Award Number:
W81XWH-11-1-0755

TITLE:
Adaptive Computer-Assisted Mammography Training for Improved Breast Cancer Screening

PRINCIPAL INVESTIGATOR:
Maciej Mazurowski

CONTRACTING ORGANIZATION: Duke University
Durham, NC 27705

REPORT DATE: March 2015

TYPE OF REPORT: Final Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
March 2015	Final Summary	15Sep2011 - 14Dec2014

4. TITLE AND SUBTITLE

Adaptive Computer-Assisted Mammography Training for Improved Breast Cancer Screening

5a. CONTRACT NUMBER
W81XWH-11-1-0755

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Maciej Mazurowski

email: maciej.mazurowski@duke.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University
2200 W Main St Ste 710
Durham, NC 27705

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this project, we propose to research the methodology for constructing adaptive computer-aided education systems for mammography. Improved mammography education could lead to improved benefit of mammography to breast cancer care and in turn to decreased mortality from the disease. The project includes: Observer studies to collect reading data from radiology trainees; Extraction of image features (human- and computer- based); Statistical modeling of the reader data to discover patterns in their error making; Development of methodology for adaptive training that utilizes the constructed models. The proposed adaptive system could improve education in mammography. This may in turn result in improved benefit of mammography in breast cancer detection and lower mortality associated the disease.

15. SUBJECT TERMS
Mammography, radiology, education, user modeling, resident, graduate medical education

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					USAMRMC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	UU	32	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

INTRODUCTION: In this project, we proposed to research the methodology for constructing adaptive computer-aided education systems for mammography. The project includes: Observer studies to collect reading data from radiology trainees; Extraction of image features (human- and computer- based) from mammograms; Statistical modeling of trainees reading data to discover patterns in their error making; Development of methodology for adaptive training that utilizes the constructed models. The proposed adaptive system could improve education in mammography. This may in turn result in improved benefit of mammography in breast cancer detection and lower mortality associated the disease.

BODY:

Overall progress:

Specific aim	Expected	Actual
1.1 Prepare the database of screening mammograms (year 1, months 1-6)	Completed	Completed
1.2 Obtain the approval for the human subject (observer) studies in tasks 1 and 3.	Completed	Completed
1.3 Perform an observer study in which residents will search for masses and architectural distortions (year 1, months 7-9). We expect 20 human subjects (observers) to participate in this study.	Completed	Completed
1.4 Utilize the user data collected in the observer study to develop machine learning-based individual user models (year 1, month 10 – year 2, month 6)	Completed	Completed
2.1 Develop a computer tool that presents the user profiles to the users and provides them with guidance regarding their performance	Not necessary based on change in Aim 3	Not necessary based on change in Aim 3
2.2 Develop an algorithm that generates an individualized training protocol based on the user model captured by the system developed in aim 1	Modified version completed based on change in Aim 3	Modified version completed based on change in Aim 3
3 Test if a trainee is presented with cases for which he/she made an error, will they benefit more than by being presented with randomly selected cases	Completed	Completed

The detailed description of progress regarding each specific aim follows.

1.1 Prepare the database of screening mammograms (year 1, months 1-6)

STATUS: completed in the 2011-2012 period in respect to the data for the first reader study

1.2 Obtain the approval for the human subject (observer) studies in tasks 1 and 3.

STATUS: completed in the 2011-2012 period

1.3 Perform an observer study in which residents will search for masses and architectural distortions (year 1, months 7-9). We expect 20 human subjects (observers) to participate in this study.

STATUS: completed in the 2011-2012 period

1.4 Utilize the user data collected in the observer study to develop machine learning-based individual user models (year 1, month 10 – year 2, month 6)

STATUS: main part completed in the 2012-2013 period. Some experiments are still in progress

2.2 Develop an algorithm that generates an individualized training protocol based on the user model captured by the system developed in aim 1

STATUS: completed with some modifications. We developed algorithms that are able to select cases that are predicted to be the most difficult for each trainee.

2.3 Test if a trainee is presented with cases for which he/she made an error, will they benefit more than by being presented with randomly selected cases

STATUS: completed

DETAILS:

The work in the 2013-2014 period focused on three aspects of the work:

- Further collection of mammograms (aim 1)
- Further development of machine learning algorithms for individual user models (aim 1)
- Reader study to test whether individually adapted educational material shows improved educational utility over randomly selected material (aim 2 & 3)

Please note that the PI has contacted Dr. Kristy Lidie on 4/22/2014 regarding a change in Aim 3 which was approved. Specifically, we decided that instead testing the specific educational system, we will test a more basic hypothesis: "if a trainee is presented with cases for which he/she made an error, will they benefit more than by being presented with randomly selected cases?". This hypothesis is a foundation on which we base our research and is very important to test. Following this change, we also adjusted the work in Aim 2 to fit the hypothesis tested in Aim 3.

The specific work accomplished in the 2013-2014 included:

- Development of new computer vision algorithms for automatic analysis of mammograms in order to predict whether a resident will or will not make a false positive error. We have submitted a journal manuscript on this concept in the 2013-2014 period.
- Revision of the 3 journal manuscripts submitted in the 2012-2013 period that led to eventual publication all 3 of them in the 2013-2014 period.

- Data collection for the final reader study. Over 400 mammographic cases were collected along with original radiology interpretations data.
- Development of a graphical user interface for the final reader study that in addition to testing, presents the reader with educational material.
- First part of the reader study (reader study still in progress): 11 subjects recruited

Below, we briefly present the design and results for the study that examined the relationship between difficulty and error:

Introduction: We developed an algorithm for prediction of false positive error making among radiology trainees. Identifying difficult locations for the trainees could allow for focusing their training and result in improvement in performance.

Methods: The proposed algorithm identifies locations that are associated with high likelihood of a trainee making a false positive error. Those locations can be identified on images previously unseen by the trainees. The algorithm first uses a Difference of Gaussian (DOG) filter to identify potential suspicious locations. Then, a random forest classifier identifies the locations with the highest probability of occurrence of a false positive error using 133 features extracted from each location identified by the DOG filter. The random forest is developed individually for each trainee, using previous locations pointed out by the trainee.

Results: The accuracy of our algorithm in identifying locations associated with false positive errors was notably higher than of an algorithm that identifies such locations randomly.

Specifically, the accuracy of our algorithm was 40% when only 1 location was selected by the algorithm for all cases for each trainee and 12% when 10 locations were selected. The accuracies for random location selection was 0% for both of these two scenarios.

Figure 3 show the performance of our algorithm and a random algorithm assessed using free-response receiver operating characteristics (from the paper below):

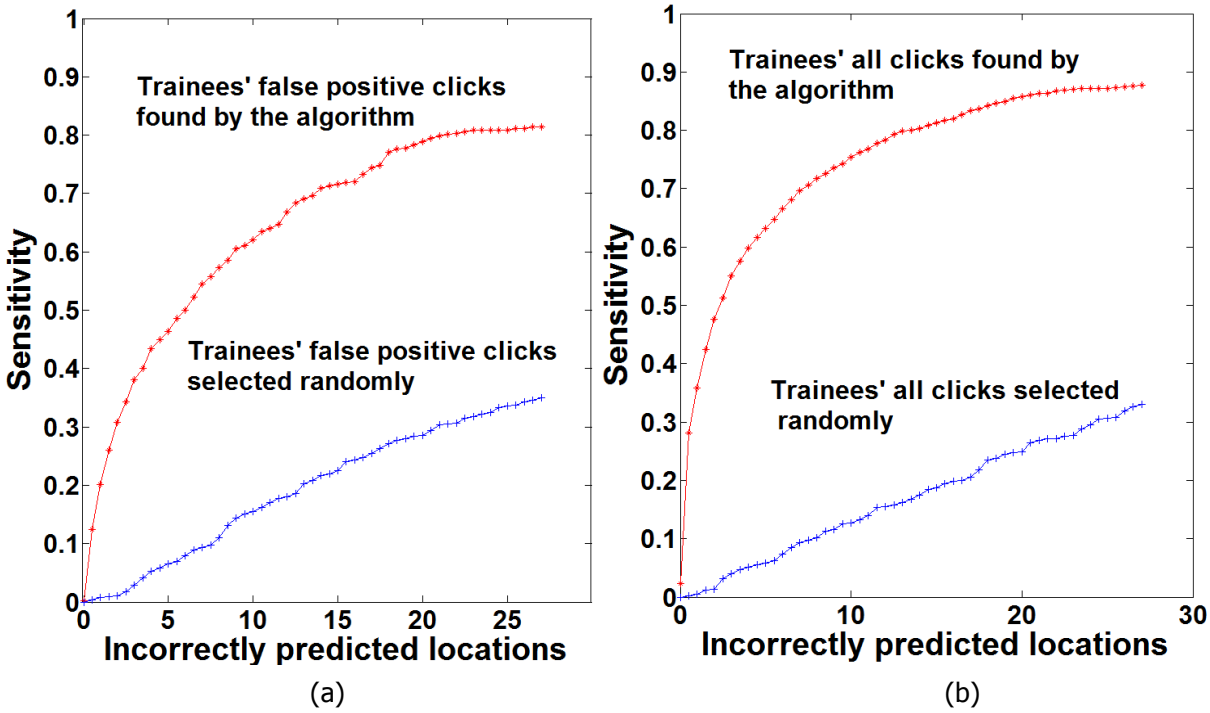


Fig. 3 FROC curves for trainees' false positive clicks and all clicks prediction. (a) shows the FROC curve

Figure 4 shows locations found by the algorithm for different trainees (from the paper attached below):

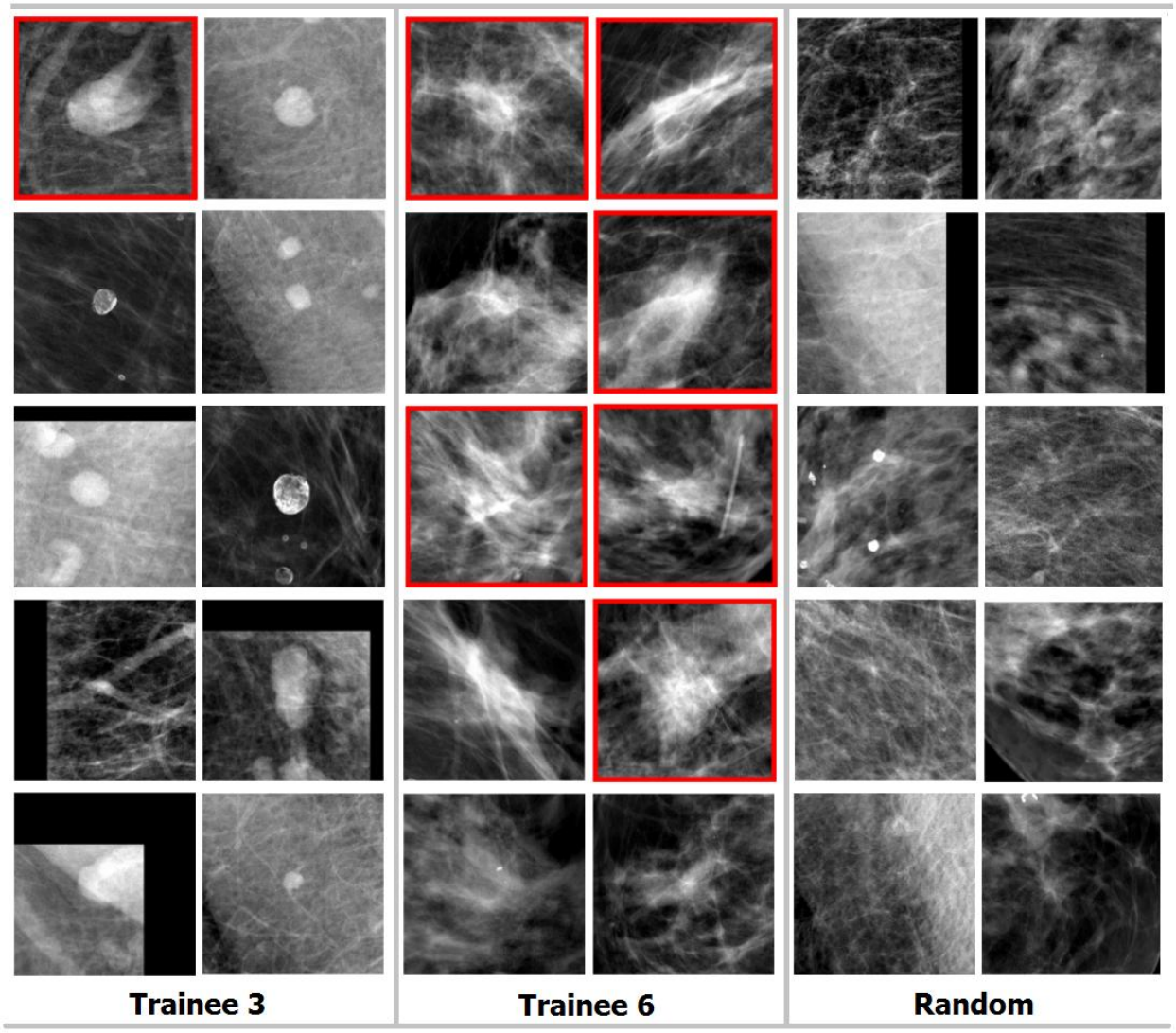


Fig. 4 The images of top 10 predicted false positives. The 10 images in the first column and the 10 images in the second column are top 10 false positives predicted by the algorithm for trainee 3 and trainee 6, respectively. The 10 images in the third column are the top 10 false positives predicted based

DIFFICULTIES:

No major difficulties have been encountered as of this point, however, we decided that instead of testing the hypothesis that was initially proposed in Aim 3, we will test a hypothesis that is more basic and potentially more generally useful in this research. This was discussed with the Scientific Officer at the DOD.

KEY RESEARCH ACCOMPLISHMENTS:

2011-2012

- Secured IRB approval for the study
- Retrospectively collected a set of mammograms for the study
- Conducted an observer study
- Conducted preliminary analysis of the observer study results
- Initiated development of a controlled dictionary for mammography education

2012-2013

- Evaluated the relationship between the concepts of self-assessed case difficulty, expert assessment of difficulty, and actual resident error (this analysis was started in 2011-2012 period).
- Implemented computer vision features for analysis of mammograms, which could be used for prediction of case difficulty/error and used them for prediction of false negative errors among radiology trainees
- We collected assessments of image features from experienced radiologists and established whether such features can be used for prediction of false negative errors among radiology trainees

2013-2014

- Development of new computer vision algorithms for automatic analysis of mammograms in order to predict whether a resident will or will not make a false positive error. We have submitted a journal manuscript on this concept in the 2013-2014 period.
- Revision of the 3 journal manuscripts submitted in the 2012-2013 period that led to eventual publication all 3 of them in the 2013-2014 period.
- Data collection for the final reader study. Over 400 mammographic cases were collected along with original radiology interpretations data.
- Development of a graphical user interface for the final reader study that in addition to testing, presents the reader with educational material.
- First part of the reader study (reader study still in progress): 11 subjects recruited

2014-2015

- Analysis of the final experiment. The analysis to date did not show conclusive results, likely due to fatigue of the trainees during this very long experiment (>400 mammographic cases).

REPORTABLE OUTCOMES:

2011-2012

- Collected a database of digital mammograms
- An extended abstract entitled "Difficulty of mammographic cases in the context of resident training: preliminary experimental data" submitted and accepted to SPIE Medical Imaging 2013 conference.

2012-2013

- Three manuscripts were submitted to journals and are currently in various review stages:
    - L. Grimm, S. V. Ghate, S.Yoon, C. M. Kuzmiak, C. Kim, **M. A. Mazurowski** (2013). 'Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features', in revision for Medical Physics, November 2013.
    - L. Grimm, C. M. Kuzmiak, S. V. Ghate, S. Yoon, **M. A. Mazurowski** (2013). 'Mammography difficulty and error making patterns in the context of resident training', submitted to Academic Radiology, October 2013
    - J. Zhang, J. Y. Lo, C. M. Kuzmiak, S. V. Ghate, S. C. Yoon, **M. A. Mazurowski** (2013), 'Using computer-extracted image features for modeling of error-making patterns in detection of mammographic masses among radiology residents', submitted to Physics in Medicine and Biology, October 2013.
- Conference proceedings paper (/extended abstract) accepted for oral presentation at SPIE Medical Imaging:
    - **M. A. Mazurowski**, J. Zhang, J. Y. Lo, C. M. Kuzmiak, S. V. Ghate, S. Yoon (2014). 'Modeling resident error-making patterns in detection of mammographic masses using computer-extracted image features: preliminary experiments', SPIE Medical Imaging 2014, in press
- Oral presentation given at SPIE Medical Imaging 2013:
    - **M. A. Mazurowski**, 'Difficulty of mammographic cases in the context of resident training: preliminary experimental data', SPIE Medical Imaging, 2013
- Conference proceedings paper published:
    - **M. A. Mazurowski**, 'Difficulty of mammographic cases in the context of resident training: preliminary experimental data', SPIE Medical Imaging, 2013

2013-2014

- Three manuscripts were published:
    - L. Grimm, S. V. Ghate, S.Yoon, C. M. Kuzmiak, C. Kim, **M. A. Mazurowski** (2014). 'Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features', Medical Physics 41, 2014.
    - L. Grimm, C. M. Kuzmiak, S. V. Ghate, S. Yoon, **M. A. Mazurowski** (2014). 'Radiology Resident Mammography Training: Interpretation Difficulty and Error-making Patterns', Academic Radiology 21, 2014

    - J. Zhang, J. Y. Lo, C. M. Kuzmiak, S. V. Ghate, S. C. Yoon, **M. A. Mazurowski** (2014), 'Using computer-extracted image features for modeling of error-making patterns in detection of mammographic masses among radiology residents', Medical Physics 41, 2014.
- One conference proceedings paper was published:
    - **M. A. Mazurowski**, J. Zhang, J. Y. Lo, C. M. Kuzmiak, S. V. Ghate, S. Yoon (2014). 'Modeling resident error-making patterns in detection of mammographic masses using computer-extracted image features: preliminary experiments', SPIE Medical Imaging 2014
- One oral presentation given by Dr. Mazurowski:
    - "Modeling resident error-making patterns in detection of mammographic masses using computer-extracted image features: preliminary experiments", SPIE Medical Imaging 2014
- One journal paper submitted:
    - J. Zhang, J. I. Silber, **M. A. Mazurowski**, "Modeling false positive error making patterns in radiology trainees for improved mammography education", Journal of Biomedical Informatics, in revision (2014)
- Grants:
    - This work inspired an application for a preliminary study on a related topic of education in digital breast tomosynthesis which is now funded by the Radiological Association of North America (PI: Mazurowski)

2014-2015 (no cost extension)

- One journal manuscript published
    - J. Zhang, J. I. Silber, **M. A. Mazurowski**, "Modeling false positive error making patterns in radiology trainees for improved mammography education", Journal 0of Biomedical Informatics, in revision (2014)

CONCLUSION:

This project is considered very successful. It resulted in a significant improvement of our understanding of the error making process in radiology. The measurable outcomes of the project include 4 manuscript published in high impact factor journals and 2 conference proceedings papers accompanied with 2 oral presentations on the topic at international conferences. This project also inspired the next step of this research (education in digital breast tomosynthesis) and was recently funded by Radiological Society of North America Research and Education Foundation.

APPENDICES: The manuscript submitted in the 2013-2014 period (in its state at the time, after revisions) is attached below. This manuscript is now published in the Journal of Biomedical Informatics (2015).

Modeling false positive error making patterns in radiology trainees for improved mammography education

Jing Zhang[1], James I. Silber[2], Maciej A. Mazurowski[1,3,4]

[1] Department of Radiology, Duke University School of Medicine, Durham, NC

[2] Department of Biomedical Engineering, Duke University Pratt School of Engineering, Durham, NC

[3] Duke Cancer Institute

[4] Duke Medical Physics Program

Corresponding Author:

Jing Zhang

Phone: 1-919-6815012

Email: jing.zhang2@duke.edu

Postal address: 2424 Erwin Road, Suite 302, Durham, NC, 27707

Abstract:

Introduction While mammography notably contributes to earlier detection of breast cancer, it has its limitations including a large number of false positive exams. Improved radiology education could potentially contribute to alleviating this issue. Toward this goal, in this paper we propose an algorithm for modeling of false positive error making among radiology trainees. Identifying troublesome locations for the trainees could focus their training and in turn improve their performance.

Methods The algorithm proposed in this paper predicts locations that are likely to result in a false positive error for each trainee based on the previous annotations made by the trainee. The algorithm consists of three steps. First, the suspicious false positive locations are identified in mammograms by Difference of Gaussian filter and suspicious regions are segmented by computer vision-based segmentation algorithms. Second, 133 features are extracted for each suspicious region to describe its distinctive characteristics. Third, a random forest classifier is applied to predict the likelihood of the trainee making a false positive error using the extracted features. The random forest classifier is trained using previous annotations made by the trainee. We evaluated the algorithm using data from a reader study in which 3 experts and 10 trainees interpreted 100 mammographic cases.

Results The algorithm was able to identify locations where the trainee will commit a false positive error with accuracy higher than an algorithm that selects such locations randomly. Specifically, our algorithm found false positive locations with 40% accuracy when only 1 location was selected for all cases for each trainee and 12% accuracy when 10 locations were selected. The accuracies for randomly identified locations were both 0% for these two scenarios.

Conclusions In this first study on the topic, we were able to build computer models that were able to find locations for which a trainee will make a false positive error in images that were not previously seen

by the trainee. Presenting the trainees with such locations rather than randomly selected ones may improve their educational outcomes.

Keywords Breast Cancer Radiology education Mammography Tumor segmentation Feature Extraction False Positive Predictive Model Random forest

1. Introduction

Mammography is the most widely used screening technique for breast cancer early detection, which plays an important role in reducing the mortality of breast cancer. However, interpretation of mammograms is a very challenging task due to overlapping tissue that might both obscure signs of cancer (false negative errors) as well as create patterns that resemble true abnormalities and unnecessarily alert a radiologist (false positive errors) (Baker and Lo 2011).

Our group has been working on the development of an adaptive computer-aided education system for mammography education. Specifically, in (Mazurowski, Baker et al. 2010), we proposed a general framework for such a system and demonstrated that image features can be used to predict errors made by a trainee. In (Mazurowski, Barnhart et al. 2012), we presented models for prediction of errors in assignment of BI-RADS features of masses and images. In (Mazurowski and Tourassi 2011), we investigated the use of collaborative filtering algorithms to model resident errors in mammography. Other work on the adaptive mammography education is limited, some related studies are available. Sun *et al.* (Sun, Taylor et al. 2008, Sun, Taylor et al. 2008) presented initial studies on developing an ontology related educational training system based on differences between radiologists. The studies by Mello-Thoms *et al.* (Mello-Thoms, Dunn et al. 2002), Tourassi *et al.* (Tourassi, Voisin et al. 2013), Voisin *et al.* (Voisin, Pinto et al. 2013) investigate visual attention and spatial frequency representations, human perception and cognition, and eye gaze tracking to study error making in mammography. Some work in computer-aided detection is also relevant to our study in terms of the computer vision methods used, such as the studies presented by Masotti *et al.* (Masotti, Lanconelli et al. 2009), Wei *et al.* (Wei, Chan et al. 1997), and Mudigonda *et al.* (Mudigonda, Rangayyan et al. 2001).

In this paper we focus on a topic largely unexplored in the context of radiology education: false positive error making. Specifically, the task that we approach is to automatically find locations that will cause a trainee to make a false positive error. For this purpose, we propose an algorithm that identifies challenging locations using computer vision algorithms and machine learning models. The models are constructed individually for each trainee based on their prior interpretations to capture their individual error making patterns.

To our knowledge, this is the first study in which future false positive locations are predicted. It differs from our previous studies in which we focused on false negative errors (Grimm, Ghate et al. 2014), errors in distinguishing benign and malignant masses (Mazurowski, Baker et al. 2010), and errors in assessment of BI-RADS features (Mazurowski, Barnhart et al. 2012). Predicting false positive locations is a difficult task as it requires analysis of the entire image and finding those locations that might cause difficulty to the trainee while dismissing all the locations that will not. While our experiments confirm the high difficulty of the task, they also show promise of our approach. One practical application of our approach is to identify locations that would result in false positive errors for each trainee, so that they can focus their training on such locations potentially improving their training.

2. Reader study and the definition of false positive errors

To validate our algorithm for predicting false positive errors, we used data from a reader study in which 10 radiology trainees along with 3 expert radiologists interpreted 100 mammographic cases independently. Among the 10 trainees, 7 were radiology residents with at least four weeks of formal breast imaging training and 3 were novices (2 medical imaging researchers and 1 medical student) with no formal training. We included the three novices to simulate radiology residents at the very beginning of their residency program. The three expert radiologists were all fellowship trained in breast imaging with 7-14 years of experience. The experts and the trainees were not aware of patients' age and medical history. The 100 mammographic cases are balanced with 50 cases originally deemed as normal and 50 abnormal cases. Each case contained 4 standard mammographic views: left craniocaudal (LCC), right craniocaudal (RCC), left mediolateral oblique (LMLO), and right mediolateral oblique (RMLO). All participants were asked to identify actionable abnormalities by clicking on them. We asked the

participants to ignore microcalcifications as the focus of our study was on masses. Institutional Review Board approval was secured for this study.

We used the marks provided by the three experts to find the actual actionable masses. Specifically, if a region contained at least two out of three experts' marks and the distance between two marks was smaller than a predefined threshold T_d, we considered this region to be associated with an actionable mass. The centers of actual actionable masses were determined as the centroids of the expert annotations. Consequently, if the distance between a trainee's mark and its nearest actionable mass center is bigger than T_d, this mark is defined as a false positive error. Otherwise, it is defined as a true positive. Because the average radius of the breast masses is 9mm (Timp, Karssemeijer et al. 2003) and the pixel spacing of the images used in the reader study was 0.0941 mm, the threshold was set to T_d = 9mm/0.0941mm=96 pixels in our study.

3. The algorithm for prediction of false positive locations

3.1 Overview

In this paper, we propose an algorithm that searches through an entire mammographic image to find locations where the trainee made false positive errors. The proposed algorithm accepts an entire mammographic image as the input and returns locations that are more likely to be associated with a false positive error as the output. The algorithm is composed of 3 steps:

Step 1: The Difference of Gaussian (DoG) filter (Babaud, Witkin et al. 1986) is adopted to identify suspicious false positive locations of an image, and then rubber band and region growing methods are used to segment suspicious false positive regions using local maximum points extracted from the DoG filter response map;

Step 2: Features are extracted to describe the properties of each region and its context; and

Step 3: A classifier is applied to predict the likelihood of a predicted location being a false positive error made by the trainee using the extracted features.

The flowchart of the proposed algorithm is shown in Fig. 1. The three steps of the algorithm are described in the subsections below.

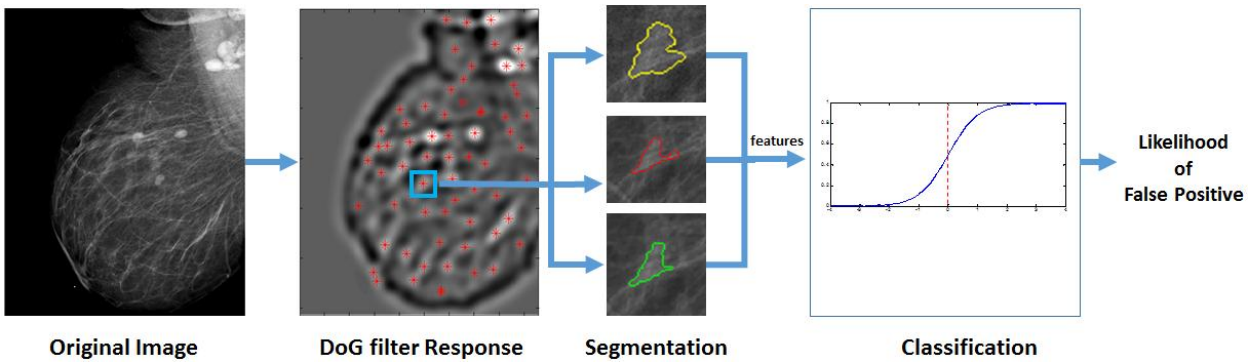


Fig. 1 Flowchart of the proposed false positive prediction algorithm

3.2 Step 1: Identifying suspicious locations

Difference of Gaussian (DoG) filter, which has been widely used for breast mass detection (Polakowski, Cournoyer et al. 1997, Catarious Jr, Baydush et al. 2006), is adopted in our study as the first step to identify the suspicious locations where the trainees may make false positive errors (i.e., click on the location). After calculating the DoG filter response for the entire image, we extract local maximum points from the DoG filter response map and consider these locations suspicious. Then, by using the identified suspicious locations as reference points, three segmentation methods (dynamic programming-based rubber band, region growing with adaptive threshold, and region growing with fixed threshold) are applied to segment the abnormality or the abnormality resembling region. These segmentations will be later used to determine features of locations. The segmentation algorithms used are described below.

The Dynamic programming-based rubber band method (Timp and Karssemeijer 2004) can transform a round image region to a rectangular region in a polar coordinate system. Gradient, size, and intensity information extracted from the image in the polar system are combined to form a cost matrix. The boundary of the region is the path that has the lowest cost in the cost matrix determined by dynamic programming. The Region growing method (Adams and Bischof 1994) segments a region by computing the similarity between the given seed region and its neighboring pixels iteratively. If the similarity is smaller than a predefined threshold, the seed region is grown by including its neighboring pixels. The

method stops when no new pixels can be included. Two seed region growing strategies of region growing method were adopted in this study: one with a fixed seed region and the other with an adaptive seed region that is updated at each iteration. The purpose of using three different segmentation methods (the two variations of the region growing algorithm was treated as two different segmentation algorithms) is to be able to compute features indicating segmentation difficulty by comparing the three segmentation results.

3.3 Step 2: Extracting features from the identified suspicious locations

Based on the segmented suspicious regions, our algorithm extracts 133 features that capture distinctive characteristics of the indicated regions and its surrounding, which we suspect relate to false positive error making. The extracted features can be grouped into two categories: region-based features and context-based features. All features (F1 to F133) are listed in Table 1. 'region growing I' and 'region growing II' represent the region growing method with the fixed seed region and the region growing method with adaptive seed region respectively. We group the image features into two different categories and briefly describe some of the features that need more explanation as follows:

(1) Region-based features: The features in this category are computed to capture the characteristics of an identified suspicious region. For example, the normal region with mass-like appearance (e.g., high intensity, sharp edge, and round shape) has a higher probability to be a false positive error made by trainees.

F4 indicates the region intensity normalized by fat tissue intensity and dense tissue intensity. Based on the rubber band method, F10 to F13 are proposed to explore the intensity changes of the suspicious region along its region boundary and inner area in the polar coordinate system, which can indicate the strength of the region boundary and whether the identified suspicious region connects with fatty tissue. F16 and F17 indicate the location of a mass based on the normalized distance transformed breast region. F18-F44 are Haralick texture features (Haralick, Shanmugam et al. 1973) computed based on a gray level

co-occurrence matrix (GLCM), including correlation, contrast, energy, entropy, etc. F44-F54 and F55-F84 are Gray scale invariant ranklet features and spatial gray level dependence (SGLD) matrix based local features are proposed for false positive reduction in mammography in Masotti *et al.* (Masotti, Lanconelli et al. 2009) and Wei *et al.* (Wei, Chan et al. 1997). F85-F87 are boundary ribbon based coherence ratio, entropy of orientation estimates, and variance of coherence-weighted angular estimates proposed for false positive analysis in mammograms in Mudigonda *et al.* (Mudigonda, Rangayyan et al. 2001).

Table 1. Extracted Features

Feature category	Feature description	
Mass-based	F1	Region area
	F2	Intensity of region centroid
	F3	Normalized intensity of region
	F4	Region circularity
	F5	Region solidity
	F6	Standard deviation of the intensities of region
	F7	Region rubber band cost
	F8	Region boundary gradient
	F9-F12	Directional intensity changes along region boundary band, between boundary and its centroid, inner area, and between its centroid and inner area
	F13	Region area (rubber band) / region area (Region growing I)
	F14	Region area (rubber band) / region area (Region growing II)
	F15-F16	Region location X and Y
	F17-F43	GLCM based Haralick texture features
	F44-F54	Gray scale invariant ranklet features
	F55-F84	SGLD based local features
	F85-F87	Ribbons based coherence ratio, entropy of Orientation Estimates, and variance of coherence-weighted angular estimates
Context-based	F88	Breast density
	F89	Dense tissue solidity
	F90	Dense tissue Euler number
	F91	Number of suspicious regions
	F92	Mean intensity of neighboring suspicious regions
	F93	Number of neighboring suspicious region
	F94	Number of suspicious regions based on intensity and size similarity
	F95	Maximum similarity of suspicious regions based on intensity and size
	F96	Average similarity of suspicious regions based on intensity and size
	F97	Number of suspicious regions based on intensity, size, and shape similarity
	F98	Maximum similarity to suspicious regions based on intensity, size, and shape
	F99	Average similarity to suspicious regions based on intensity, size, and shape
	F100	Number of local suspicious regions
	F101	Local Suspicious region area ratio

F102	Local breast density
F103	Mean intensity of region / Mean intensity of neighboring area
F104	Intensity-based area under the ROC curve
F105	Is region in pectoral muscle
F106	DoG filter response
F107-F133	GLCM based Haralick texture features extracted from context

(2) Context-based features: The features in this category are computed to capture the relationships between the suspicious region and its context. Examples include if the intensity of a suspicious region is higher than its neighboring area or not; and whether a suspicious region is similar to many other regions in the breast or not.

F94 to F96 measure the similarities among suspicious regions and other neighboring suspicious regions by computing the average similarity value and the biggest similarity value using size and intensity, and the number of regions whose similarities are bigger than a threshold. F97 to F99 measure the similarities using neighboring suspicious regions as well, but one more feature, region solidity, is used for similarity computation. F100 and F101 indicate the number of suspicious regions inside a surrounding circular area and the area ratio between the suspicious region and all suspicious regions inside this surrounding area. By using the intensity of pixels as predictors and assigning pixels inside and outside the suspicious region two different classes, F104 computes the area under curve (AUC) of a receiver operating characteristic (ROC) curve. F106 is the DoG filter response and F107-F133 are Haralick texture features extracted from the surrounding context of a suspicious region.

3.4 Step 3: Classification

The task of our classifier is to determine, for each location found by the DoG filter in step 1 of our algorithm, whether the trainee will make a false positive error for the location (i.e., click on the location) or not. Using a classifier will allow for elimination of some incorrect predictions. The input to the classifier is the set of features extracted from the location and its surrounding in Step 2 of our algorithm described in Section 3.3. The classifier was trained individually for each trainee. For the classifier training, we used the previous annotation data of the trainee (i.e., their clicks). For the classification purpose, we defined positive and negative instances (i.e., examples, or samples) in the following way:

Positive instances: (1) the locations that were indicated by the DoG filter and predicting the false positive errors marked by a trainee correctly (i.e., they were within 96 pixels from an actual trainee's false positive mark); and (2) the locations that were marked by the trainee but were not indicated by the DoG filter.

Negative instances: locations that were indicated by the DoG filter but did not predict false positive errors made by the trainee correctly.

A Random forest classifier (Breiman 2001) was used in our study. Random forest is a popular machine learning method that utilizes an ensemble of decision trees for classification. The Matlab function "TreeBagger" was used to create the random forest. 500 trees were used in the forest and 5 variables were selected randomly for each decision tree.

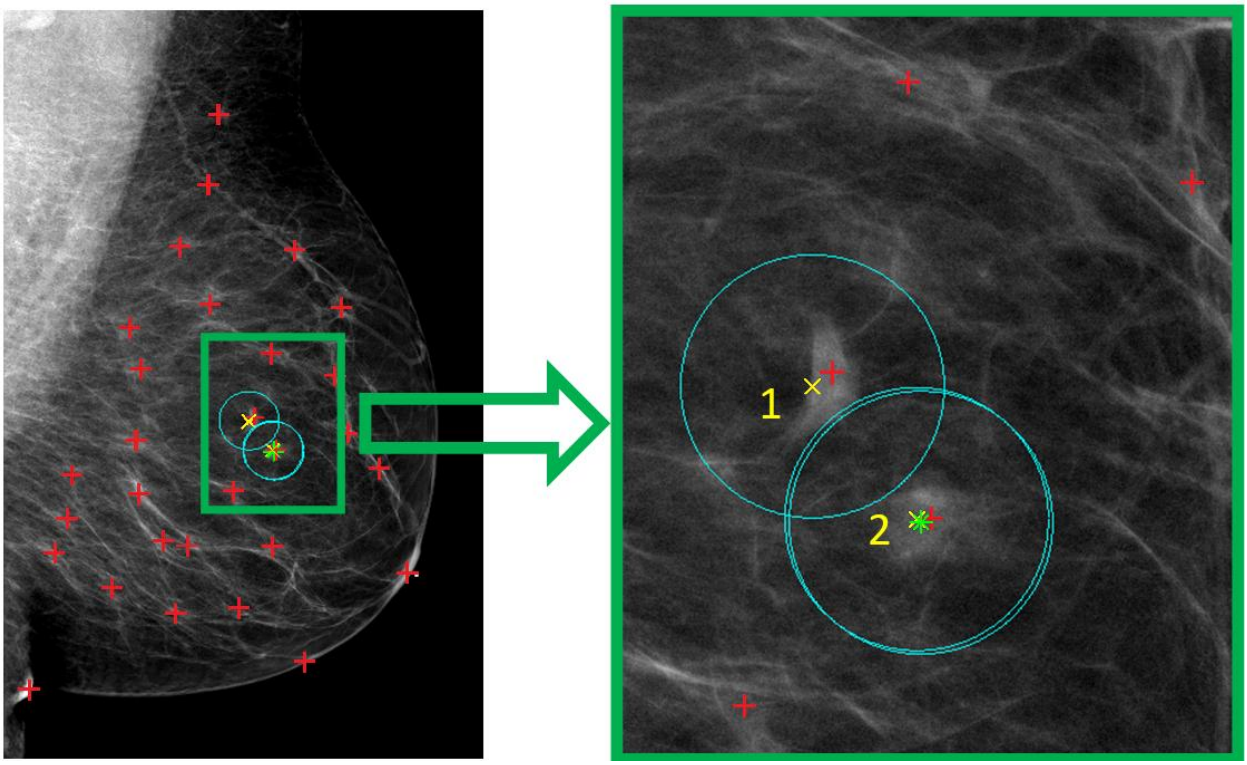
4 Evaluation of the algorithm

4.1 Determining whether the algorithm's prediction is correct

Note that as we focus on prediction of false positive errors in this study, in our evaluation we excluded the locations found by the algorithm that indicated true positive marks made by a trainee. Such locations cannot be considered correct predictions of false positive location but we believe that they should also not be considered incorrect predictions as they still correctly predict the trainees' click and therefore demonstrate that the trainee's behavior was modeled correctly. When counting the number of correctly and incorrectly predicted false positives, such locations will be simply excluded from the analysis. Hence, the following description does not apply to such locations.

To determine whether a location indicated by our algorithm correctly predicts a false positive error, we used the following criterion: if the distance between the location predicted by the algorithm and any trainee's false positive location is smaller than the threshold T_{d}, (96 pixels) we consider this predicted location to be a correct prediction. Otherwise, we consider it to be an incorrect prediction. Fig. 2-a

illustrates an example of a mammographic image and Fig. 2-b is the close up of the green box in Fig.2-a. In this image, one mass was marked by the experts (green star) and two locations were marked by the trainee (yellow cross). One is a true positive (yellow cross #2) and the other is a false positive error (yellow cross #1). The suspicious locations found by the first step of our algorithm are shown as red plus signs. The cyan circles indicate the round areas centered at yellow crosses and green stars with the radius T_d. We can see that the trainee's clicks are detected by the red plus signs successfully using the criteria described above. As an example of such situation, see the red plus sign closest to the yellow cross #2. Thus, the location found by the algorithm closest to the trainee's false positive (yellow cross #1) is the only correct prediction.



(a) (b)

Fig. 2 Definition of correct false positive prediction provided by the algorithm. (a) is an example of a mammogram and (b) is the close up of the green box in (a). The green star is a mass centroid computed using experts' marks. Yellow crosses are clicks made by the trainee (yellow cross #1 is a false positive mark and yellow cross #2 is true positive mark). Red plus points are suspicious locations found by the algorithm. The cyan circles show the round areas centered at yellow crosses and green stars within a radius of 96 pixels.

4.2 Cross validation

To minimize bias in our evaluation of the algorithm, we applied a leave-one-case-out cross validation approach (Kearns and Ron 1999). Specifically, for each individual trainee, one case was excluded from the dataset and the remaining cases were used for development of the algorithm. Please note that while many parts of the algorithm (such as the DoG filtering) do not change based on the training data, what will change is the random forest classifier. After the algorithm is trained, it is applied to the images of the case that was left-out and it provides a set of locations for each image along with a likelihood of becoming a false positive location provided by the classifier. This process is repeated multiple times such that each case is excluded from training set and is assigned a set of locations once.

4.3 Evaluation metrics

The purpose of the algorithm is to find normal locations that will be erroneously identified as abnormal by a radiology trainee. For this purpose we employed two evaluation metrics: (1) free-response receiver operating characteristic (FROC) (Chakraborty 1989) to show the percentage of "detected" false positive clicks given a different number of incorrectly predicted false positive clicks; and (2) positive predictive value (PPV) to show, at a given number of predicted locations, the proportion of such locations that will actually result in false positive errors committed by the trainee.

FROC allows us to see how well the algorithm is doing at finding the locations that will be incorrectly marked by a trainee at a given average number of locations indicated by the algorithm but not corresponding to a trainee's error. In our FROC evaluation, sensitivity is defined as follows:

$$Sensitivity = \frac{\text{Number of false positive locations found by the algorithm}}{\text{Number of false positive errors made by trainee}}$$

Higher sensitivity with lower number of predictions means better performance of the algorithm. Besides FROC for false positive clicks, we additionally evaluated our algorithm for the complete set of clicks for all trainees (false positive and true positive clicks) see how well the algorithms predict trainees marking behavior in general. The FROC curves were calculated individually for each trainee and the curves of the 10 trainees were averaged to show the overall performance of our algorithm.

Furthermore, we evaluated performance of our algorithm in respect to a particular practical scenario in which a certain limited number of locations is selected to be presented to a trainee. Specifically, the

locations were selected based on the likelihood of being a false positive location provided by our algorithm. We calculated how many of such locations found by the algorithm actually correspond to a trainee's false positives. Formally, this performance measure is positive predictive value of our algorithm (PPV):

$$PPV = \frac{\text{Number of correctly predicted false positives provided by the algorithm}}{\text{Number of predicted false positives provided by the algorithm}}$$

The PPVs were calculated individually for the 10 trainees and then average to provide the overall performance of our algorithm.

As we are not aware of any previous algorithms approaching the task tackled in this paper, we compared our algorithm to a "chance" algorithm, which finds the locations randomly. To simulate such an algorithm, we first selected the same number of random locations as the number of locations indicated by the DoG filter, and then each random location was assigned a random value that is used as its likelihood to be a trainee's false positive error. Showing the superiority of our algorithm over the "chance" algorithm will allow us to establish that the false positive error making in radiology trainees is not entirely random but rather it is driven by a pattern that we were able to capture in our algorithm.

5 Experimental results

Analysis of the marks indicated by experts showed that there were total of 154 mass locations that could be used in our analysis based on a majority vote, among which 97 locations were indicated by all 3 experts and 57 locations were indicated by 2 out of 3 experts. The image-based Cohen's Kappa scores were computed using the 400 images in the dataset. The image-based Cohen's Kappa calculated for all 400 images were 0.5059 between Expert 1 and Expert2, 0.7003 between Expert 1 and Expert 3, and 0.4570 between Expert 2 and Expert 3. This indicates that the three experts achieved moderate to substantial agreement.

Table 2 lists the number of true positive and false positive annotations made by 10 trainees (R1 to R7 are the 7 residents and N1 to N2 are the 3 novices) for all 400 images in the database. One can see wide variability in the number of false positive errors made which ranges from 12 to 341.

Table 2. Number of true positives and false positives of trainees

Subject	R1	R2	R3	R4	R5	R6	R7	N1	N2	N3
Number of True	91	67	103	98	61	106	93	63	74	72

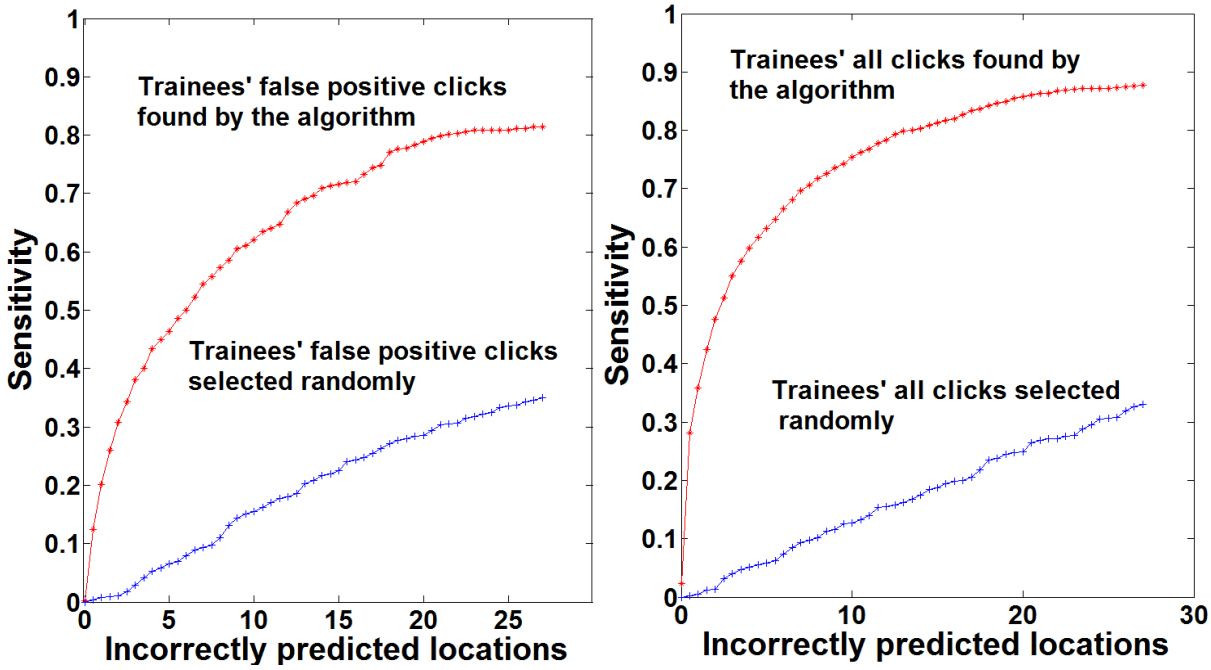
Positives										
Number of False Positives	41	12	49	51	64	341	46	70	124	158

Table 3 shows the detection rates of the DoG filter for all clicks (true positive + false positive) and false positive clicks only for the 10 trainees (R1 to R7 are the 7 residents and N1 to N3 are the 3 novices).

Table 3. Detection rates of DoG filter for true positive + false positive (TP+FP) and false positive (FP)

	R1	R2	R3	R4	R5	R6	R7	N1	N2	N3	Mean
TP+FP	0.89	0.87	0.88	0.83	0.91	0.89	0.87	0.87	0.91	0.86	0.88
FP	0.88	0.67	0.82	0.73	0.91	0.85	0.80	0.76	0.91	0.81	0.81

Figure 3 shows FROC curves for our algorithm and randomly selected locations. The red curves are FROC generated using algorithm predicted locations and the blue curves are FROC generated using random locations.



(a) (b)

Fig. 3 FROC curves for trainees' false positive clicks and all clicks prediction. (a) shows the FROC curve for the trainees' false positive clicks (red *) found by the algorithm and the FROC curve for trainees' false positive clicks (blue +) selected randomly. (b) shows the FROC curve for trainees' all clicks (red *) found by the algorithm and the FROC curve for trainees' all clicks (blue +) selected randomly.

By comparing the red and blue curves in Fig. 3-a and b, we can see that our algorithm has much higher sensitivities than the random selection, which means that it is able to locate a much higher number of

locations that will be associated with trainees' false positive errors. When the number of incorrectly predicted locations is equal to 20, the sensitivities of red curves are nearly 4 times higher than those of blue curves and when the number of incorrectly predicted locations is 5, the sensitivities of red curves are about 10 times higher than those of blue curves. Therefore, the proposed algorithm can predict where trainees are likely to commit false positive clicks much better than chance.

The PPV of algorithm predicted locations are illustrated in Fig. 4 with 1, 10, 20, 30, 40, 50, and 60 predicted false positives. For the location with the highest likelihood for each trainee, 4 out of 10 locations catch the false positive clicks made by trainees correctly. When the number of predicted false positives is from 10 to 60, the PPVs are higher than or equal to 10%. The comparison of PPVs of algorithm predicted locations and random locations are listed in Table 4. Clearly the algorithm predicted locations have notably higher PPV than random locations.

Table 4. PPV for locations found by the algorithm and locations selected randomly

Number of predicted False Positives	1	10	20	30	40	50	60
PPV: locations found by the algorithm	0.4	0.12	0.115	0.12	0.105	0.10	0.10
PPV: Random locations	0	0	0	0	0.0025	0.0020	0.0017

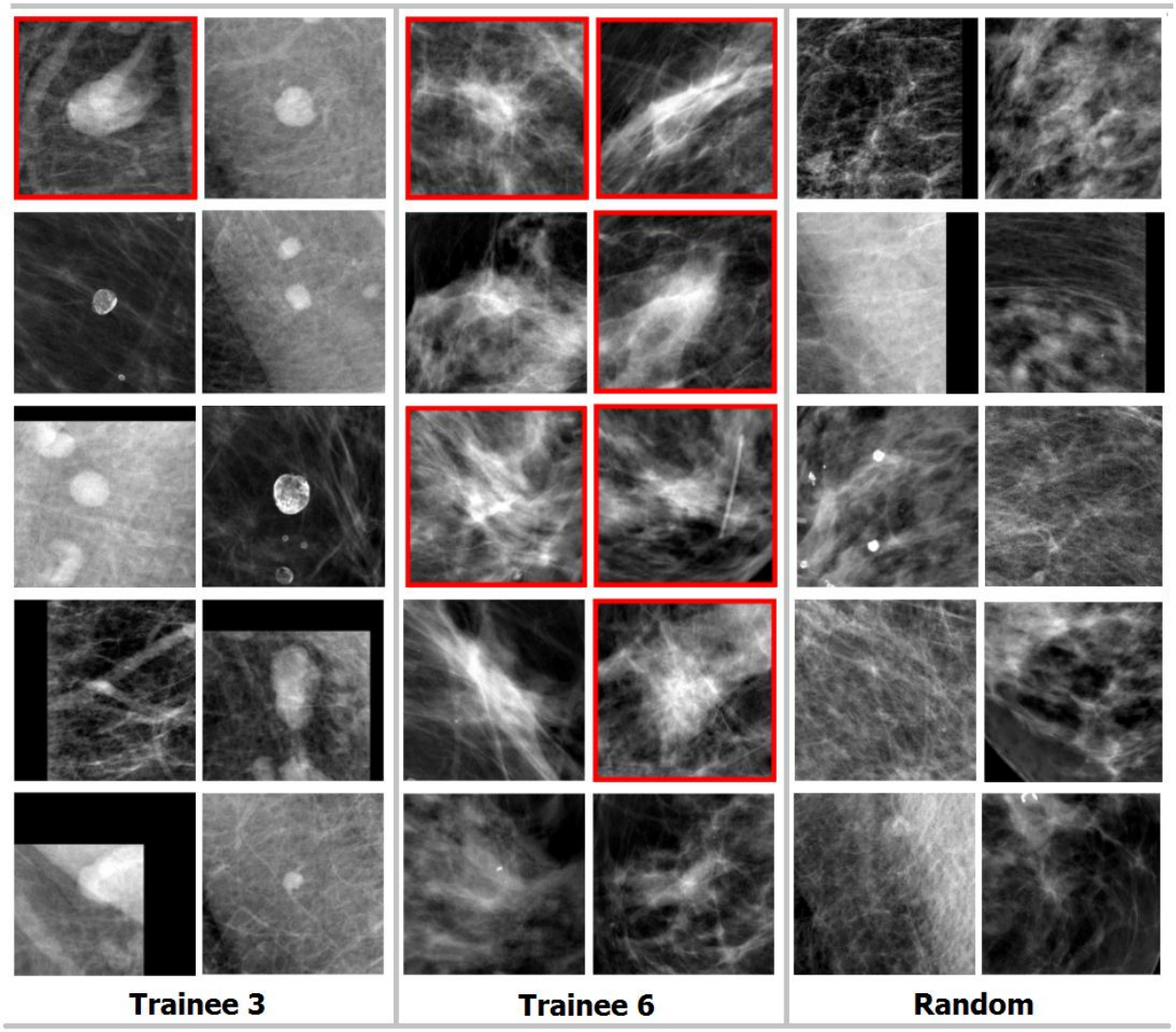


Fig. 4 The images of top 10 predicted false positives. The 10 images in the first column and the 10 images in the second column are top 10 false positives predicted by the algorithm for trainee 3 and trainee 6, respectively. The 10 images in the third column are the top 10 false positives predicted based on random values of randomly selected locations. The images with red frames are correctly predicted false positives.

The images of top 10 predicted false positives for trainee 3 and trainee 6 are shown in the first two columns in Fig. 4. For comparison purpose, the images of top 10 random locations are also shown in the third column. The images with red frame are correctly predicted false positive errors made by trainees. We can see that all images for trainee 3 and trainee 6 contain regions that resemble mass regions that are more likely to be associated with false positive errors. Because trainee 6 has 341 false positive clicks and trainee 3 has only 49 false positive clicks as listed in Table 4, it is not a surprise that a much higher

proportion of the locations identified by our algorithm are associated with an actual false positive click for the trainee 6 than for the trainee 3. While the predicted locations for the trainee 3 and the trainee 6 are different, examining the top 10 locations for other trainees revealed that there is a notable overlap between the selected locations between trainees suggesting that there are both common and individual error making patterns among the trainees.

The analysis of the importance of individual features in predicting false positive errors (based on single run of random forest for each trainee and the "OOBPermutedVarDeltaError" parameter) showed that Intensity-based area under the ROC curve (F103), Region circularity (F4), and Normalized intensity of region (F3) are the most important features for false positive prediction. Among the three features, F103 was the most important feature for 5 trainees, F4 for 4 trainees, and F3 for 1 trainee.

6 Conclusions and Discussion

In this first exploratory study on modeling false positive error making behavior for radiology trainees from the perspective of mammography education, we were able to build computer models that are able to find unseen locations that were more likely to be associated with false positive errors. Our algorithm used computer vision and machine learning-based approaches to identify suspicious false positive locations, segment suspicious locations, extract image features, and predict the likelihood of false positive error in a fully automatically manner without human intervention. The FROC and PPV based evaluations on the database containing 100 mammographic cases demonstrate that the proposed algorithm can provide much more accurate prediction of false positive locations than chance.

Although much better than those of randomly selected locations, the FROC and PPV of locations predicted by the algorithm are not high. This result is expected for a number of different reasons. First, the number of false positive errors is very low for some trainees which results in: (1) a lower number of training samples for our model (to identify patterns in a trainee's error making) and a lower likelihood of identifying the actual false positives by our algorithm as the trainees with lower number of false positives will dismiss many locations that might appear challenging to our algorithm. Please note that the performance of our algorithm for the trainee for whom many false positive errors were available (trainee

6) is very good. The PPVs of the algorithm for this trainee at 1, 10, 30, and 60 selected locations are 1, 0.6, 0.5, and 0.5167, which are much higher than the average values shown in Table 4.

Second, while we believe that there is a pattern in error making among radiology trainees, we also believe that part of error making cannot be explained by properties of the images but is rather caused by other factors, such as the level of distraction and fatigue of the trainee as well as other factors. As we do not measure these factors, they are considered random (noise) from our perspective. Future educational systems could measure these factors through eye tracking and other modern technologies and incorporate them in the error modeling. Finally, even though the proportion of actual false positive errors among the locations identified by the algorithm was low for some trainees, it does not mean that the remaining locations were not challenging for the trainee. In other words, not all challenging locations are expected to result in false positive errors (i.e., a good trainee is expected to dismiss a large portion of challenging locations). Visual inspection of the locations selected by the algorithms suggests that many of the selected locations might pose increased challenge to the trainees. The final test of usefulness of the locations indicated by our algorithm would be through evaluation of their educational benefit.

Future work on improvement of the algorithm's performance will focus on two issues. First, new segmentation algorithms and feature extraction algorithms can be applied to improve characterization of the suspicious regions indicated by the DoG filter. Modern machine learning methods such as deep learning can also be applied for this purpose. Second, we will collect more data. Availability of data from more readers and for more cases will improve performance of the classifiers that determine locations with high and low likelihood of false positive mark. Finally, we will combine non-imaging features, such as patient's age and medical history, with imaging features used in our current study to explore the impact of non-imaging features on error. Further research could also investigate locations that were commonly indicated by trainees but not experts and incorporate expert-generated explanations in the training process.

Practical applications of predicting challenging locations for radiology trainees that learn mammography are significant. We envision a system that displays such more difficult negative locations to each trainee for more targeted training. Specifically, our system will learn the trainee's individual weaknesses and construct a user model based on his/her previous image interpretations. Then the algorithm will search the database of available images in order to identify images and locations in those images that might

pose a challenge to the trainee. Such locations will be presented to the trainee in order to focus their training on challenging cases/locations. Our fully automated approach makes such a system possible. Focusing of more challenging locations rather than random selection of material is likely to improve efficiency of the trainee's training.

Conflict of interest

Dr. MM receives grant funding from the Department of Defense Breast Cancer Research Program. He also receives consulting fees from American College of Radiology Image Metrix (contractor to GE) for his services as a scientific consultant

Reference:

Adams, R. and L. Bischof (1994). "Seeded region growing." Pattern Analysis and Machine Intelligence, IEEE Transactions on **16**(6): 641-647.

Babaud, J., A. P. Witkin, M. Baudin and R. O. Duda (1986). "Uniqueness of the Gaussian kernel for scale-space filtering." Pattern Analysis and Machine Intelligence, IEEE Transactions on(1): 26-33.

Baker, J. A. and J. Y. Lo (2011). "Breast tomosynthesis: State-of-the-art and review of the literature." Academic radiology **18**(10): 1298-1310.

Breiman, L. (2001). "Random forests." Machine learning **45**(1): 5-32.

Catarious Jr, D. M., A. H. Baydush and C. E. Floyd Jr (2006). "Characterization of difference of Gaussian filters in the detection of mammographic regions." Medical physics **33**: 4104.

Chakraborty, D. P. (1989). "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data." Medical physics **16**: 561.

Grimm, L. J., S. V. Ghate, S. C. Yoon, C. M. Kuzmiak, C. Kim and M. A. Mazurowski (2014). "Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features." Medical physics **41**(3): 031909.

Haralick, R. M., K. Shanmugam and I. H. Dinstein (1973). "Textural features for image classification." Systems, Man and Cybernetics, IEEE Transactions on(6): 610-621.

Kearns, M. and D. Ron (1999). "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation." Neural Computation **11**(6): 1427-1453.

Masotti, M., N. Lanconelli and R. Campanini (2009). "Computer-aided mass detection in mammography: False positive reduction via gray-scale invariant ranklet texture features." Medical physics **36**: 311.

Mazurowski, M. A., J. A. Baker, H. X. Barnhart and G. D. Tourassi (2010). "Individualized computer-aided education in mammography based on user modeling: Concept and preliminary experiments." Medical physics **37**: 1152.

Mazurowski, M. A., H. X. Barnhart, J. A. Baker and G. D. Tourassi (2012). "Identifying error-making patterns in assessment of mammographic BI-RADS descriptors among radiology residents using statistical pattern recognition." Academic Radiology **19**(7): 865-871.

Mazurowski, M. A. and G. D. Tourassi (2011). Exploring the potential of collaborative filtering for user-adaptive mammography education. Biomedical Sciences and Engineering Conference (BSEC), 2011, IEEE.

Mello-Thoms, C., S. Dunn, C. F. Nodine, H. L. Kundel and S. P. Weinstein (2002). "The perception of breast cancer: what differentiates missed from reported cancers in mammography?" Academic radiology **9**(9): 1004-1012.
Mudigonda, N. R., R. M. Rangayyan and J. Leo Desautels (2001). "Detection of breast masses in mammograms by density slicing and texture flow-field analysis." Medical Imaging, IEEE Transactions on **20**(12): 1215-1227.
Polakowski, W. E., D. A. Cournoyer, S. K. Rogers, M. P. DeSimio, D. W. Ruck, J. W. Hoffmeister and R. A. Raines (1997). "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency." Medical Imaging, IEEE Transactions on **16**(6): 811-819.
Sun, S., P. Taylor, L. Wilkinson and L. Khoo (2008). Individualised training to address variability of radiologists' performance. Medical Imaging, International Society for Optics and Photonics.
Sun, S., P. Taylor, L. Wilkinson and L. Khoo (2008). An ontology to support adaptive training for breast radiologists. Digital Mammography, Springer**:** 257-264.
Timp, S. and N. Karssemeijer (2004). "A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography." Medical Physics **31**: 958.
Timp, S., N. Karssemeijer and J. Hendriks (2003). Analysis of changes in masses using contrast and size measures. Digital Mammography, Springer**:** 240-242.
Tourassi, G., S. Voisin, V. Paquit and E. Krupinski (2013). "Investigating the link between radiologists' gaze, diagnostic decision, and image content." Journal of the American Medical Informatics Association.
Voisin, S., F. Pinto, G. Morin-Ducote, K. B. Hudson and G. D. Tourassi (2013). "Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography." Medical Physics **40**: 101906.
Wei, D., H.-P. Chan, N. Petrick, B. Sahiner, M. A. Helvie, D. D. Adler and M. M. Goodsitt (1997). "False-positive reduction technique for detection of masses on digital mammograms: Global and local multiresolution texture analysis." Medical Physics **24**: 903.