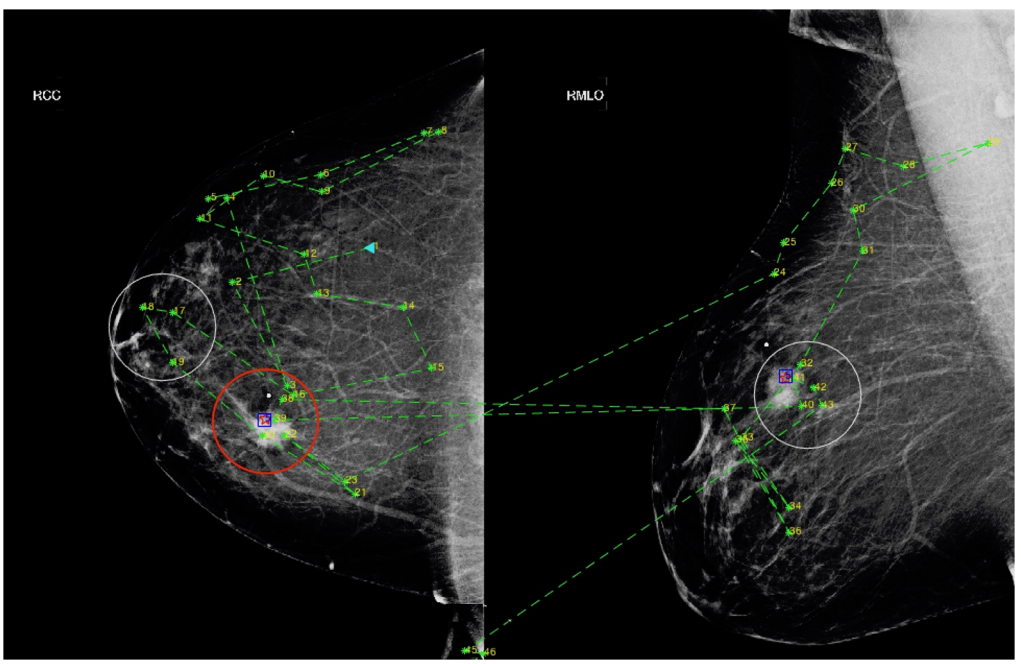

Fig. 2 PCs are the breast areas within 2.5-deg radial angle, from a location where a decision was made by radiologists, consisting of <3 temporally sequential fixations. In this figure, the area shown in red circle is an example of PC. PC, in this example, is TP. For details of the figure annotations, please refer to Fig. 1 legend.

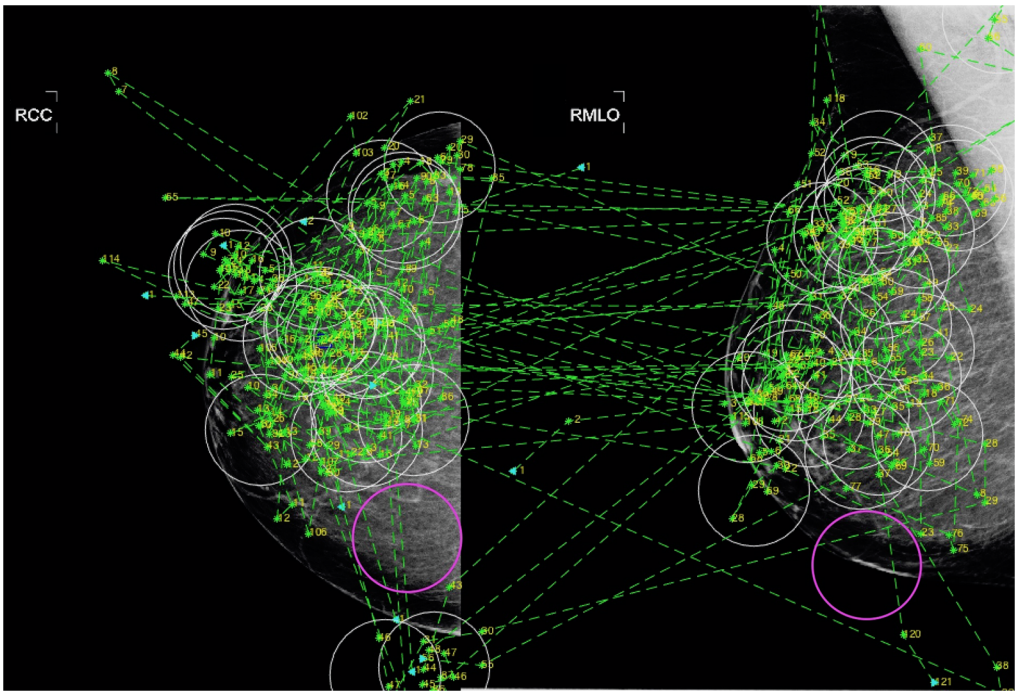

Fig. 3 NFCs are the breast areas that did not receive any fixation by any of the eight radiologists. This figure overlays visual search behavior of all radiologists for the case indicating areas that did not receive any attention by any of the radiologists. Example of NFC area is shown in pink circle. For details of the figure annotations, please refer to Fig. 1 legend.

(1) Decision outcome,

(2) confidence in the decision, and

(3) attentional level (i.e., foveal, peripheral, or none) obtained by an area of the mammogram.

2.2 *Data Processing and Analysis*

2.2.1 *Preparing the dataset*

The distribution of clusters per category, in our case, was non-uniform (as shown in Table 1). Using the entire dataset would

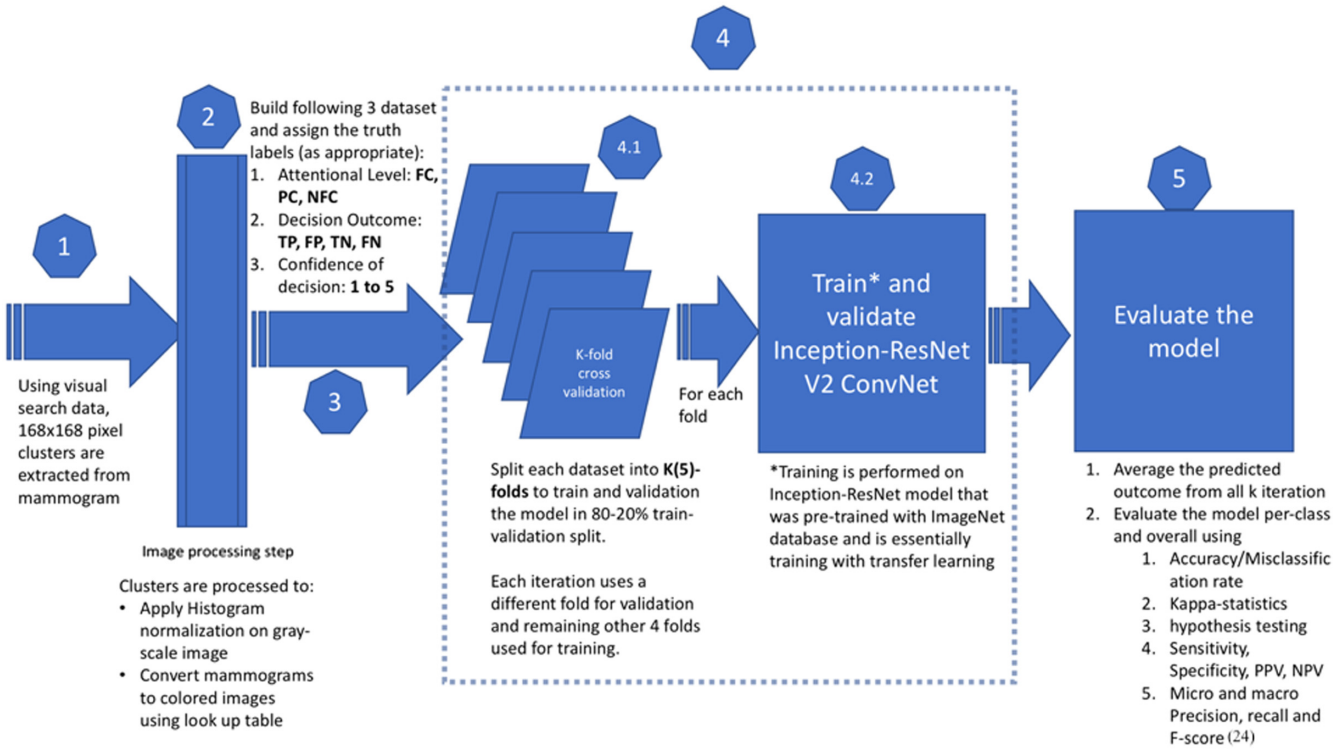

Fig. 4 Details of the workflow used to model radiologists' visual search behavior and their decisions.

Table 1 Details of the dataset used in modeling radiologists' visual search behavior.

			Randomized dataset used in modeling	
Dataset	Category	Total data available in the category	Total count	Training and validation split (training and validation)
Attentional level	FC	10458	1200	(960, 240)
	PC	1196	1196	(957, 239)
	NFC	240	240	(192, 48)
Decision	TN	9866	1500	(1200, 300)
	TP	224	224	(179, 45)
	FN	147	147	(118, 29)
	FP	1417	1417	(1134, 283)
Confidence	1	660	330	(264, 66)
	2	362	362	(290, 72)
	3	210	210	(168, 42)
	4	209	209	(167, 42)
	5	224	224	(179, 45)

lead to a very high null accuracy (a.k.a. "no information rate," that is, the accuracy when the model has no training (dumb model) and always predicts one category that has highest amount of data in the dataset). For example, for attentional level (ref. Table 1), the FC category includes 10,458 clusters while NFC only has only 240, leading to null accuracy of 88% $[= 10458/(10458 + 240 + 1196)]$, thus increasing the risk of bias or dumb model. For this reason, using a random sampling approach, the distribution of data for each model was normalized (i.e., brought to be approximately of the same order as other categories) and only a subset of all available data was used in training. These are detailed in Table 1.

2.2.2 *Preprocessing*

Adhering to the useful field of view (2.5 deg) radial angle, as described in Ref. 6, all clusters (obtained from processing visual search data of radiologists) were 160×160 pixel images. These gray-scale images that represent areas of the breast were then processed to be converted to colored images using the lookup-table approach. Prior to color conversion, histogram normalization was applied to avoid any loss of information. The results of normalization and color conversion are shown in Fig. 5. This step was necessary because ConvNets are designed to work with natural images that have three channels.[22]

2.2.3 *Modeling visual search behavior of radiologists and their decisions*

The models were trained and validated using the following approach.

K-fold cross validation. We used a fivefold (k-fold wherein $K = 5$) cross validation approach to match the 80% to 20% split of the training and validation samples. The final results were calculated based on the average prediction matrix of these fivefold training/validation outcomes.

Modeling visual search behavior of radiologists and their decisions.. Deep ConvNet architectures, as used in this study, are layered ConvNets of different configurations and filter sizes (7×1, 1×7, 1×3, 3×1, 3×3, and 1×1). The "Inception-ResNet-v2"[21] (the deep ConvNet architectures used in this study) combines three residual networks (ResNet) containing 1 Inception v4 network (Fig. 6). This network has shown 3.08% top-5 error in the ImageNet dataset[21] and has been shown to outperform Inception-v4 (albeit by a thin margin) and thereby all its predecessors. The hyperparameters used were

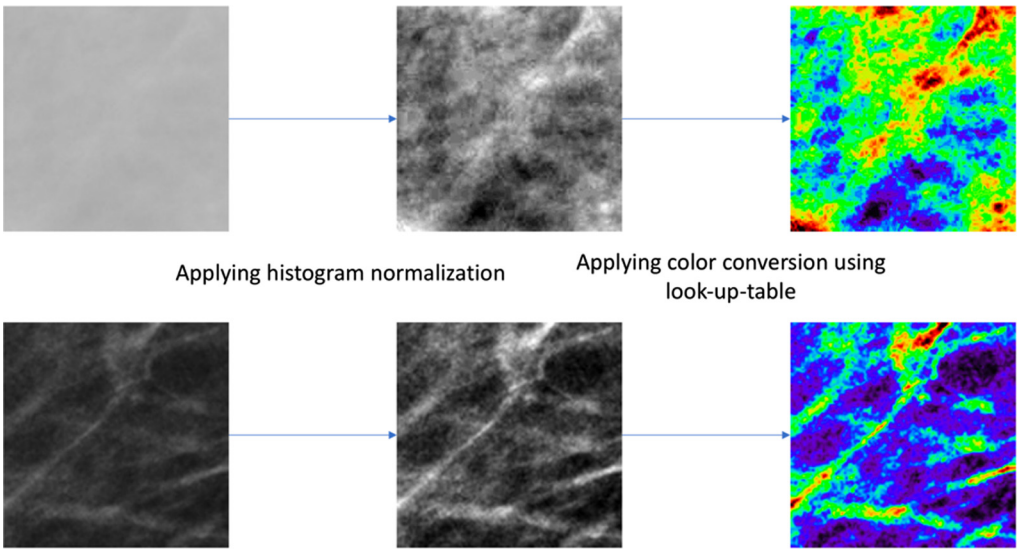

Fig. 5 Stepwise results of preprocessing on clusters aimed to convert grayscale cluster images to colored images using the lookup table approach.

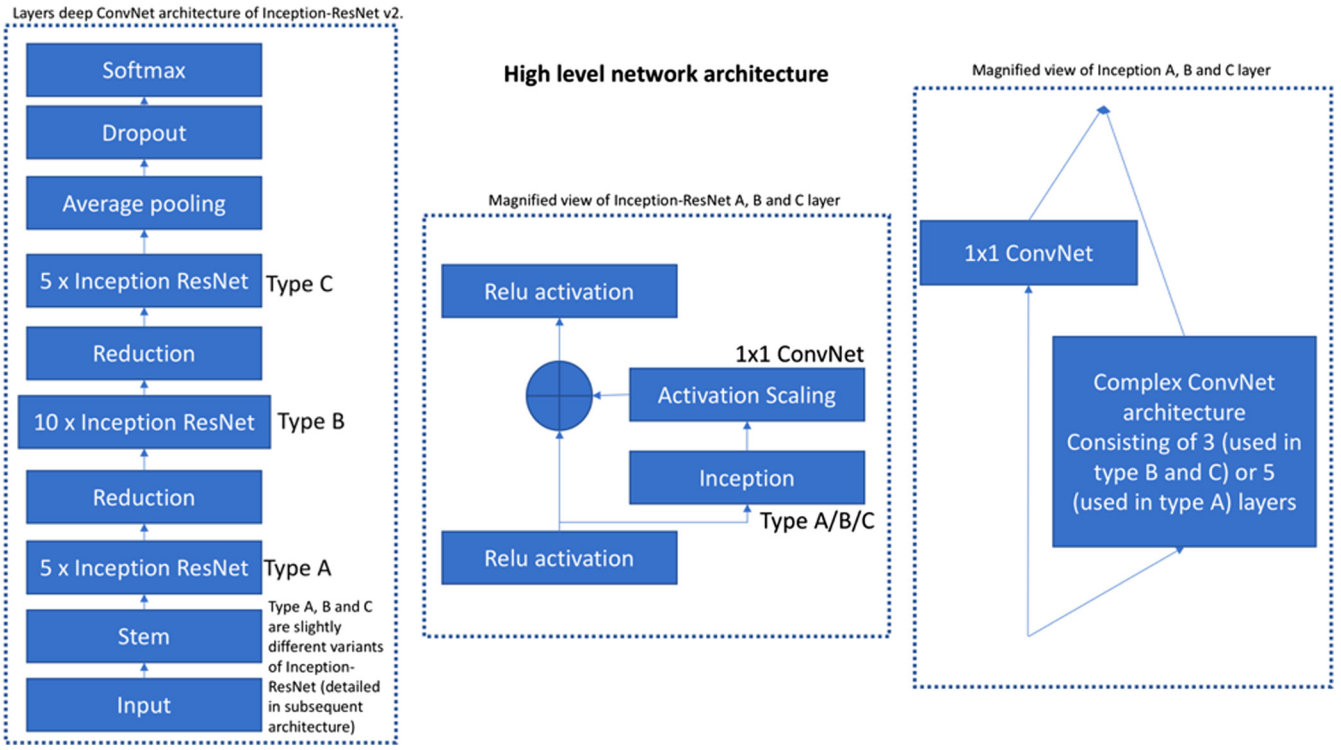

Fig. 6 High-level architecture of deep convolution network used in modeling the behavior of radiologists.

(1) Learning rate: exponential decay function, initial rate 0.01, end rate 0.0001;

(2) Optimization function: RMSProp [(Root mean square), an adaptive learning rate method proposed by Geoff Hinton[23]] with moment 0.9, epsilon 1.0, decay 0.9;

(3) Follow the regularized leader: accumulator value 0.1, L1, and L2 strength 0.

To avoid overfitting and improve model performance, image augmentation was also applied. Distortion in color (by changing the hue, contrast, and saturation) and random rotation of images was used for augmentation. Random cropping was not used to avoid any loss of information [the objective of the study was to retain the cluster with information of the useful field of view, which is essentially the size of cluster images (160×160 pixels)].

A very large amount of data is required to train a deep learning model; however, in our case, the dataset is relatively small. For this reason, we have reinforced our analysis using transfer learning techniques. In transfer learning, a model that was previously trained to perform a specific task, T1, is retrained (a.k.a. fine-tuned) to perform task T2. It has been shown that models are able to use the knowledge they have gained to perform T1 into performing task T2.[18,19] The ImageNet dataset was used to train this model, which was then trained further to learn specific task of predicting the (1) attentional level, (2) decision outcome, and (3) confidence in the decision made on a given location.

During transfer learning/fine-tuning, the softmax layer (Fig. 6) of the pretrained model was dropped and replaced with new layer due to mismatch in the number of outputs (categories) of the pretrained (1000 categories, as trained with ImageNet) and the desired (fine-tuned) model [i.e., models for decision (four categories), confidence in decision (five categories), and attention level (three categories)]. This is a standard practice when using transfer learning.

Most of the analysis was performed using Tensorflow and R-language. Graphical processing unit (GPU) NVIDIA GRID K520 was utilized to accelerate the training and validation durations.

2.2.4 *Evaluation of the model*

Model performance. The confusion matrix obtained from the model formed the basis of the evaluation. The averaged (of k-fold predictions) confusion matrix was analyzed to obtain

- **Per-category evaluation**: We analyzed sensitivity, specificity, positive (PPV) and negative (NPV) predictive values, and accuracy [i.e., (1 − Misclassification Rate)] of the model in predicting a specific category. These measures indicate how well the model understands and categorizes breast areas for a given specific category.
- **Overall evaluation**: To evaluate the overall performance of the model, we analyzed accuracy [i.e., (1 − Misclassification Rate)] and confidence interval (95% CI) of accuracy. We also compared the model against null accuracy using hypothesis testing with H1: accuracy of current model is better than "dumb model." To look at the agreement between truth and predicted class, we performed Cohen's Kappa analysis. Lastly, we also analyzed microprecisions and macroprecisions,[24] recall[24] and F-score[24] for our multiclass classifiers models.

Bias and variance analysis.. Bias and variance of decision outcome, confidence in the decision and attentional level models

Table 2 Results from modeling radiologists' decision outcome.

	Average outcome of $K(5)$-fold cross validation	True category of decision outcome on breast area				Measures based on average of $k(5)$-fold evaluation	Overall model measure
		TP decision	TN decision	FP decision	FN decision		
Predicted category of decision outcome on breast area	TP decision	30	3	6	0	Sensitivity: 0.77 Specificity: 0.98 PPV: 0.67 NPV: 0.99 Accuracy: 0.67 Misclassification rate: 0.33	Accuracy: 0.92 95% CI: (0.8924, 0.9363) Misclassification: 0.08 Null-accuracy: 0.46 P-value: $<2\times10^{-16}$ Kappa: 0.86 Precision_{μ}: 0.92 Recall_{μ}: 0.92 F1-score_{μ}: 0.92 Precision_{M}: 0.79 Recall_{M}: 0.83 F1-score_{M}: 0.81
	TN Decision	3	285	2	7	Sensitivity: 0.96 Specificity: 0.96 PPV: 0.95 NPV: 0.97 Accuracy: 0.95 Misclassification rate: 0.05	
	FP Decision	11	10	270	5	Sensitivity: 0.91 Specificity: 0.96 PPV: 0.95 NPV: 0.93 Accuracy: 0.95 Misclassification rate: 0.05	
	FN Decision	1	2	5	17	Sensitivity: 0.68 Specificity: 0.98 PPV: 0.59 NPV: 0.99 Accuracy: 0.59 Misclassification rate: 0.41	

were also calculated using the misclassification rates, i.e., the error estimates of their respective k-iteration (of k-fold) training. The bias is defined as

$$\text{Bias} = \frac{\sum_{i=1}^{i=k} (\text{misclassification rate})_i}{k}.$$

The variance in the error estimates is defined as

$$\text{Variance} = \frac{\sum_{i=1}^{i=k} [(\text{misclassification rate})_i - \text{Bias}]^2}{k}.$$

3 Results

Our results from modeling the radiologists' visual search behavior and decisions are as follows.

3.1 *Decision Outcome*

We noted 92% accuracy in modeling radiologists' decisions using their visual search behavior. This model was found to be statistically significantly better (p-value $\cong 0$) than the dumb model (Table 2). We have also noted a very high agreement ($k = 0.86$) between the true decision outcome and the predicted decision outcome.

Standard deviation and variance in sensitivity (standard deviation $= 0.13$ and variance $= 0.01$) and specificity (standard deviation $= 0.02$ and variance < 0.001) for all decision outcome categories were low. The least sensitivity was obtained for the false negative category of decision prediction; perhaps not coincidently, this category had the smallest dataset.

3.2 *Confidence in the Decision*

We noted 66% accuracy in modeling radiologists' confidence in their decisions using their visual search behavior. This model was found to be statistically significantly (p-value $\cong 0$) better than the dumb model (Table 3). We have also noted moderate agreement ($k = 0.56$) between the true confidence level on the decisions and the predicted confidence level on the decisions. Standard deviation and variance in sensitivity (standard deviation $= 0.10$ and variance $= 0.01$) and specificity (standard deviation $= 0.03$, variance < 0.001) for all confidence level categories were low.

3.3 *Attentional Level*

We noted 90% accuracy in modeling deployment of radiologists' attentional level using their visual search behavior. This model was found to be statistically significantly (p-value $\cong 0$) better than the dumb model (Table 4). We have also noted very high agreement ($k = 0.82$) between the true attentional level (deployed on a cluster) and the predicted attentional level (Figs. 7 and 8).

Standard deviation and variance in sensitivity (standard deviation $= 0.02$, variance < 0.001) and specificity (standard deviation $= 0.03$, variance < 0.001) for all attentional-level categories were low. The lowest sensitivity was obtained for NFC category of attentional-level prediction;

Table 3 Results from modeling radiologists' confidence in their decision.

	Average of $K(5)$ fold	True confidence in the radiologist's decision on breast area					Measures based on average of $k(5)$-fold evaluation	Overall model measure
		1	2	3	4	5		
Predicted confidence in the radiologist's decision on breast area	1	44	12	7	6	6	Sensitivity: 0.59 Specificity: 0.89 PPV: 0.67 NPV: 0.85 Accuracy: 0.67 Misclassification rate: 0.33	Accuracy: 0.66 95% CI: (0.5989, 0.7159) Misclassification: 0.34 Null-accuracy: 0.27 P-value: $<2 \times 10^{-16}$ Kappa: 0.56 Precision_{μ}: 0.66 Recall_{μ}: 0.66 F1-score_{μ}: 0.66 Precision_{M}: 0.65 Recall_{M}: 0.69 F1-score_{M}: 0.67
	2	15	51	9	5	4	Sensitivity: 0.61 Specificity: 0.89 PPV: 0.71 NPV: 0.83 Accuracy: 0.71 Misclassification rate: 0.29	
	3	3	4	24	3	3	Sensitivity: 0.65 Specificity: 0.92 PPV: 0.58 NPV: 0.94 Accuracy: 0.57 Misclassification rate: 0.43	
	4	2	3	1	27	2	Sensitivity: 0.77 Specificity: 0.94 PPV: 0.64 NPV: 0.96 Accuracy: 0.64 Misclassification rate: 0.36	
	5	2	2	1	1	30	Sensitivity: 0.83 Specificity: 0.94 PPV: 0.67 NPV: 0.97 Accuracy: 0.67 Misclassification rate: 0.33	

Table 4 Results from modeling radiologists' attentional level.

	Average of $K(5)$ fold cross validation	True level of attention deployed on breast area			Measures based on average of $k(5)$-fold evaluation	Overall model measure
		FC	PFC	NFC		
Predicted level of attention deployed on breast area	FC	221	13	8	Sensitivity: 0.91 Specificity: 0.93 PPV: 0.92 NPV: 0.93 Accuracy: 0.92 Misclassification rate: 0.08	Accuracy: 0.90 95% CI: (0.8705, 0.9238) Misclassification: 0.10 Null-accuracy: 0.46 P-Value: $<2 \times 10^{-16}$ Kappa: 0.82 Precision_{μ}: 0.90 Recall_{μ}: 0.90 F1-score_{μ}: 0.90 Precision_{M}: 0.82 Recall_{M}: 0.89 F1-score_{M}: 0.86
	PFC	17	224	11	Sensitivity: 0.89 Specificity: 0.95 PPV: 0.94 NPV: 0.90 Accuracy: 0.94 Misclassification rate: 0.06	
	NFC	2	2	29	Sensitivity: 0.88 Specificity: 0.96 PPV: 0.60 NPV: 0.99 Accuracy: 0.60 Misclassification rate: 0.40	

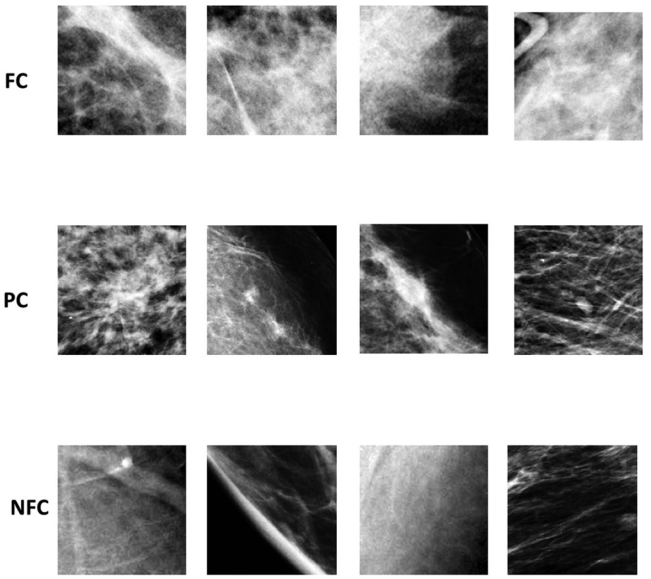

Fig. 7 Example of breast regions that were classified as FC, PC, and NFC. These are the true FC, PC, and NFC regions.

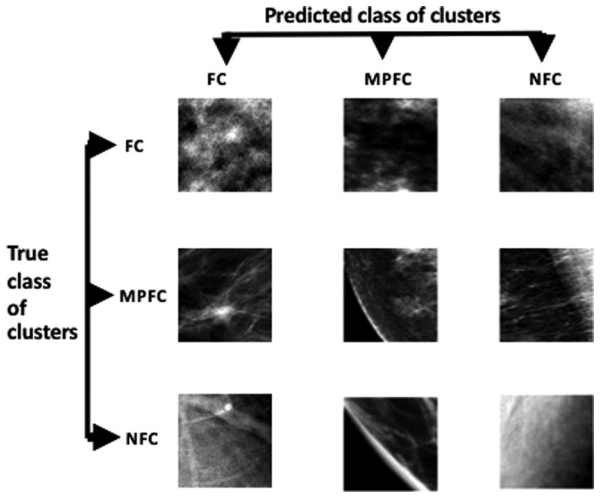

Fig. 8 Example of breast regions FC, PC, and NFC in truth and the predicted region types (as determined by the classifier) shown in a confusion matrix format.

perhaps not coincidently, this category had the smallest dataset (240 clusters only).

3.4 *Bias and Variance of these Models*

Bias and variance are the two tradeoffs of a model. Bias can be represented as the misclassification rate, i.e., the error estimates of the model, and it is minimized by choosing a sufficiently large training set. The variance in the error estimates of the model indicates the ability to generalize (based on test dataset). In other words, high variance in the error estimates implies that the performance of the model is highly dependent on the training data and the model does not generalize well (i.e., does not find too many outliers that effect the model performance).

The benefit of using k-fold cross-validation technique is that all the data have been used for both training and validation purposes in k-iteration of modeling the radiologists' behavior. Although there is an overlap in the training set in each of the k iterations (wherever $k > 2$), the validation set remains unique per iteration. Our results of bias and variance analysis based on all k-fold iterations of training of all the three models are shown in Table 5. We observed standard deviations of 0.015, 0.06, and 0.015 in error estimates of fivefold cross validation for decision outcome, confidence in decision and attentional-level modeling, respectively. The variance in error estimate of all our models was observed to be <0.004.

4 Discussion

Detection and identification of malignancy in radiographic imaging is a learnt skill that radiologists acquire over the course of time. In analogy, machine learning techniques learn to perform a specific task based on the information (learning dataset) it is provided with. Machine learning techniques, specifically deep convolution neural networks, in the last 10 years have evolved to be really efficient in detecting and identifying everyday objects. From initial 17% top-5 error rate[25] when the ImageNet large scale visual recognition challenge first reported the use of ConvNet (AlexNet[25]) to now at about <4% (e.g., Inception-ResNet[21]), the reduction in the error rate is promising. The layered architecture of deep ConvNet is a multitiered multilayer perceptron that simulates how information is processed in the human visual cortex. Use of various hidden layers in such a network has also previously been compared with how radiologists process information.[26] In both cases, recollection of all the steps and the weight of the factors that contributed to the final decision are not explicit and some factors are always hidden/endogenous.[26]

The benefits of understanding radiologists' visual search behavior and being able to predict some aspects of search,

Table 5 Result of bias and variance analysis of misclassification rates (probability of error) for each iterator of $k(5)$-fold cross validation.

	Misclassification error of each iteration of $K(5)$-fold cross validation					Average misclassification rate of k-fold validation	Bias	Standard deviation	Variance
	1	2	3	4	5				
Decision outcome	0.10	0.07	0.10	0.07	0.08	0.08	0.084	0.015	0.0002
Confidence in decision	0.25	0.32	0.38	0.40	0.35	0.34	0.339	0.060	0.0036
Attentional level	0.08	0.09	0.10	0.11	0.12	0.1	0.101	0.015	0.0002

such as the selection of the regions to which the foveal vision is deployed and the characteristics of regions that influence radiologists' decisions, are manifold. For example, it can be used to predict which lesions are likely to be missed during search and where an erroneous decision is likely to be made. This information can thus be used in providing more efficient training programs and second opinions during mammography interpretation —leading to increased accuracy of interpretation and improved health care experience. Artificial neural networks (ANN) have previously been used to predict the decision outcome on foveally fixated (FC) regions using energy profile characteristics of the regions.[3,14] This ANN model, built using feature engineering (handcrafted features), had about 67% accuracy in predicting TP decisions. Error in predicting all decision outcome categories varied from about 2% to 33%. In our study, we have shown that deep ConvNet (Inception-ResNet V2) self-learnt feature network can be trained to predict decision outcome based on visual search behavior and that high accuracy and high agreement ($k = 0.86$) in such predictions can be achieved.

Kundel and Nodine's focal/global model[12] describes a multistage process wherein radiologists build a holistic view of the image (in <2 s[27]), identify perturbations, gather information through foveal vision, and make a decision and terminate search reporting suspected cancer or absence of abnormalities. Radiologists' holistic view of the mammographic image is based on information gathered using peripheral vision only as fixation/foveal attention is deployed afterward. Peripheral attention is very much a covert operation, continually occurring and assisting in foveal deployments, thereby enabling efficient extraction of information.[28] Also, peripheral vision despite being less detailed, at the expense of increased latency,[29,30] can assist in identification.[2,29] The role peripheral vision takes in identification of suspected lesions in mammography is largely unexplored; however, it has been shown that, in mammography, areas that receive direct (foveal), indirect (peripheral), or no attention at all are different from each other.[6] In this study, we show that a ConvNet model does learn about characteristics that play a critical role in attention deployment and the level of attention a location is likely to receive (or not receive at all) can be predicted with high accuracy. This information can be used in the identification of malignancies that may be missed (FN) and it can also be used to improve CAD algorithms that sample the whole image in their search strategies.

Confidence level in radiologists' (binary) decision (cancer or non-cancer) is a probability score (of five levels) and is provisional—it is not an everyday practice that radiologists observe in the clinic. It is a laboratory measure that is asked so that the area under the trapezoidal receiver operating characteristic curve can be plotted. It has been shown that radiologists' binary decisions do not necessarily agree with the confidence levels reported.[31] In this study, only moderate agreement between the true and predicted confidence in radiologist's decision could be achieved. We theorize that perhaps it is more the endogenous factors of radiologists that influence the confidence in their decisions, thereby making it harder to model using visual search behaviors.

4.1 *Limitations*

In this study, the areas that received direct or indirect attention were extracted from each of eight radiologists' visual search behavior and were pooled together to form the available dataset for behavior modeling. Out of this larger dataset, using a random sampling approach [to avoid building dumb model (as described in Sec. 2)], a subset of the dataset was partitioned and used in the modeling. It is possible that for some categories (such as FC, TN) the same area (or overlapping areas) has been used more than once. This, if at all true, would have only occurred for FC and TN categories as there was an overlap in these categories among radiologists. We minimized the occurrence of such influences using random sampling but, if occurred, this may have adversely impacted the training or validation outcomes for said categories.

5 Conclusion

We have shown the radiologists decision outcome (and the confidence in such decisions), and attentional level received at a given area can successfully be modeled, and high accuracy in such predictions can be achieved. We have also shown that there is very high agreement between the predicted outcome and true decision ($k = 0.86$) and attentional level ($k = 0.82$) and that all these models are statistically significantly better than "dumb" models. In addition, these models possess knowledge related to the radiologists' search characteristics and decision making, suggesting that these are "smart" models that learn about the radiologists' behaviors.

Disclosures

The authors have no relevant financial interests in the article and no other potential conflicts of interest to disclose.

Acknowledgments

We would like to thank the radiologists that participated in our experiment. This work has been presented at SPIE Medical Imaging conference in Houston in February 2018 under the title "A deep (learning) dive into visual search behaviour of breast radiologists," paper number 1057708.

References

1. C. Zetzsche, "Natural scene statistics and salient visual features," Chapter 37 in *Neurobiology of Attention*, pp. 226–232, Elsevier Inc., Cambridge, Massachusetts (2005).
2. H. L. Kundel, C. F. Nodine, and L. Toto, "Searching for lung nodules. The guidance of visual scanning," *Invest Radiol.* **26**(9), 777–781 (1991).
3. C. Mello-Thoms et al., "The perception of breast cancer: what differentiates missed from reported cancers in mammography?" *Acad. Radiol.* **9**(9), 1004–1012 (2002).
4. C. Mello-Thoms et al., "The perception of breast cancers—a spatial frequency analysis of what differentiates missed from reported cancers," *IEEE Trans. Med. Imaging* **22**(10), 1297–1306 (2003).
5. C. Mello-Thoms, C. F. Nodine, and H. L. Kundel, "Relating image-based features to mammogram interpretation," *Proc. SPIE* **4686**, 80–83 (2002).
6. S. Mall, P. Brennan, and C. Mello-Thoms, "Fixated and not fixated regions of mammograms: a higher-order statistical analysis of visual search behavior," *Acad. Radiol.* **24**(4), 442–455 (2017).
7. H. D. Nelson et al., "Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. preventive services task force recommendation," *Ann. Intern. Med.* **164**(4), 244–255 (2016).
8. P. T. Huynh, A. M. Jarolimek, and S. Daye, "The false-negative mammogram," *Radiographics* **18**(5), 1137–1154 (1998).
9. H. L. Kundel, "Peripheral vision, structured noise and film reader error," *Radiology* **114**(2), 269–273 (1975).
10. H. L. Kundel, C. F. Nodine, and D. Carmody, "Visual scanning, pattern recognition and decision-making in pulmonary nodule detection," *Invest Radiol.* **13**(3), 175–181 (1978).

11. D. P. Carmody, C. F. Nodine, and H. L. Kundel, "An analysis of perceptual and cognitive factors in radiographic interpretation," *Perception* **9**(3), 339–344 (1980).
12. E. Samei and E. A. Krupinski, *The Handbook of Medical Image Perception and Techniques*, Cambridge University Press, Cambridge (2010).
13. E. P. A. Alberdi, L. Strigini, and P. Ayton, "CAD: risks and benefits for radiologists' decision," in *The Handbook of Medical Image Perception and Techniques*, E. Samei and E. A. Krupinski, Eds., pp. 326–330, Cambridge University Press, Cambridge (2010).
14. C. Mello-Thoms et al., "Using computer-assisted perception to determine the characteristics of missed and reported breast cancers," *Proc. SPIE* **4324**, 64–67 (2001).
15. M. W. Pietrzyk, D. Rannou, and P. C. Brennan, "Implementation of combined SVM-algorithm and computer-aided perception feedback for pulmonary nodule detection," *Proc. SPIE* **8318**, 831815 (2012).
16. Z. Gandomkar et al., "A model based on temporal dynamics of fixations for distinguishing expert radiologists' scanpaths," *Proc. SPIE* **10136**, 1013606 (2017).
17. Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.* **2**(1), 1–127 (2009).
18. I. Arel, D. C. Rose, and T. P. Karnowski, "Research frontier: deep machine learning—a new frontier in artificial intelligence research," *IEEE Comput. Intell. Mag.* **5**(4), 13–18 (2010).
19. H. Greenspan, B. V. Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique," *IEEE Trans. Med. Imaging* **35**(5), 1153–1159 (2016).
20. D. M. Mount et al., "Fast nearest neighbour search (Wraps Arya and Mount's ANN: a library for approximate nearest neighbor searching)," 2015, https://cran.r-project.org/web/packages/RANN/RANN.pdf; https://www.cs.umd.edu/~mount/ANN/.
21. C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," arXiv:1602.07261 (2016).
22. Y. K. Tsehay et al., "Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images," *Proc. SPIE* **10134**, 1013405 (2017).
23. G. Hinton, "RMSProp: neural networks for machine learning 2012," 2017, https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
24. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.* **45**(4), 427–437 (2009).
25. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM* **60**(6), 84–90 (2017).
26. B. Jacob, H. L. Kundel, and R. L. Van Metter, *The Nature of Expertise in Radiology*, pp. 1–2, SPIE Press, Bellingham, Washington (2000).
27. H. L. Kundel et al., "Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms," *Acad. Radiol.* **15**(7), 881–886 (2008).
28. M. Carrasco and B. McElree, "Covert attention accelerates the rate of visual information processing," *Proc. Natl. Acad. Sci. U. S. A.* **98**(9), 5363–5367 (2001).
29. H. W. Lee, G. E. Legge, and A. Ortiz, "Is word recognition different in central and peripheral vision?" *Vision Res.* **43**(26), 2837–2846 (2003).
30. A. Albonico et al., "Temporal dissociation between the focal and orientation components of spatial attention in central and peripheral vision," *Acta Psychol.* **171**, 85–92 (2016).
31. D. P. Chakraborty, "Measuring agreement between rating interpretations and binary clinical interpretations of images: a simulation study of methods for quantifying the clinical relevance of an observer performance paradigm," *Phys. Med. Biol.* **57**(10), 2873–2904 (2012).

Suneeta Mall is a PhD candidate at University of Sydney and is researching in the field visual search behavior and breast imaging. Her thesis title is "Modelling radiologists' visual search behaviour and interpretation of digital mammograms using high order statistics and deep machine learning techniques". She has a bachelor of technology (2007) and has active interest in medical imaging specifically breast imaging (mammography and digital breast tomosynthesis), visual search, and machine learning.

Patrick C. Brennan is a professor of diagnostic radiography at the University of Sydney in Australia. His research interests are in image perception, diagnostic reference levels, and medical image optimization.

Claudia Mello-Thoms is a research associate professor of radiology at the University of Iowa in USA and an honorary associate professor at the University of Sydney in Australia. Her research interests are in image perception and interpretation, machine learning, and human learning.