Using Eye-Gazing Data to Predict Radiologists' Cognitive Behavior during Breast Cancer Screening

Shelby Anne Becker

Saint Mary-of-the-Woods College

Research Alliance in Math and Science

Computing and Computational Sciences Directorate

Oak Ridge National Laboratory

Mentors: Georgia Tourassi, Ph.D and Barbara Beckerman, M.B.A.

Draft

Author Note

This research is supported by The Research Alliance in Math and Science program and sponsored by the Office of Advanced Scientific Computing Research, U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. De-AC05-00OR22725. This work has been authored by a contractor of the U.S. Government, accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

Abstract

Breast cancer is one of the leading causes of death of women in the United States, but early detection can greatly improve survival. Mammography is the most common modality to screen for breast cancer. Mammographic screening has greatly reduced breast cancer mortality, but limitations still exist due to the complexity of the diagnostic task. Research has shown that 10% to 30% of breast lesions are missed and a large portion of them are visible retrospectively. Commercial computer-assisted diagnostic (CAD) systems are used to reduce human error, but the systems ignore each radiologist's perceptual and cognitive needs. There are two types of diagnostic errors associated with mammographic detection of breast cancers: (i) perceptual (failure to see lesions) and (ii) cognitive (failure to report as lesion, a visually-perceived parenchymal structure). Using eye-tracker data from 6 radiologists who viewed 20 mammographic cases, artificial intelligence techniques are applied to capture the radiologists' perceptual and cognitive patterns. Specifically, the use of data mining algorithms investigated whether eye-tracking data could be used to predict the radiologists' diagnostic errors. The findings showed that eye-tracking data could be used to predict a radiologist's error and his decision. Group and individual level modeling tended to be better at predicting behavior when more variables are included, but findings trended to show that while individual modeling offered some advantages, it was not a consistent finding. This research could lead to the next generation of CAD systems that are user-adaptive to each radiologist and case under review.

Draft

1. Introduction

The cost of error in identifying lesions in mammographic screening has fatal consequences. Breast cancer has the second highest death rate in cancers for women in the United States (1) with over 200,000 new cases in 2012 for women and almost 40,000 deaths this year so far (2). Mammographic screening has become the most cost effective and widely-used method for detecting breast cancer, but due to the complexity of the breast parenchyma, interpretation of the breast is difficult.

Mammograms are x-rays of the breast tissue and make it possible to detect lesions that are not felt and calcium deposits that can indicate cancerous lesions (2). The breast parenchyma, or connective tissue of the breast, is difficult to interpret due to the fatty tissue and dense tissue in the breast, leading to a high density breast (3). The dense tissue, like the lesions, appears white on the mammogram, making the cancer harder to detect. The fatty and dense tissue could potential affect detection by masking the lesion or giving the lesion a way to blend in to the parenchyma. The screening has shown to reduce mortality in breast cancer by 30%-39% in women over the age of 50 and by roughly 24% in women over the age of 40 (4). However, 10% to 30% of lesions are missed in the screening process and those missed are also visible in retrospectively (3, 4). Early detection of lesions can greatly change the prognosis, and the use of technology has significantly improved cancer detection though higher image resolution and computer-assisted diagnosis systems (CAD) to pinpoint smaller lesions and more accurate findings (5). Despite mammography being sensitive to breast lesions, the method has

Draft

a low positive predictive value (6) and most cases sent for biopsy are found to be benign (5). Only 10-34% of the biopsy cases are diagnosed as malignant (10).

Better identification of lesions could prevent costly procedures that are also uncomfortable and potentially harmful to the patient. The CAD systems have the potential to decrease the mortality rates, but the effects of using CAD input and radiologists' decisions are not fully understood (3). Studies have shown that a highly sensitive, highly specific CAD system aids mammography, but low sensitivity and specificity allows for greater error (5). This is important in detection of lesions as is a radiologist's ability to scan and interpret the image.

To search out lesions in a mammographic image requires the ability to interpret the complex image. The errors in missing the lesions are due to search and integration errors as well as training and the image (5). Studies have tracked radiologists' eye movement and have shown that roughly 62% (7) to 70% (4) of missed cancers have fixated upon (6). Mello-Thoms theorized that visual dwell is strongly related to visual attention in her paper The Perception of Breast Cancers-A Spatial Frequency Analysis of What Differentiates Missed from Reported Cancers." Eye-tracking studies have also shown that malign tumors are dwelled upon longest as are lesion-free areas, which they report as having a lesion, but lesions that are not reported have shorter dwell time (8). This combined with CAD systems that ignore the radiologist's error increases the likely hood of a missed lesion due to perceptual and cognitive error. The radiologist's ability to detect a lesion is in response to the relationship and placement of a lesion against the background of the breast tissue (9). The increased capability of a CAD scheme and the radiologists' eye movements could greatly reduce the number of missed lesions. In regards of this paper, perceptual error is defined as the failure of the radiologist to

fixate or detect the lesion and cognitive error is defined as a failure to report a visually perceived lesion.

The objective of this paper is to determine whether or not eye tracking data can be used to predict the radiologists' cognitive behavior and to differentiate between perceptual and cognitive error. The paper will also determine if prediction of errors is more effective at the individual level or group level with experience playing a key role. The study will also determine if predictive modeling is more effective at the individual or group level.

Draft

2. Methods and Materials

A. Observer Selection

Six radiologists participated in this data collection. Three of the radiologists were experienced mammographers with large number of case readings in the background. The three other participants were radiology students from the same institution and had a limited number of case-reviews and experience.

B. Experimental Protocol

The radiologists reviewed a set of 20 cases. Each case had cranicaudel and mediolateral oblique views. The cases were pulled from previous cases and reviewed by an experienced radiologists not participating in the study (11). The cases were digitized into the computer and the participants were to identify malignant lesions until they were confident in their diagnosis and the radiologists were tracked using a head mounted eye-tracking system. This system

calculated eye-gazing with an accuracy of 1 degree while it monitors corneal and pupil reflections (12).

C. Statistical Analysis

The diagnostic decision points taken from the study were analyzed using Gabor wavelets to derive textural signals and the eye-tracking data was recording in terms of length of dwell on each image, number of returns, and total dwelling upon the image. Classification was performed using a leave-one-out cross validation system to determine if the objectives stated previously could be supported. The data was analyzed using statistical and data mining software: Microsoft Excel, JMP, and Weka. In WEKA or Waikato Environment for Knowledge Analysis, predictive modeling was carried out using logistic regression and artificial neural networks (ANN). The ANN then tested and validated the modeling using a leave-one-out cross validation scheme. The leave-one-out cross validation scheme keeps one of the observations out while the rest of the data is used as training data and then this process is repeated until each data point is used to validate the model. K-Fold validation was used in WEKA, where K is the number of points in the sample. This was run using samples as all observations, per group of radiologists based on experience, and individual readers to determine the accuracy of the model and comparison. In addition, trend analysis was also used on the levels mention previously using correlation and Analysis of Variance (ANOVA) to determine the relationship of variables between each other. Correlation analysis determined the strength of the relationship between the various variables. Most correlations were run with JMP and Excel to verify the data given by Excel. In the data set, the readers are defined as numbers 1-6 to differentiated

Draft

the reader and prevent bias. The experienced mammographers are label 4, 5, and 6. The inexperienced radiologists are labeled as 1, 2, and 3.

3. Results

A.

Running the data through WEKA, results were taken from the classification rate or percentage of the validated data being correct.

B. Can eye tracking data be used to predict radiologists' cognitive behavior?

The results showed trends that the eye tracking data can predict a radiologist's cognitive behavior. Using the variables of "total dwell", "returns", and "dwell length" on "error" in an overall grouping analysis, groups by experience, and individuals, the following is summarized below:

y=f (dwell, initial, return) in % Correct on Error

% Error	**Table 1**	
	Logistic	ANN
All	79.67%	81.6400%
Per Group		
(1-3)	80%	77.14%
(4-6)	78.65%	75.7300%
Individual		
1	78.79%	81.82%
2	75%	83.33%
3	72.22%	72.22%
4	75.68%	65.68%
5	82.86%	82.86%
6	35.49%	45.16%

Draft

%Error	Table 2			
	Logistic		ANN	
	No	Yes	No	Yes
All	77.80%	80.60%	81.30%	82.00%
Per Group				
(1-3)	80%	79%	82.22%	73.30%
(4-6)	73.10%	84.30%	71.40%	79.60%
Individual				
1	73.30%	83.30%	78.60%	84.20%
2	73.70%	76.50%	92.90%	77%
3	72.00%	72.70%	76.20%	66.70%
4	71.40%	81.30%	66.70%	68.40%
5	85.00%	80.00%	88.90%	76.50%
6	0.00%	47.80%	22.20%	54.50%

Table 1 displays in percent correctly classified in the overall model. Table 2 displays the percentage that is correctly labeled as an error or not. In the data table, if an error has occurred, then the error is labeled as yes and if no error has occurred, then it is classified as no. Included in the tables are the classification accuracies for both the logistic regression modeling and artificial neural network modeling. From this, the ANN proved to be more accurate, but both models tended to show above 70% in the classification. The individual modeling had a great deal of variability ranging from about 35% to 82%. This could be caused by small sample sizes, when the data points pull by reader of each radiologist was around 33 points whereas the overall group was 207 data points. The group models predicted the trainees or radiology residents more accurately than the experts. This follows previous studies. The tables also show a trend that the modeling more accurately predicts no error in the overall and group models, but better predict the errors in the individuals.

Draft

C. Can eye tracking data be used to differentiate perceptual and cognitive errors?

The models run in WEKA using logistic regression and ANN used the variables of "total dwell", "returns", and "dwell length" on "decision" in an overall grouping analysis, groups by experience, and individuals, to determine whether or not the eye tracking data could be used to differentiate between perceptual and cognitive error. The findings are summarized below:

y=f (dwell, initial, return)

% Decision		
	Logistic	ANN
All	74.40%	75.36%
Per Group		
(1-3)	73.34%	70.48%
(4-6)	75.30%	78.65%
Individual		
1	72.73%	72.73%
2	69.44%	63.89%
3	75%	75%
4	72.93%	70.27%
5	80%	82.86%
6	51.62%	41.94%

Draft

% Decision	Logistic				ANN			
	fn	fp	tn	Tp	fn	fp	tn	tp
All	0%	0%	74.50%	74.30%	66.70%	0%	73.50%	77.40%
Per Group								
(1-3)	0%	0%	75.50%	71.40%	50.00%	0%	73%	69.00%
(4-6)	0%	0%	70.00%	81.10%	0%	0%	76%	79%
Individual								
1	0%	0%	64.30%	78.90%	0%	0%	62.50%	82.40%
2	50.00%	0%	75.00%	68.80%	0%	0%	66.70%	64.70%
3	0%	0%	82%	68.40%	0%	0%	76%	73%
4	0%	0%	66.70%	81.30%	0%	0%	66.70%	85.70%
5	0%	0%	80%	80.00%	0%	0%	76.50%	88.90%
6	0%	0%	42.90%	59.10%	0%	0%	30.00%	52.60%

In the tables, table 3 follows like table 1 in summarizing the overall percent classification accuracy of the modeling using the logistic regression and ANN functions in WEKA. Table 4 further goes into detail in displaying the percent accuracies for the decision of all data points, by group experience, and individual reader. The objective of these models was to determine if the data could differentiate between a cognitive and perceptual error. The data trends for the overall accuracy in differentiating between the data was 74.4% and for the groups, 73% for radiologists 1-3 and 75% for radiologists 4-5. The individual runs showed a range off 41-82% for both the logistic regression and ANN models. The artificial neural network (ANN) tended to have a higher accuracy rate then the logistic regression and the data could better predict the accuracy rate for the experienced radiologists than the less experienced mammographers. The modeling also tended to show higher accuracy rates for true negative and true positive decisions. True positive decisions mean that the radiologist correctly perceived and identified the case as a

lesion. True negative decisions are defined as that the radiologist correctly perceived and cognitively identified the case as normal. The higher accuracy rates could be caused by the small sample size per individual, as some of the readers lacked a decision for a false positive or negative or simply had one point for the false positives and false negatives. The accuracy rates for the groups versus the individual showed that curtained individuals had higher accuracy rates then the group accuracy rates. The data for the individuals also showed that often the accuracy rate for the true negative was higher than that of the true positive and inversely for the higher true positives to lower percentages of true negatives.

Draft

D. Determine whether predictive modeling is more effective at the individual level rather than the group level.

Determining whether the predictive modeling is more effective at the individual level rather than the group level depends on the sample size for each and how detailed the modeling is. In the predictive modeling based on the error, the group level had a percentage of 79.67% and 81.64% for logistic and ANN respectively and the individual percent accuracies ranged from 35%-82% for the logistic regression and 45%-83% for the artificial neural network. Two of the three inexperienced radiologists had higher accuracy rates than that of the group accuracy rate and the third radiologist (73%) was within four percent of the group accuracy rate (77%). The accuracy rate for the group of experienced radiologist (75%) was more accurate than the individual rates. The accuracy rates for correct versus incorrect from table 2 displayed some different results. The individual reader, number 6, had very low accuracy rates and was the outlier in all tests. In comparing

individual rates to group rates, overall accuracy rates were more efficient and accurate to predict correct or incorrect, whereas the individual accuracy rates were better than that of the group accuracy rates.

In the predictive modeling based on the decision, the group accuracies rates, especially the ANN percentages, tended to be more effective in prediction than that of the overall accuracies. Some individuals had very high accuracy rates of 80%-82% (reader 5), but some had very low accuracy scores of less than 51% (reader 6). In the more detailed modeling, displayed on table 4, the accuracy rates of individuals were more often higher than that of the groups or overall accuracy rates in the true negative logistic regression. The accuracy rate for the experienced readers was higher than that of the overall accuracy rates and the individual rates in the logistic regression, but the group accuracy for the experienced radiologist tended to lower than that of many of the individual readers in the ANN. In all cases and decisions, the group accuracy rates for the experienced radiologists were higher than that of the inexperienced radiologists. Individual radiologist accuracy scores also seemed to be more effective for inexperienced radiologist than group or overall accuracy rates. Overall accuracies for false positive decisions are more effective at the overall level.

4. Conclusion

From analysis, predictive modeling has been carried out to interpret the eye-glazing data from that of a radiologist to predict behaviors and decisions. Predictive modeling was done on the basis of a radiologist's error and decisions using the variables of total time, dwell, and number of returns to an area on the image. The objectives of this paper sought

Draft

to offer an answer to whether eye-tracking data can predict a radiologist's cognitive behavior, differentiate a radiologist's perceptual and cognitive errors, and determine if individualization of models is more effective or not that the group levels.

The modeling found that the data could predict a radiologist's cognitive behavior, especially if the radiologist correctly a diagnosis the image. The overall models were just as effective at predicting the behavior the cognitive behavior of the inexperienced radiologists as the group level. However, at the group level, predicting the correct decision of the experienced mammographers was more effective than at the overall level and individual levels. This suggests experienced radiologist have a higher chance of correctly diagnosing the image. This supports previous studies and suggests the way radiologists are training in a specific search pattern, method, and experience could change how a radiologist scans mammograms for lesions.

Predictive modeling of differentiating between the radiologist's perceptual and cognitive errors was possible with above 70% accuracy in the overall and group levels. Group level modeling had higher accuracy rates on both logistic regression and the ANN. However, individual level modeling tended to be much more effective for many of the radiologists, especially the experienced mammographers, for diagnosing the image as true positive. The modeling was also more effective at predicting the true negative and positive decisions, but could not identify the false positives at any level. Only the ANN could detect false negatives at the overall and group levels. This could be due to the surety of radiologists seeing a normal or easy to see lesion or the radiologist does not see the details that suggest a lesion or not.

Draft

The determination as to whether or not the individual modeling was more effective than the group or global models was inconsistent. Some individuals had very high predictive accuracies and others did not. Reader number 6 was a challenging exception all around and even though some of the data offered some advantages of individual predictive models, there were not enough parameters to determine if it was truly effective.

5. Future Work

The biggest limitation upon this study was a small sample size. The modeling that was conducted using all data points was 207 and had much more of an impact upon training the artificial neural networks as compared to the smaller sample sizes at the group (107 data points) and individual level modeling (between 33-37 data points). By recruiting more radiologists to review more cases as well as to expand the set of eye-gazing parameters, more accurate and elaborate models can be created. In addition, in expanding the models support vector machine predictive modeling could be used to improve predictive modeling. Lastly, based on the findings, a development of a computerized decision support system that adapts to each radiologist's eye-gaze pattern seems promising.

6. Acknowledgements

This work was supported by The Research Alliance in Math and Science program and sponsored by the Office of Advanced Scientific Computing Research, U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. De-AC05-00OR22725. This author would like to thank Oak

Draft

Ridge National Laboratories, Georgia Tourassi, Barbara Beckerman, and the RAMS program for their support and input.

Draft

7. References

(1) Chakraborty, D., Yoon, H.-J., & Mello-Thoms, C. (2007; 14). Spatial Localization of Radiologist in Free-Response Studies: Inferring Percetual FROC curves from Mark Rating Data. *Academy of Radiology*, 4-18.

(2) National Cancer Institute. (2012). *Breast Cancer*. Retrieved June 20, 2012, from National Cancer Institute at the National Institute of Health: http://www.cancer.gov/cancertopics/types/breast

(3) Mello-Thoms, C. (2010). Visual Search Characterisitics in Mammogram Reading: SFM vs. FFDM. *Medical Imaging*.

(4) Mello-Thoms, C., Dunn, S. M., Nodine, C. F., & Kundel, H. L. (2003; 22). The Perception of Breast Cancer-A Spatial Frequency Analysis of What Differentiates Missed From Reported Cancers. *IEEE Transactions on Medical Imaging*, 1297-1305.

(5) Mello-Thoms, C., Dunn, S., Nodine, C. F., Kundel, H. L., & Weinstein, S. P. (2002;179). The Perception of Breast Cancer: What Differentiates Missed From Reported Cancers in Mammography? *Academy of Radiology*, 1004-1012.

(6) Kopans, D. (1992). THe Positive Predictive Value of Mammography. *American Roeentgen Ray Society*, 521-526.

(7) Mello-Thoms, C., Dunn, S. M., Nodine, C. F., & Kundel, H. L. (2001). An Analysis of Perceptual Errors in Reading Mammograms Using Quasi-Local Spatial Frequency Analysis. *Journal of Digital Imaging; 14*, 117-123.

(8) H.L Kundel, C. N. (1989). Searching for Lung Nodules: Visual Dwell indicates Locations of false-positives and false negarive decisions. *Investigative Radiology*, 472-478.

(9) Robinson, P. A. (1997; 70). Radiology's Achilles Heel: Error and Variation in the Interpreation of the Rontgen Image. *The British Journal of Radiology*, 1085-1098.

(10) Floyd Jr., Carey E., J. Y. Lo, and G. D. Tourassi (2000, 175). Case-Based Reasoning Computer Algorithm that Uses Mammographic Findings for Breast Biopsy Decisions. *American Journal of Roentgenology*, 1347-1352.

(11) Tourassi, G. D., V. Paquit, and E. Krupinski (2012). Machine Learning Analysis of Radiologists' Eye-Gaze Data, Decisions, and Local Image Texture to Reduce Diagnositic Errors in Screening Mammography.

(12) Applied Science Laboratories. (2009). *Applied Science Laboratories*. Retrieved June 2012, 2012, from Applied Science Laboratories: http://www.asleyetracking.com/Site/.

Draft