

# Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

## Modeling sequential context effects in diagnostic interpretation of screening mammograms

Folami Alamudun  
Paige Paulus  
Hong-Jun Yoon  
Georgia Tourassi

**SPIE.**

Folami Alamudun, Paige Paulus, Hong-Jun Yoon, Georgia Tourassi, "Modeling sequential context effects in diagnostic interpretation of screening mammograms," *J. Med. Imag.* **5**(3), 031408 (2018), doi: 10.1117/1.JMI.5.3.031408.

# Modeling sequential context effects in diagnostic interpretation of screening mammograms

Folami Alamudun,<sup>a,b,\*</sup> Paige Paulus,<sup>c</sup> Hong-Jun Yoon,<sup>a,b</sup> and Georgia Tourassi<sup>a,b,c</sup>

<sup>a</sup>Oak Ridge National Laboratory, Computational Sciences and Engineering Division, Oak Ridge, Tennessee, United States

<sup>b</sup>Oak Ridge National Laboratory, Health Data Sciences Institute, Oak Ridge, Tennessee, United States

<sup>c</sup>University of Tennessee, Department of Mechanical, Aerospace, and Biomedical Engineering, Knoxville, Tennessee, United States

**Abstract.** Prior research has shown that physicians' medical decisions can be influenced by sequential context, particularly in cases where successive stimuli exhibit similar characteristics when analyzing medical images. This type of systematic error is known to psychophysicists as sequential context effect as it indicates that judgments are influenced by features of and decisions about the preceding case in the sequence of examined cases, rather than being based solely on the peculiarities unique to the present case. We determine if radiologists experience some form of context bias, using screening mammography as the use case. To this end, we explore correlations between previous perceptual behavior and diagnostic decisions and current decisions. We hypothesize that a radiologist's visual search pattern and diagnostic decisions in previous cases are predictive of the radiologist's current diagnostic decisions. To test our hypothesis, we tasked 10 radiologists of varied experience to conduct blind reviews of 100 four-view screening mammograms. Eye-tracking data and diagnostic decisions were collected from each radiologist under conditions mimicking clinical practice. Perceptual behavior was quantified using the fractal dimension of gaze scanpath, which was computed using the Minkowski–Bouligand box-counting method. To test the effect of previous behavior and decisions, we conducted a multifactor fixed-effects ANOVA. Further, to examine the predictive value of previous perceptual behavior and decisions, we trained and evaluated a predictive model for radiologists' current diagnostic decisions. ANOVA tests showed that previous visual behavior, characterized by fractal analysis, previous diagnostic decisions, and image characteristics of previous cases are significant predictors of current diagnostic decisions. Additionally, predictive modeling of diagnostic decisions showed an overall improvement in prediction error when the model is trained on additional information about previous perceptual behavior and diagnostic decisions. © The Authors. Published by SPIE under a Creative Commons Attribution 3.0 Unported License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: [10.1117/1.JMI.5.3.031408](https://doi.org/10.1117/1.JMI.5.3.031408)]

Keywords: perceptual behavior; eye tracking; mammography; cognitive bias; context effect.

Paper 17366SSR received Dec. 18, 2017; accepted for publication Feb. 19, 2018; published online Mar. 19, 2018.

## 1 Introduction

In clinical practice, timely and accurate diagnostic decisions carry important implications for subsequent monitoring, treatment, prognosis, and quality of life. Decisions on disease occurrence are primarily based on well-defined criteria but, ultimately, the judgment of the practitioner trained for this purpose and shaped by years of experience plays the decisive role. However, studies suggest that, even with experienced practitioners, medical decision-making may also be influenced by psychological factors including perception, cognition, emotion, and other similar factors.<sup>1</sup>

Humans have an innate ability for detecting patterns when presented with static visual stimuli. In addition, when exposed to sequential or dynamic stimuli, humans have an advanced ability to detect sequential patterns.<sup>2</sup> While the capabilities for processing visual patterns are historically advantageous for certain adaptations for survival, they can pose problems when performing tasks of an independent but sequential nature. As such, when humans are faced with uncertainty while engaging in predictive or discrimination tasks, which are sequential in

nature, there is a tendency for perceived patterns to subconsciously influence or bias human decisions based on previous states within the sequence.<sup>3–5</sup>

In addition to being sequential in nature, medical decisions involve individual cases whose pathologies are independent of one another, thereby creating additional potential for bias, in which the practitioner may erroneously (subconsciously) assume that future probabilities are affected by current outcomes (such as mentally adjusting the prior probability distribution for a given pathology after the occurrence of a streaky sequence). This phenomenon is described in literature as context effect.<sup>6,7</sup>

### 1.1 Factors Affecting Diagnostic Accuracy

Several factors influence the occurrence of error in the decision process in medical diagnosis. These errors can affect subsequent treatment and have a significant negative impact on patients. In mammography, the characteristics of a case, such as breast density,<sup>8,9</sup> physician experience levels, and metrics describing the visual search process have been associated with the occurrence of diagnostic error during the interpretation of mammograms.<sup>10–14</sup> Studies have examined the perceptual and cognitive dimensions in visual diagnosis.

Perceptual errors, i.e., errors where the observer fails to recognize an abnormality in an image during the detection phase,

\*Address all correspondence to: Folami Alamudun, E-mail: [alamudunff@ornl.gov](mailto:alamudunff@ornl.gov)

are a primary cause of missed diagnoses in radiology.<sup>15,16</sup> These types of errors result from cognitive biases,<sup>17–19</sup> faulty visual search,<sup>20,21</sup> or a combination of both. Since there are interrelated cognitive contributors to visual errors, understanding the cognitive processes involved during the interpretation of images can shed light on inherent biases that occur and how these impact diagnostic decisions.

## 1.2 Local Sequential Context Bias

Prior research has shown that medical decisions can be influenced by sequential context, particularly in cases where successive stimuli exhibit similar characteristics. Laming<sup>22</sup> investigated systematic errors in diagnostic decisions in screen cervical smears. Laming reported that experienced pathologists made systematic errors resulting from psychological interactions in successive cases and offered recommendations on prevention. This type of systematic error, peculiar to sequential stimulus with similar characteristics (as is the case in mammography), indicates that judgments are influenced by features of and decisions about the preceding case in the sequence of examined cases, rather than being based solely on the peculiarities unique to the present case. This type of cognitive bias, known to psychophysicists as sequential context effect,<sup>23–25</sup> is studied in radiology literature and referred to in some cases as context bias.<sup>4,5</sup>

## 1.3 Global Sequential Context Bias

In contrast to local sequential context bias, global sequential context bias refers to the ability to detect patterns in sequences, such as prevalence in a specific type of case, and subconsciously adjusting *a priori* probabilities accordingly. In clinical practice, the sequential nature of diagnostic imaging, combined with the ability to detect patterns increases the possibility for biased decisions particularly in borderline cases where the practitioner may be less certain of the diagnostic decision. While previous studies have focused on the occurrence of bias resulting from case prevalence recognized within the case sequence (which may be a test-specific phenomena), fewer have studied the effect of prior diagnostic decisions on subsequent ones.

Sequential biases may occur in combination particularly in borderline cases where the contextual similarities with a previous case and the inference of case prevalence in case sequence combine to influence diagnostic decisions. In this work, we examine the diagnostic decisions of radiologists with varied levels of experience to determine if and to what extent contextual biases occur and if these occurrences can be predicted. To this end, we explore correlations between current and present perceptual behavior, diagnostic decisions, and case characteristics. We hypothesize that radiologist's prior perceptual behavior and cognitive decisions are predictive of current decisions. To test our hypothesis, we tasked 10 image readers of varied experience to conduct blind reviews of 100 x-ray images. Eye-tracking data and diagnostic decisions were collected from each reader while reviewing four-view mammographic cases under conditions mimicking clinical practice.

## 2 Methods

### 2.1 Data Collection Protocol

A description of methods for this study is provided in detail in a previous work.<sup>10</sup> Briefly, we selected 100 screen-film

mammograms from a corpus of mammographic cases available from the University of South Florida's Digital Database for Screening Mammography<sup>26</sup> (DDSM). Each case included the craniocaudal (CC) and mediolateral oblique views (MLO) of both breasts with associated ground truth using BI-RADS<sup>TM</sup> lexicon.<sup>27</sup> Fifty percent of the selected cases were malignant, whereas the remaining 50% was evenly split between normal and benign pathologies. The parenchymal density ranged between 1 (fatty) to 4 (dense), according to the BI-RADS<sup>TM</sup> lexicon.

For this study, each of the 10 participants (three board certified radiologists and seven radiology residents) performed a review of all 100 cases and provided a report on location and corresponding BI-RADS<sup>TM</sup> rating of any suspicious mass. Reviews were conducted independently, and markings and BI-RADS<sup>TM</sup> ratings were reported by each reader through a custom-built graphical user interface, which allowed for image manipulation on dual displays as customary in clinical practice. Readers were outfitted with H6 head-mounted eye-tracking device, with a 60-Hz sampling rate, designed with eye-head integration from Applied Science Laboratories (ASL, Bedford, Massachusetts). Through the remote eye-tracking device, readers' eye-position data were recorded to within 0.5-deg visual accuracy. Institutional review board approval was obtained prior to conducting this study. Human subject recruitment and data collection were performed according to a protocol approved by the Oak Ridge Site-Wide Internal Review Board. All readers reviewed and signed an informed consent form prior to participation.

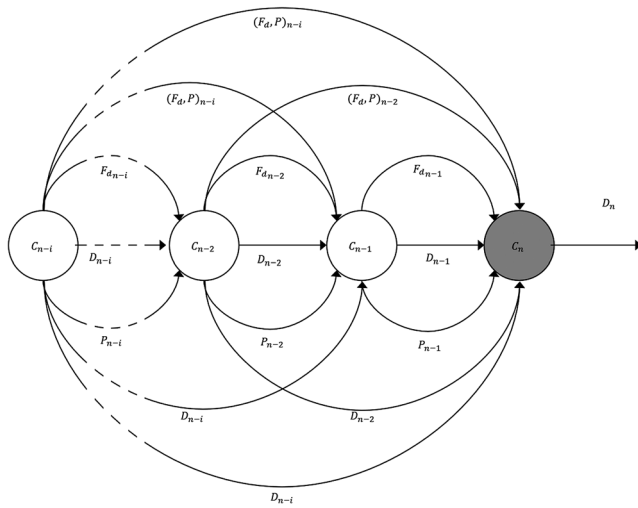
Cases for each reader were presented in a randomized order. Although readers examined an identical set of cases, ordering resulting from the randomization process differed for each one. In addition, readers were permitted to complete the study in multiple sessions to accommodate personal preferences and scheduling conflicts. Some readers completed the experiment in two or more sessions on the same day, whereas others completed the experiment over two or more days.

### 2.2 Data Processing

Details describing data preprocessing is provided in a previous work.<sup>10</sup> In summary, eye-position data for each reader was aggregated from mammographic views across two displays and first preprocessed to extract a time ordered sequence of fixations along with other associated measures (including fixation duration, interfixation degree, and pupil size) using EyeNAL analysis software from ASL.<sup>28</sup> Data provided through EyeNAL software can be further processed to extract more informative features for a deeper analysis.

For our analysis, we used raw gaze data from eye movements to characterize overall visual behavior. Specifically, we characterize the trajectory of the scanpath using a metric known as fractal dimension (FD). The scanpath describes the trajectory of eye movements formed when time-ordered fixations or gaze points are connected. From the scanpath, we computed the FD) using the Minkowski–Bouligand box-counting method<sup>29</sup> as a measure of complexity in search behavior for each case. Information describing the computation and analysis of this metric can be found in detail in a previous work.<sup>10</sup>

The 10 participating readers were grouped into one of three experience levels according to years of training and practice: new trainee resident (N) for those residents with at most two mammo rotations, advanced trainee resident (A) for those



**Fig. 1** Illustrates potential sequential dependence between diagnostic decisions ( $D_n$ ) for a case ( $C_n$ ) and the visual behavior (FD  $F_{n-i}$ ), breast parenchymal density ( $P_{n-i}$ ), and diagnostic decisions ( $D_{n-i}$ ) during  $i$  previous cases ( $C_{n-i}$ ).

residents with greater than two mammo rotations, and expert radiologist (E) for board-certified mammographers. Reported diagnostic decisions for each case were mapped into one of the three case pathologies (normal, benign, and malignant) based on BI-RADS™ rating. Cases without markings (i.e., no scores were given) were designated normal (N); BI-RADS™ ratings {2 and 3} as benign (B); and BI-RADS™ ratings {4A, 4B, 4C, and 5} as malignant (M). Three breast parenchyma density groupings were formed by combining heterogeneous and dense parenchymal cases in the same density grouping because of a small sample size.

Data for each reader were structured into overlapping pairs such that the  $n$ 'th case ( $C_n$ ) was paired with  $k$  previous cases ( $C_{n-i}$ , where  $i$  is an integer between 0 and 5) with each  $n$ 'th case representing the current visual behavior and diagnostic decision, and the  $i$  preceding cases representing previous visual behavior and diagnostic decisions. Each pair represents the reader's diagnostic decision for the current case ( $C_n$ ), and the previous behavior and diagnostic decisions ( $\bigcup_{i=0}^5 C_{n-i}$ ), which may have potentially impacted it.

Our analysis was performed on each pair by extracting case descriptors. Namely, visual behavior, computed as the FD of scanpath, diagnostic decisions determined as BI-RADS™ ratings reported by each reader for the respective cases, and image characteristics/properties for which we utilize the breast parenchymal density (Fig. 1).

### 2.3 Statistical Methods

To perform an initial analysis into the temporal relationship in behavior during mammographic screening, we used a statistical tool: autocorrelation function. Autocorrelation captures the Pearson correlation between values of a sequential process at different times, as a function of these times or the lag between them. It enables finding the correlation between previous measures and measures at subsequent time intervals.

In regression analysis using time series data, autocorrelation in a variable of interest is typically modeled with an autoregressive moving average model (ARMA). We represent the sequence of mammographic cases seen by each radiologist as

a time series ( $C_1, C_2, \dots, C_n$ ). The autocorrelation function gives the correlation (y axis) between the value of measurements at  $C_n$  and the value of measurements at  $C_{n-k}$  as a function of  $k$ , where  $k$  is an integer representing the lag (x axis).

To determine any combined or individual dependencies between the current diagnostic decisions and the radiologists' current and previous visual behavior, previous diagnostic decisions, and characteristics of the current and previous images, we performed a multifactor fixed-effects ANOVA with six levels for diagnostic decision (based on the BI-RADS™ rating provided), three levels for breast parenchyma density, and the FD of scanpaths for eye movements.

## 3 Results

Data from all 10 radiologists who participated in reading mammographic cases were included in this analysis. Each radiologist interpreted 100 cases, resulting in 1000 individual interpretations. We performed analysis on data from each radiologist independently.

### 3.1 Autocorrelation Tests for Randomness

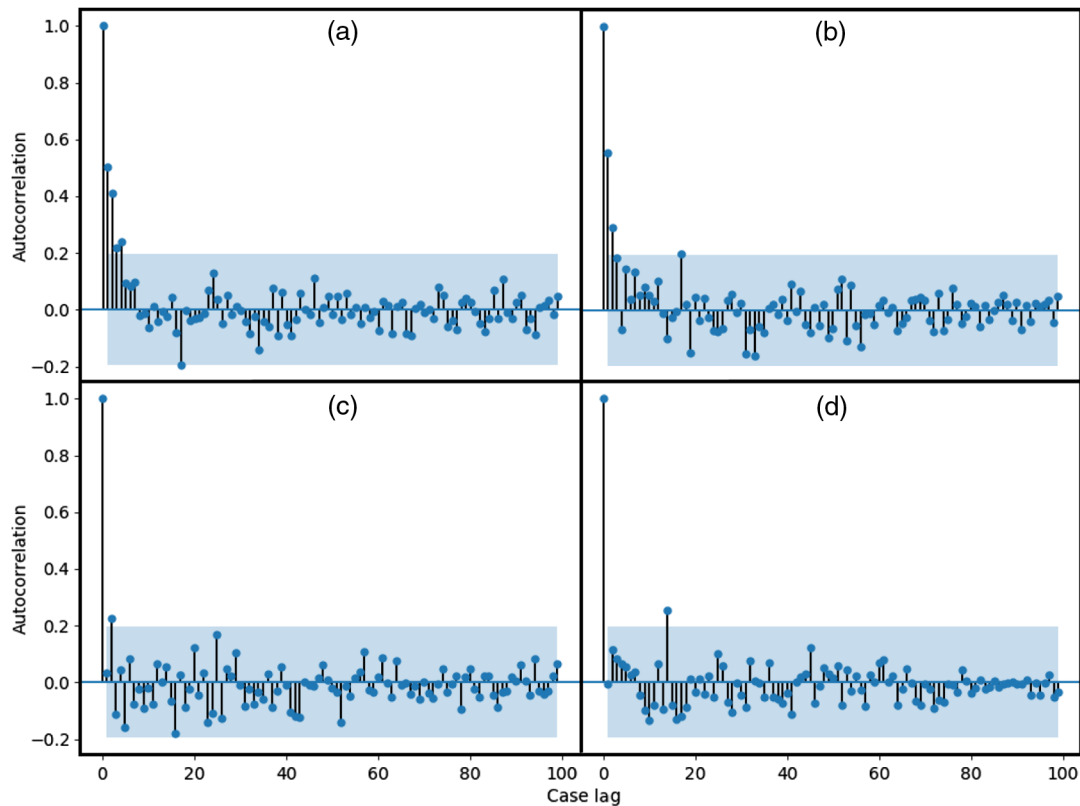
Autocorrelation is the correlation between a time series lagged one or more time periods and itself. The autocorrelation plot provides a visual representation of the degree of dependence (or independence) between a given time series or sequence and a lagged version of itself over successive intervals. In Fig. 2, we illustrate the autocorrelation function graphs of the FD of scanpaths and of the diagnostic decisions for a subset of readers.

The autocorrelation coefficient summarizes the strength of the relationship between present and past values. The shaded region in the graph represents a 95% confidence interval for significance (i.e., those points within this shaded region do not meet this criteria for statistical significance). From Fig. 2, we note a number of nonsignificant correlation peaks, which tend to indicate a random average distribution. However, in a number of cases we identify several significant ( $p < 0.05$ ) peaks, which show both strong ( $r_k \geq 0.5$ ) and weak ( $0.5 > r_k > 0.2$ ) correlation between sequences at their respective time lags.

### 3.2 Analysis of Variance

For our analysis, we restricted previous behavior and diagnostic decisions to five preceding cases. Therefore, each pair of observations consisted of the current diagnostic decision (dependent response variable) and a set of the current visual behavior and breast parenchymal density, along with the diagnostic decision, visual behavior, and breast parenchymal density of the five preceding cases (factors).

A multifactor fixed-effect ANOVA was performed to determine any individual or combined dependencies of the current diagnostic decision on current and previous behavior and diagnostic decisions. The ANOVA was performed individually for each reader. First, we mapped the diagnostic decision for each case to one of seven categories based on the provided BI-RADS™ rating: cases without markings (i.e., no scores or marks were given); BI-RADS™ ratings {2 and 3}; and BI-RADS™ ratings {4A, 4B, 4C, and 5}. We formed three breast parenchymal density groupings (fatty, fibroglandular, and heterogeneous/dense). This was achieved by combining heterogeneous and dense cases in a single group, which became necessary resulting from a small sample size.



**Fig. 2** Illustrates autocorrelation function plots for some readers: (a, b) FD of scanpath and (c, d) diagnostic decision.

Next, we formed pairs of observations with the current diagnostic decision at case  $n$  ( $D_n$ ) as the response variable, and the current and previous estimates for the FD of scanpath ( $F_n, F_{n-1}, F_{n-2}, \dots, F_{n-5}$ ) (continuous factor), the parenchymal density for the current and previous cases ( $P_n, P_{n-1}, P_{n-2}, \dots, P_{n-5}$ ), and previous diagnostic decisions ( $D_{n-1}, D_{n-2}, \dots, D_{n-5}$ ), as independent factors. In Table 1, we summarize the ANOVA results by presenting a subset of statistically significant individual and combined effects.

### 3.3 Predictive-Model-Based Analysis

As described in Sec. 3.1, each data point consists of a pair: diagnostic decision ( $D_n$ ) at case  $C_n$  as target, and reader's diagnostic decision and behavior for  $C_{n-i}$  previous cases, where  $0 \leq i \leq 5$  (our analysis was limited to a maximum of five previous cases) as features. This process results in six distinct datasets representing  $\bigcup_{i=0}^5 (C_n, C_{n-i})$  pairings with a maximum of  $m = 100 - i$  samples for each subject in each dataset. During the experimental protocol (described in Sec. 2.1), readers were permitted to complete the experiment over multiple sessions. The final number of pairs for each reader varied 85 to 98 (with a median value of 94) dependent upon the number of sessions, resulting in  $m = 100 - i \times (s + 1)$  pairs, where  $s$  is the total number of sessions.

To assess the predictive potential of the observed dependencies between current diagnostic decisions and previous behavior, we developed a predictive model using a random forest classification model. The model was evaluated using a within-subject leave-one-case-out cross-validation model evaluation scheme. To achieve this, data were first partitioned into 10

independent datasets, representing response data from each of the 10 readers. Leave-one-case-out cross-validation scheme partitions each dataset into complementary subsets, where all but a single data sample is utilized in model-training (training set), which analysis is validated on the excluded sample (test case). This process is repeated until every sample in the dataset is utilized as the singular test case, resulting in  $m$  models (where  $m$  is the total number of samples).

The aggregated predictive values over all rounds serve as the final performance evaluation of the predictive model for each user. In this paper, we report the weighted-average  $f$ -score (the harmonic mean of precision and recall) performance metrics for predicting diagnostic decision on a current case using features derived from previous behavior and diagnostic decisions. In Table 2, we report the  $f$ -score for each user by  $i$  (number of previous cases) and, for comparison purposes, Table 2 also includes performance metrics for a random chance classifier (a random classifier provides a simple baseline reflecting the predictive outcome of a random guess).

## 4 Discussions and Conclusion

In this study, we examined the potential for bias in diagnostic by analyzing visual behavior and diagnostic decisions of 10 radiologists viewing mammograms for breast cancer screening. To this end, we performed autocorrelation tests, analyses of variance, and predictive-model-based analysis on the diagnostic decision and visual behavior (characterized using the FD of scanpath) for individual readers' case sequences to determine if temporal dependencies exist between current and previous behavior and decisions.

**Table 1** Multifactor fixed-effect ANOVA test results with FD ( $F$ ) of scanpath, prior diagnostic decision ( $D$ ), and breast parenchymal density ( $P$ ) for readers by experience level: new (N) and advanced (A) radiology resident, and expert (E) radiologists.

Reader	Factor	$F$	$p > F$
E3	$F_0:P_0:F_1:D_1:F_2:P_2:D_2$	13.76	$<1e^{-3}$
N1, N2, A3, E1	$F_0:F_1:F_3:F_4:F_5$	9.17	$<1e^{-3}$
A3, E1	$F_0:F_2:F_3:F_4$	8.39	0.01
A3	$F_0:P_0:F_1:P_1:D_1:D_2$	6.97	0.01
N1, N2, A3, E1	$F_0:F_1:F_2:F_3:F_4:F_5$	6.83	0.01
A2, A3	$F_0:F_3:F_4:F_5$	7.55	0.01
N1, E2	$P_0:F_0:F_1:P_1:D_1:F_2:P_2:D_2:F_3:P_3:D_3$	6.40	0.01
E1	$P_0:F_0:F_1:P_1:D_1:P_2:D_2$	6.35	0.01
E3	$P_0:F_1:P_1:D_1$	6.34	0.01
A2, E2	$F_0:F_1:P_1:D_1:F_2:P_2:D_2$	6.20	0.02
N3, A1	$P_0:F_0:F_1:P_1:D_1:F_2:P_2:D_2:P_3:D_3$	6.08	0.02
A4	$D_1:D_2:D_3:D_4$	6.02	0.02
N1, N2, A4, E3	$P_0:F_0:F_1:P_1:D_1$	5.94	0.02
A2, E3	$F_0:D_1:F_2:P_2:D_2$	5.88	0.02
A4	$F_0:P_1:D_1:F_2:P_2:D_2$	5.42	0.02
A2	$P_0:F_0:F_1:P_1:D_1:F_2:P_2:D_2:P_3$	5.33	0.02
N3, A1, A2	$F_0:F_1:F_3:F_4$	5.63	0.03
N1	$P_0:F_0:F_1:P_1:D_1:F_2:P_2:D_2:F_3:P_3:D_3:F_4:P_4:D_4$	5.18	0.03
E2	$D_1:D_2:D_3:D_5$	5.21	0.03
N1, E1, A1, A3	$P_0:F_0:F_1:P_1:D_1:F_2:P_2:D_2:F_3:P_3:D_3:F_4:P_4:D_4:P_5:F_5$	5.11	0.03
E2	$P_0:F_0:F_1:P_1:D_1:F_2:P_2:D_2:F_3:D_3$	4.36	0.04
N1	$F_1:P_1:D_1:F_2:P_2:D_2$	4.37	0.04
N1, A4	$P_0:F_1:D_1:F_2:P_2:D_2$	4.21	0.04

The results from the autocorrelation tests show significant ( $p < 0.05$ ), strong ( $r_k \geq 0.5$ ), and weak ( $0.5 > r_k > 0.2$ ) correlation  $r_k$ , in sequences of both visual behavior and diagnostic decisions for 7 of the 10 subjects reported in this study. We also note the occurrence of stronger sequential dependencies, indicated by large correlation coefficient ( $r_k$ ) in visual behavior in comparison to diagnostic decisions, which show a weaker correlation.

The ANOVA tests conducted determine a number of significant individual and combined effects on current diagnostic decisions. These results, presented in Table 2, suggest that visual search behavior during previous cases can serve as combined predictors of current diagnostic decision. In addition, visual search behavior, breast parenchymal density, and diagnostic decisions during previous cases are combined predictors of current diagnostic decisions. This observation is consistent across readers with a variation in the combination of factors.

The predictive-model-based analysis was performed on a random forest classification model trained using a leave-one-out cross-validation scheme on each individual reader. Our results show an  $f$ -score of 0.54 averaged over all case-pairing datasets, which represents a 64% improvement over random chance (0.33). The notably higher predictability of the current diagnostic decision of new radiology residents (0.66  $f$ -score averaged over all case-pairing datasets) using data from previous visual behavior and diagnostic decisions, compared with advanced radiology residents (0.48  $f$ -score averaged over all case-pairing datasets), and experienced radiologists (0.49  $f$ -score averaged over all case-pairing datasets), indicates that new residents may be more susceptible to sequential context bias. However, data from the current experiments are not sufficient to establish the significance of this observation. The predictive performance results also show, with the exception of two readers (both advanced radiology residents), an increase in

**Table 2** Performance (*f*-score) of behavior-based prediction task for readers by experience: new (N) and advanced (A) radiology resident and expert (E) radiologists.

Reader	$C_n$	$C_{n-1}$	$C_{n-2}$	$C_{n-3}$	$C_{n-4}$	$C_{n-5}$	Random chance
N1	0.39	0.47	0.42	0.50	0.53	0.49	0.24
N2	0.85	0.90	0.87	0.85	0.89	0.87	0.27
N3	0.61	0.67	0.68	0.63	0.60	0.61	0.37
<b>Avg.</b>	<b>0.62</b>	<b>0.68</b>	<b>0.66</b>	<b>0.66</b>	<b>0.67</b>	<b>0.66</b>	<b>0.29</b>
A1	0.30	0.39	0.29	0.31	0.34	0.40	0.29
A2	0.55	0.47	0.50	0.53	0.46	0.50	0.33
A3	0.57	0.59	0.63	0.63	0.68	0.67	0.33
A4	0.51	0.43	0.42	0.44	0.42	0.42	0.38
<b>Avg.</b>	<b>0.48</b>	<b>0.47</b>	<b>0.46</b>	<b>0.48</b>	<b>0.48</b>	<b>0.50</b>	<b>0.33</b>
E1	0.51	0.57	0.53	0.49	0.52	0.48	0.29
E2	0.41	0.43	<b>0.46</b>	0.52	0.45	0.48	0.44
E3	0.38	0.51	0.54	0.43	0.50	0.53	0.28
<b>Avg.</b>	<b>0.43</b>	<b>0.50</b>	<b>0.51</b>	<b>0.48</b>	<b>0.49</b>	<b>0.50</b>	<b>0.34</b>

performance when information from previous cases is introduced into the model [ $M_n < \bigcup_{i=0}^5 (M_n, M_{n-i})$ ].

This study represents a step in the investigation of contextual biases in medical diagnosis through the incorporation of sensor-based measurements of visual behavior in the analysis of biases in mammographic readings. Our findings suggest that new radiology residents are most susceptible to context bias, causing their diagnostic decisions to be predictable given previous behavior and diagnostic decisions. However, our findings also suggest that more advanced radiologists experience some level of context bias.

There are limitations of this study, which need to be considered. The first and foremost is the inherent nature of experiments of this nature. The results from this test indicating contextual bias may be an artifact of the experimental design. We are unable to determine whether bias observed in participating radiologists reflects their individual altered expectations of case prevalence within the study or, if it is applicable, in clinical practice.

Further, interruptions during multiple sessions in all participating readers resulted in the exclusion of ~15% of data collected in this study. Our analysis treats these discontinuities as the beginning of a new experiment by ignoring the first few samples, the number of which is dependent on the number of previous cases being considered. However, it is possible that a number of sessions were interrupted for the purpose of clinical diagnosis as experiments were performed in the clinician's work place. Our analyses do not factor the influence of these intermediary cases on diagnostic decisions after the reader resumes the experiment.

In conclusion, breast parenchymal density, visual behavior, and diagnostic decisions in previous cases may serve as predictors of current diagnostic decisions indicating contextual bias in radiologists' review and diagnosis of mammographic images in

testing situations. We are unable to ascertain the veracity of these claims in clinical practice. However, in the event that these observations apply in clinical practice, a deeper understanding of how these biases occur, and additional factors, which improve predictability of these biases, will be invaluable in improving training methodology and reducing the occurrence of errors in diagnostic imaging.

### Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

### Acknowledgments

This paper has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for the United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan.

### References

1. R. Adolphs, "The neurobiology of social cognition," *Curr. Opin. Neurobiol.* **11**(2), 231–239 (2001).
2. H. A. Simon and K. Kotovsky, "Human acquisition of concepts for sequential patterns," *Psychol. Rev.* **70**(6), 534–546 (1963).
3. J. W. Beckstead, "Modeling sequential context effects in judgment analysis: a time series approach," *Judgment Decis. Making* **3**(7), 570–584 (2008).

4. J. W. Beckstead et al., "Stimulus sequence features influence physicians' response tendencies in radiological image interpretation," *Appl. Cognit. Psychol.* **27**(5), 625–632 (2013).
5. P. D. Frederick et al., "The influence of disease severity of preceding clinical cases on pathologists' medical decision making," *Med. Decis. Making* **37**(1), 91–100 (2016).
6. W. J. Matthews and N. Stewart, "Psychophysics and the judgment of price: judging complex objects on a non-physical dimension elicits sequential effects like those in perceptual tasks," *Judgment Decis. Making* **4**(1), 64 (2009).
7. J. W. Beckstead et al., "Sequential dependencies in categorical judgments of radiographic images," *Adv. Health Sci. Educ.* **22**(1), 197–207 (2017).
8. R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology* **184**(3), 613–617 (1992).
9. P. A. Carney et al., "Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography," *Ann. Intern. Med.* **138**(3), 168–175 (2003).
10. F. Alamudun et al., "Fractal analysis of visual search activity for mass detection during mammographic screening," *Med. Phys.* **44**(3), 832–846 (2017).
11. F. Alamudun et al., "Shapelet analysis of pupil dilation for modeling visuo-cognitive behavior in screening mammography," *Proc. SPIE* **9787**, 97870M (2016).
12. F. T. Alamudun et al., "Geometry and gesture-based features from saccadic eye-movement as a biometric in radiology," in *Int. Conf. on Augmented Cognition*, pp. 123–138, Springer International Publishing, Cham (2017).
13. C. Nodine and H. Kundel, "The cognitive side of visual search in radiology," in *Eye Movements from Physiology to Cognition*, pp. 573–582, Elsevier (1987).
14. G. Tourassi et al., "Investigating the link between radiologists' gaze, diagnostic decision, and image content," *J. Am. Med. Inf. Assoc.* **20**(6), 1067–1075 (2013).
15. J. J. Donald and S. A. Barnard, "Common patterns in 558 diagnostic radiology errors," *J. Med. Imaging Radiat. Oncol.* **56**(2), 173–178 (2012).
16. A. Van der Gijp et al., "How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology," *Adv. Health Sci. Educ.* **22**(3), 765–787 (2017).
17. C. S. Lee et al., "Cognitive and system factors contributing to diagnostic errors in radiology," *Am. J. Roentgenol.* **201**(3), 611–617 (2013).
18. L. P. Busby, J. L. Courtier, and C. M. Glastonbury, "Bias in radiology: the how and why of misses and misinterpretations," *RadioGraphics* **38**(1), 236–247 (2017).
19. G. A. Taylor, "Perceptual errors in pediatric radiology," *Diagnosis* **4**(3), 141–147 (2017).
20. H. L. Kundel and P. S. La Follette Jr., "Visual search patterns and experience with radiological images," *Radiology* **103**(3), 523–528 (1972).
21. G. Jager, J. Fütterer, and M. Rutten, "Cognitive errors in radiology: 'Thinking fast and slow'," in *European Congress of Radiology*, Vienna (2014).
22. D. Laming, "Screening cervical smears," *Br. J. Psychol.* **86**(4), 507–516 (1995).
23. N. Stewart, G. D. Brown, and N. Chater, "Absolute identification by relative judgment," *Psychol. Rev.* **112**(4), 881–911 (2005).
24. L. M. Ward and G. Lockhead, "Sequential effects and memory in category judgments," *J. Exp. Psychol.* **84**(1), 27–34 (1970).
25. L. T. DeCarlo, "Intertrial interval and sequential effects in magnitude scaling," *J. Exp. Psychol.: Hum. Percept. Perform.* **18**(4), 1080 (1992).
26. M. Heath et al., "Current status of the digital database for screening mammography," in *Digital Mammography: Nijmegen, 1998*, N. Karssemeijer et al., Eds., pp. 457–460, Springer, Netherlands, Dordrecht (1998).
27. American College of Radiology, *Breast Imaging Reporting and Data System (BI-RADS)*, American College of Radiology, Reston, Virginia (1998).
28. Applied Science Laboratories, "ASL EyeTrac 6: EyeNal Analysis Software," Applied Science Laboratories, Bedford, Massachusetts (2007).
29. F. Hausdorff, "Dimension und äußeres Maß," *Math. Ann.* **79**(1), 157–179 (1918).

**Folami Alamudun** is a postdoctoral research associate in the Health Data Sciences Institute and the Biomedical Sciences, Engineering and Computing Group at the Oak Ridge National Laboratory (ORNL). He holds his BS degree in electrical engineering from the University of Texas at El Paso and his PhD in computer science from Texas A&M University. His current research interests include deep learning, intelligent text understanding, and virtual reality applications in psychotherapy and education. He is a member of SPIE.

**Paige Paulus** recently graduated from the University of Tennessee, Knoxville, Tennessee, USA, with a degree in biomedical engineering. Her research interests include prosthetics and biomechanics. She hopes to pursue a career doing research and clinical work as a physical therapist in these areas.

**Hong-Jun Yoon** is a staff scientist of the Health Data Sciences Institute and the Biomedical Science and Engineering Center at the ORNL. He received his MS degree in electrical engineering from the University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA, in 1999, and his PhD in electrical engineering from the University of Pittsburgh, Pittsburgh, Pennsylvania, USA, in 2011. From 2003 to 2012, he was a research staff at the University of Pittsburgh Medical Center, participated in eye tracking and gaze perception analysis. He won an R&D 100 award in 2014 and was selected as finalist an R&D 100 award in 2015.

**Georgia Tourassi** is a director of the Health Data Sciences Institute at ORNL. Her research interests include medical imaging, biomedical informatics, clinical decision support systems, and data-driven biomedical discovery. She is a fellow of the American Institute of Medical and Biological Engineering, the American Association of Medical Physicists, and SPIE. She holds a BS in physics from University of Thessaloniki, Greece, and a PhD in biomedical engineering from Duke University.