MIPS XX

August 6 – 9, 2024
Vanderbilt University

Psychology Department, Wilson Hall
111 21st Ave South
Nashville, TN

Organizers:
Frank Tong, Vanderbilt University
Elizabeth Krupinski, Emory University

Medical Image Perception Society

mips

Schedule

Date & Time	Title	Presenter
TUESDAY August 6th		
5:30 – 7:30 pm	**Opening Reception** **Wilson Hall Lobby**, 111 21st Ave South	
WEDNESDAY August 7th		
Breakfast	**Hotel Breakfast Area (Hyatt House)**	**On your own** **(*included in hotel fee*)**
Meeting	**Vanderbilt's Psychology Department**	**Wilson Hall room 115**
9:00 – 9:10 am	Welcome	Frank Tong
Session 1	*Session Chair*	*Todd Horowitz*
9:10 – 9:30 am	Quantifying the Margin of Error for a High-Stakes Diagnostic Decision Using Response Models	Martin Pusic
9:30 – 9:50 am	Using global scoring to evaluate performance on screening mammograms within a simulation setting	Lonie Salkowski
9:50 – 10:10 am	Sensitivity and Specificity Confidence Intervals (CIs) vs. Joint Confidence Region (CG)	Yulei Jiang
10:10 – 10:40 am	**Coffee Break**	
Session 2	*Session Chair*	*Todd Horowitz*
10:40 – 11:00 am	Effect of prevalence rate on perceptual learning of lung nodule detection in initially naïve observers	Frank Tong
11:00 – 11:20 am	The Impact of Response Mode and Recalibration on Crowdsourced Medical Image Labels	William Holmes
11:20 – 11:40 am	The impact of dose information on radiographer's subjective interpretation of image quality	Mark McEntee on behalf of Shauna Lane
11:40 – 11:55 am	**Group Photo**	
12:00 – 1:00 pm	**Lunch (mediterranean cuisine from Sadie's)**	
Session 3	*Session Chair*	*William Holmes*
1:00 – 1:20 pm	Trials and tribulations of implementing interpretive skills simulation system for resident breast imaging education	Lonie Salkowski
1:20 – 1:40 pm	Predicting Radiologists' Visual Attention for Mammograms	Hantao Liu
1:40 – 2:00 pm	Breast Cancer Survivors' Perceptual Map of Breast Reconstruction Appearance Outcomes	Haoqi Wang (MIPS scholar)
2:00 – 2:20 pm	Mitigating search errors with 3D image stacks	Miguel Eckstein
2:20 – 2:40 pm	**Coffee Break**	
Session 4	*Session Chair*	*Megan Mills*
2:40 – 3:00 pm	Using a Limited Field of View to Improve Nodule Perception and Reduce Eye Strain	William Auffermann
3:00 – 3:20 pm	Assessing visual hindsight bias in radiologist eye tracking metrics	Jacky Chen (MIPS scholar)
3:20 – 3:40 pm	Mitigate, don't litigate: Dealing with "Look but Fail to See" errors in Radiology	Jeremy Wolfe
5:30 – 7:00 pm	**Social Event: Johnny Cash Museum** Hyatt House lobby at **5:00 pm**, Uber to museum	
7:00pm - whenever	**Dinner on your own**	

THURSDAY August 8th		
Breakfast	**Hotel Breakfast Area**	**On your own** **(*included in hotel fee*)**
Meeting	**Vanderbilt's Psychology Department**	**Wilson Hall room 115**
Session 5	*Session Chair*	*Frank Tong*
9:00 – 10:00 am	**Keynote:** How domain-general object recognition may help us understand medical expertise	Isabel Gauthier
10:00 – 10:10 am	**Coffee Break**	
10:10 – 10:50 am	MIPS business meeting	Elizabeth Krupinski
Session 6	*Session Chair*	*Jay Hegdé*
10:50 – 11:10 am	Gamification for Image Perception and Emergency Radiology Education	Soham Banerjee
11:10 – 11:30 am	Global gist signal differs between specific genetic subtypes of breast cancer	Karla Evans & Roisin Bradley
11:30 – 11:50 am	The impact of radiation dose information on subjective evaluation of CT image quality	Mark McEntee on behalf of Conor Lee
11:50 - 12:10 pm	Identifying the Textural Components of the Global Gist Signal of the Abnormal	Cameron Kyle-Davidson (MIPS scholar)
12:10 – 1:20 pm	**LUNCH (boxed order from Chef's Market)**	
Session 7	*Session Chair*	*Karla Evans*
1:20 – 1:40 pm	Prediction and generalization performance for ramp-spectrum forced-localization tasks	Craig Abbey
1:40 – 2:00 pm	Multi-reader multi-case AUC analysis methodology for studies involving Artificial Intelligence Applications	Stephen Hillis
2:00 – 2:20 pm	A Dual-Branch Deep Learning Model for MRI Image Quality Assessment	Hantao Liu on behalf of Yueran Ma
2:20 – 2:40 pm	Exploring the Explainability of a Machine Learning Model for Prostate Cancer: Do Lesions Localize with the Most Important Feature Maps?	Murray Loew
2:40 – 3:00 pm	**Coffee Break**	
Session 8	*Session Chair*	*Cameron Kyle-Davidson*
3:00 – 3:20 pm	The Perceptual Bias Cascade in the Medical-AI Information Value Chain	Jennifer Trueblood
3:20 – 3:40 pm	Medical Image Processing by Artificial Intelligence Software: Outcome Quality Vetting by Human Observers	Jay Hegdé
3:40 – 4:00 pm	EyeSee: Integrating Deep Learning for Image Analysis	Bin Wang & Ulas Bagci
6:00 – 8:30 pm	**Conference Dinner** **19th floor of Zeppos Tower** (near 2415 West End Ave) Meet at 5:45pm in Hyatt House Lobby, walk to dinner site (~15min walk)	

FRIDAY August 9th		
Breakfast	**Hotel Breakfast Area**	**On your own** **(*included in hotel fee*)**
Meeting	**Vanderbilt's Psychology Department**	**Wilson Hall room 115**
Session 9	*Session Chair*	*Jennifer Trueblood*
9:00 – 9:20 am	Training ophthalmologists to recognize novel retinal markers identified using artificial intelligence	Ipek Oruc
9:20 – 9:40 am	Using Expert Gaze for Self-Supervised and Supervised Contrastive Learning of Glaucoma from OCT Data	Kaveri Thakoor & Wai Tak Lau (MIPS scholar)
9:40 – 10:00 am	Eye-Tracking to Evaluate AI Use	Elizabeth Krupinski
10:00 – 10:20 am	False-Color Images to Facilitate Image Perception and Radiology Education	William Auffermann
10:20 – 10:40 am	**Coffee Break**	
Session 10	*Session Chair*	*Elizabeth Krupinski*
10:40 – 11:00	Using Deep Learning to Predict Radiologists' Decisions when Reading Mammograms	Karthika Kelat (MIPS scholar)
11:00 – 11:20	A phantom image quality study comparing ultra-low and standard dose computed tomography protocols for the investigation of non-accidental injury	Mark McEntee on behalf of Lisa Kingston
11:20 – 11:40	Convolutional Neural Network Model Observer During Search in Virtual Digital Breast Tomosynthesis	Miguel Eckstein
11:40 - 12:00 pm	**Wrap-up and adjourn**	

WEDNESDAY MORNING, August 7th
Session 1 – 9:10 am to 10:10 am

Quantifying the Margin of Error for a High-Stakes Diagnostic Decision Using Response Models

Martin V. Pusic MD PhD[1], David A Cook[2] MD MHPE, Matthew Lineberry[3] PhD,
Joseph Bennett[4] MD, Laura Penalo[5] RN PhD, Julie Friedman[6] MD, David Rhee[5] MD,
Jeffrey Lorin[6] MD, Barry Rosenzweig[6] MD, Rose Hatala[7] MD PhD.
Departments of Pediatrics[1] & Emergency Medicine[1], Harvard Medical School
Department of Internal Medicine[2], Mayo Clinic
Department of Population Health[3], University of Kansas
Departments of Emergency Medicine[4], Internal Medicine[5], Cardiology[6], New York University
Department of Internal Medicine[7], University of British Columbia

Rationale

Clinicians often find themselves making diagnoses with incomplete information, thus under conditions of uncertainty. It can be difficult to disentangle what is uncertainty due to incomplete mastery and what is structural uncertainty where even a fully trained expert would have difficulty distinguishing between two alternative diagnoses. When the stakes of such a decision are high, clinicians may opt to build-in a margin for error. In this study, we explore a statistical approach using item-response models to examine how this margin for error varies with expertise.

Methods

We assembled an image bank of 100 ECGs that had varying degrees of ST elevation that could be confused with ST Elevation Myocardial Infarction (STEMI) with there being a roughly equal ratios of Normal:Distractor Diagnosis:STEMI diagnoses. 100 raters of varying expertise diagnosed the cases as to whether they represented a STEMI for which urgent catheterization should be considered. We modeled the resulting fully crossed data (100 raters x100 ECGs) using an Item Response Model, reporting the decision-thresholds and graphic probability tracelines. We demonstrate the margin for error using individual case locations on the underlying scale.

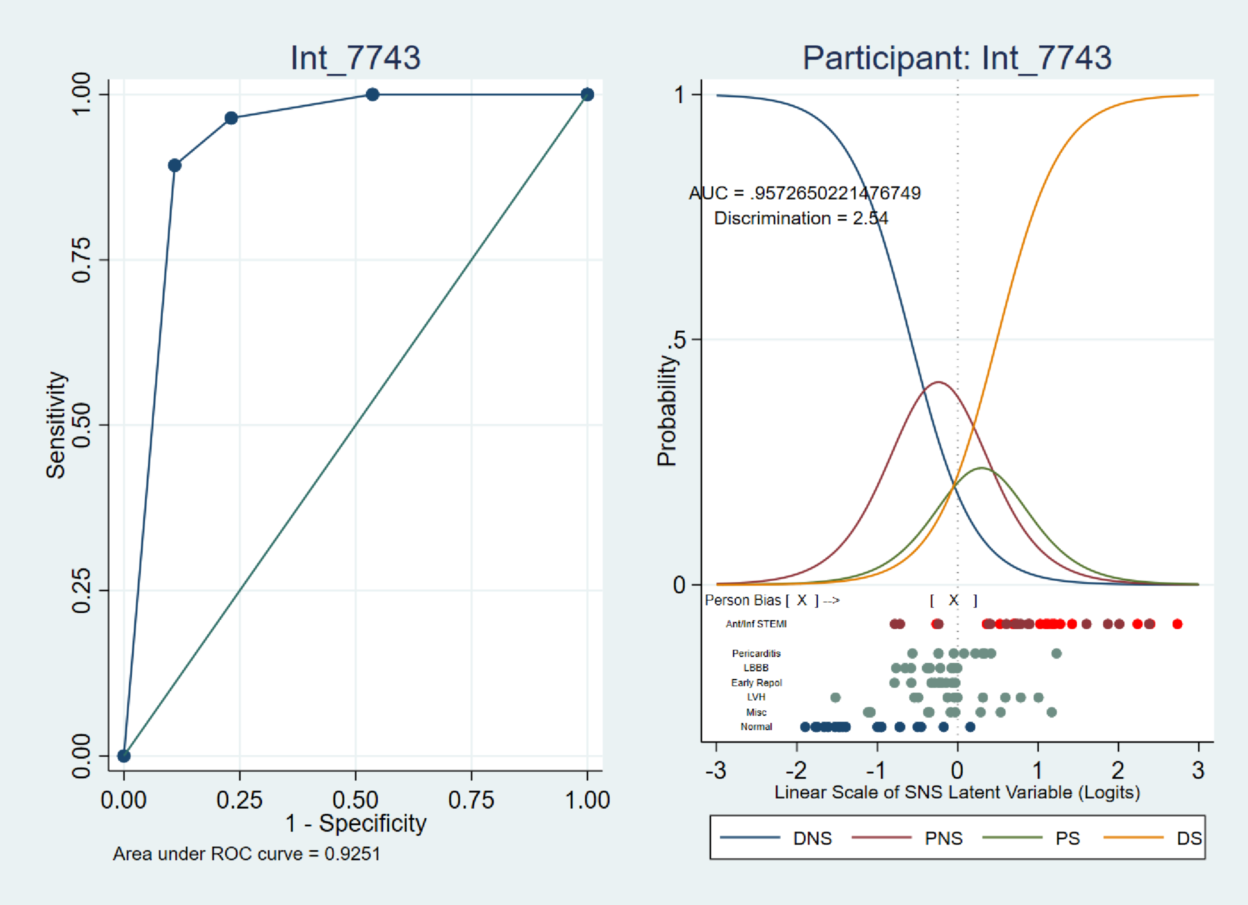
Results

Under the item response model, the 100 cases demonstrated a full range of diagnostic uncertainty, from -2.0 logits (Not STEMI) to +3.0 logits (+STEMI). There was considerable overlap between case diagnoses making absolute discrimination difficult. Modelling practitioners decision thresholds, we found that they demonstrated considerable practice variation in where they set their decision cut points (See Figure for an example).

Conclusions

Item response modeling, when aligned with an important clinical distinction as in this case, can be used to provide meaningful feedback as to a clinician's overall tendencies when faced with uncertain cases. The concept of an acceptable margin for error can be estimated using this approach but full certainty is not possible.

Figure: ROC Curve and Probability traceline for one clinician rating 100 potential STEMI ECGs.



At left is the ROC curve for one clinician rating ECG cases. At right are the corresponding Item Response Model decision threshold tracelines. The case locations are represented by the coloured circles at the x-axis (red=STEMI by reference standard; blue=Normal; greenish=Other Diagnosis). A case at the zero point would be predicted to be equally likely to be assigned either category ("STEMI" or "Not STEMI") by a clinician of average bias, as in this case. Clinicians whose traceline mid-point is to the left of zero show a larger margin of error; those to the right show a bias towards a smaller margin of error.

Using global scoring to evaluate performance on screening mammograms within a simulation setting

LR Salkowski MD PhD[1], DM Bolt, PhD[2], AM Fowler PhD MD[1], EA Krupinski PhD[3], MA Elezaby MD[1]
[1]University of Wisconsin-Madison SMPH, [2]Univeristy of Wisconsin-Madison, [3]Emory University

Rationale

Simulation is an effective method to enhance and assess progress in an area of interest. Interpretive skills simulation of screening mammograms is one way to assess progress and provide constructive feedback within a radiology residency training program. Residents have only 12-weeks of breast imaging to become proficient in all aspects of breast imaging. There is no cross training in other rotations nor call experience to support their knowledge development in breast imaging. Within a simulation setting, there is a vast amount of data. Developing a global scoring framework for each case can facilitate analysis.

Methods

This IRB-approved study used machine learning techniques to construct simulation case sets to be implemented in the 3rd or 4th week of a 4-week breast imaging rotation. Each simulation session had 50 screening mammography cases. A global score was developed that included six (6) elements: breast tissue density, need for a technical recall, lesion identification, lesion overcall, BI-RADS assessment, and appropriate management. Each case was assigned a score against each element. Performances of an initial 15 participants were analyzed using individual case scores averaged across elements. Next, we evaluated performance on known cancer cases compared to no cancer.

Results

Fifteen (15) of the 19 (79%) invited residents completed the 50-case set screening simulation. The average time to complete the simulation was 4.5 hours (SD=1 hr 6 min). Using global scoring, preliminary principal components analysis distinguished on a primary dimension participants with respect to their relative performances on cases with no lesion cases with more than one lesion. Two subjects demonstrated data 'bias' with one subject performing disportionately better in cases without a lesion, and a second subject performing better on cases with lesions. A secondary dimension appears to be an overall performance dimension.

Conclusions

A global scoring framework can be effectively used within a screening simulation system to evaluate the relative performance on cases and individuals compared to each other. Participant performance appears sensitive to the presence and number of lesions in the simulated cases. Additional participant data should enable stronger conclusions regarding the psychometric quality of individual simulated cases.

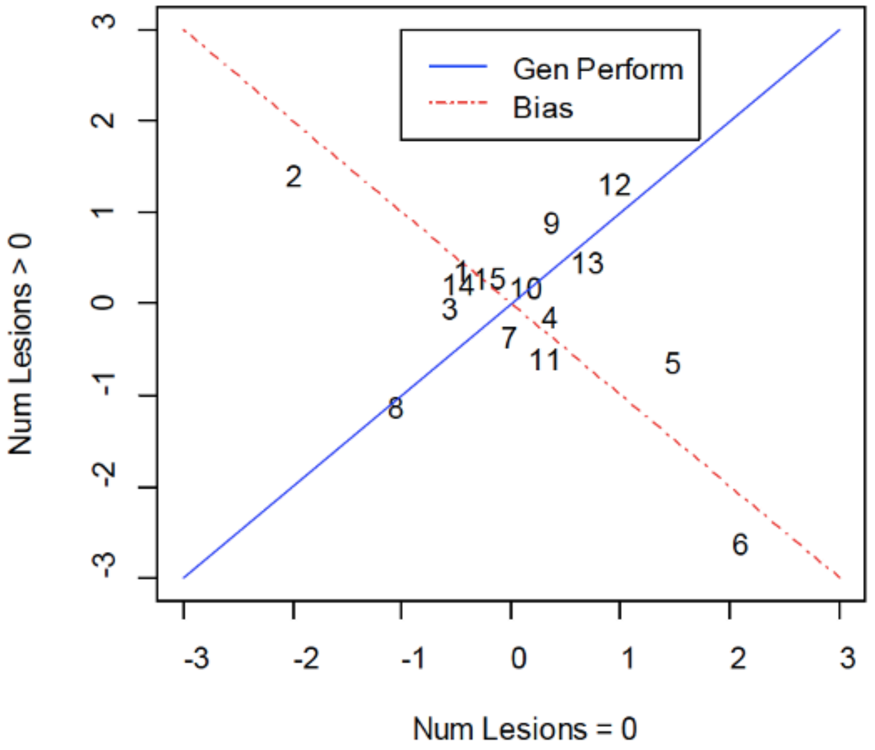


Figure: Relative performance of subjects comparing when there are no lesions (x-axis) compared to more than one lesion (y-axis). The axes are represented in z scores. The blue line represents general performance, and dashed red line introduces a bias when subjects 2 and 6 are included. Note: The numbers on the graph represent individual subjects.

Sensitivity and Specificity Confidence Intervals (CIs) vs. Joint Confidence Region (CG)

Yulei Jiang, Ph.D., Department of Radiology
The University of Chicago

Rationale

In imaging and radiology studies, we often are interested in sensitivity and specificity of certain imaging modality, technique, observer, or combinations thereof. Sensitivity is the probability of an actually positive patient receiving a true-positive diagnosis and specificity is the probability of an actually negative patient receiving a true-negative diagnosis, given any particular disease or condition of interest. Confidence intervals (CIs) quantify the uncertainty or precision in the estimate of sensitivity and specificity from clinical case series or reader studies. However, by convention CIs are calculated separately for sensitivity and specificity. This leads to problems in interpretation. We discuss these problems and propose a joint sensitivity-specificity confidence region (CG) for improvement.

Methods

Referring to an ROC curve, which plots sensitivity vs. one minus specificity, we recognize that sensitivity and specificity must be interpreted in the context of each other and not each by itself. Therefore, given a CI for sensitivity (or vice versa for specificity), the implication for specificity (or sensitivity) is ambiguous—whether it is a fixed value (zero-width CI) or have an independent (non-zero-width) CI or else. A joint sensitivity-specificity CG overcomes this ambiguity by specifying a set of sensitivity-specificity value pairs. This CG can be obtained by extending the same principle for estimating the CIs. Given an observed sensitivity (or specificity) value, we estimated its CI from the likelihood function related to the binomial probability density function. Integrating the likelihood function by rank order gave rise to CI estimates. Similarly, given an observed sensitivity-specificity value pair, integration of the joint likelihood function gave rise to joint CG estimates. We validated these methods with simulations.

Results

A likelihood-based CI or CG represents the probability that the unknown population value or population value pair is within a CI or CG estimate, respectively. A CG is ellipsoid-like in shape and its longest orthogonal dimensions are beyond the rectangular region defined by separate CIs for sensitivity and specificity. Simulation results showed that empirical probabilities that a population value or population value pair is within CI or CG estimates agree with nominal probabilities to an arbitrary accuracy limited by numerical-integration precision.

Conclusions

Joint sensitivity-specificity confidence region estimated from the likelihood function is a better alternative to conventional confidence intervals estimated separately for sensitivity and specificity.

Effect of prevalence rate on perceptual learning of lung nodule detection in initially naïve observers

Frank Tong[1], PhD, Hui-Yuan Miao[1], BS, Zoe Armstrong[1], BS,
Anthony Micetich[2], MD, and Edwin F. Donnelly[3], MD, PhD
[1] *Psychology Department, Vanderbilt University*
[2] *Department of Radiology, Vanderbilt University Medical Center*
[3] *Department of Radiology, The Ohio State Wexner Medical Center*

Rationale

Does the acquisition of expertise at a challenging diagnostic task, such as the detection of lung nodules in 2D chest radiographs, depend more on learning from positive examples as compared to negative examples? If so, then low prevalence rates in the clinic could impede the efficacy of radiology training. Previously, we have shown that undergraduate observers show progressive improvements at detecting simulated nodules in 2D chest X-rays over a series of training sessions, with successful generalization to real nodule cases (Tong et al., MIPS 2022). Here, we evaluated whether prevalence has an impact on the rate at which initially naïve observers can learn and improve at tasks of nodule detection and localization.

Methods

We recruited 60 undergraduate participants who completed the 4-session study; each participant was randomly assigned to 1 of 3 prevalence groups. We obtained pre- and post-training measures of nodule localization accuracy using a challenging set of real and simulated nodule cases. For nodule detection training, participants were presented with chest X-rays that contained visually realistic simulated nodules at prevalence rates of either 5%, 20% or 50%, and evaluated 600 examples over 3 sessions. Post-training performance was evaluated in session 4. Payment was scaled to reward more accurate performance.

Results

All three groups showed significant improvements in detection performance across training sessions, with the low prevalence group exhibiting conservative decisional bias as expected. D-prime measures of detection sensitivity revealed a significant Group × Session interaction effect, due to greater improvement across sessions at higher prevalence ($F(4,114) = 2.53$, $p < 0.05$). While not all group comparisons proved statistically significant, planned comparisons revealed better post-training localization of simulated nodules for high vs. low prevalence training ($t(38) = 2.49$, $p < 0.05$). Moreover, only the high prevalence group showed significantly better localization of real nodules following training ($t(19) = 2.33$, $p < 0.05$). All groups showed improved performance at localizing simulated nodules post-training, but these improvements were more prominent for high prevalence groups ($F(2,57) = 4.46$, $p < 0.05$).

Conclusions

Our results strongly suggest that initially naïve observers can acquire more information from positive than negative examples when learning to perform a challenging radiology task. Thus, while training in the clinic is essential for the development of radiological expertise, it may be beneficial to supplement this training with exposure to numerous positive examples to boost performance. Such training will not only lead to shifts away from conservative bias (cf. Wolfe & Van Wert, 2010; Evans et al., 2013), they may lead to improved learning of relevant internal templates or heightened sensitivity as defined by signal detection theory.

The Impact of Response Mode and Recalibration on Crowdsourced Medical Image Labels

William R. Holmes, PhD; Andrew Caplin, PhD; Erik Duhaime, PhD; Gunnar Epping, BA; Daniel Martin, PhD; Jennifer S. Trueblood, PhD

Mathematics and Cognitive Science, Indiana University; Department of Economics, New York University; Centaur Labs; Psychological and Brain Sciences, Indiana University; Department of Economics, University of California Santa Barbara; Psychological and Brain Sciences, Indiana University

Rationale

We investigate different approaches to crowdsourcing medical image labels and the effects of those approaches on the properties of AI models trained on those data. Ideally, experts should annotate biomedical images for model training, but costs can be prohibitive. Crowdsourcing has emerged as a solution to this bottleneck. Biomedical data annotation company, Centaur Labs, employs skilled annotators who label content through a gamified app, DiagnosUS. These annotators have diverse backgrounds and locations, with many being medical students from different countries. While medical students have some medical training, they lack specialized knowledge in biomedical image classification. Thus, Wisdom of the Crowd (WoC) is used to reduce dependence on individuals and generate highly accurate annotated datasets.

Methods

We study the impact of response mode (binary choice, BC versus elicited beliefs, EB) and recalibration methods on WoC accuracy in a white blood cell annotation task. Annotators (N = 175) were recruited through the DiagnosUS app and were randomized into BC/EB conditions. Annotators decided whether a cell image was a “blast” or a “non-blast”. EB responses were on a 0 to 100 scale. While all images (549) in our data set have expert labels, we split the images into gold standard (GS, 249) and “unlabeled” (300) images to test the efficacy of different WoC labeling methods.

Standard ResNet type classification models were trained using different types of annotation with and without data recalibration. Models were assessed on their accuracy on unseen data.

Results

We had an average of 62.4 BC responses per unlabeled image and 52.2 EB responses per unlabeled image. The average accuracy at the individual level was 0.70 in BC and 0.69 in EB (EB score binarized at 0.5). We examined WoC accuracy for the two response modes after applying recalibration methods (Platt scaling using GS responses) to both individuals and the crowd. Figure 1 shows the WoC accuracies (in red) for different labeled datasets for different crowd sizes.

We then separately trained classification models on these different labeled datasets. As shown in the figure (in blue), models trained on recalibrated EB outperformed models trained on BC (even when BC was recalibrated at the group-level). Further, model accuracy was higher than WoC accuracy, suggesting that the models denoise the crowd.

Conclusions

These results show the importance of response mode and recalibration methods on the accuracy of crowdsourced medical image labels and their impact on AI models trained using these data.

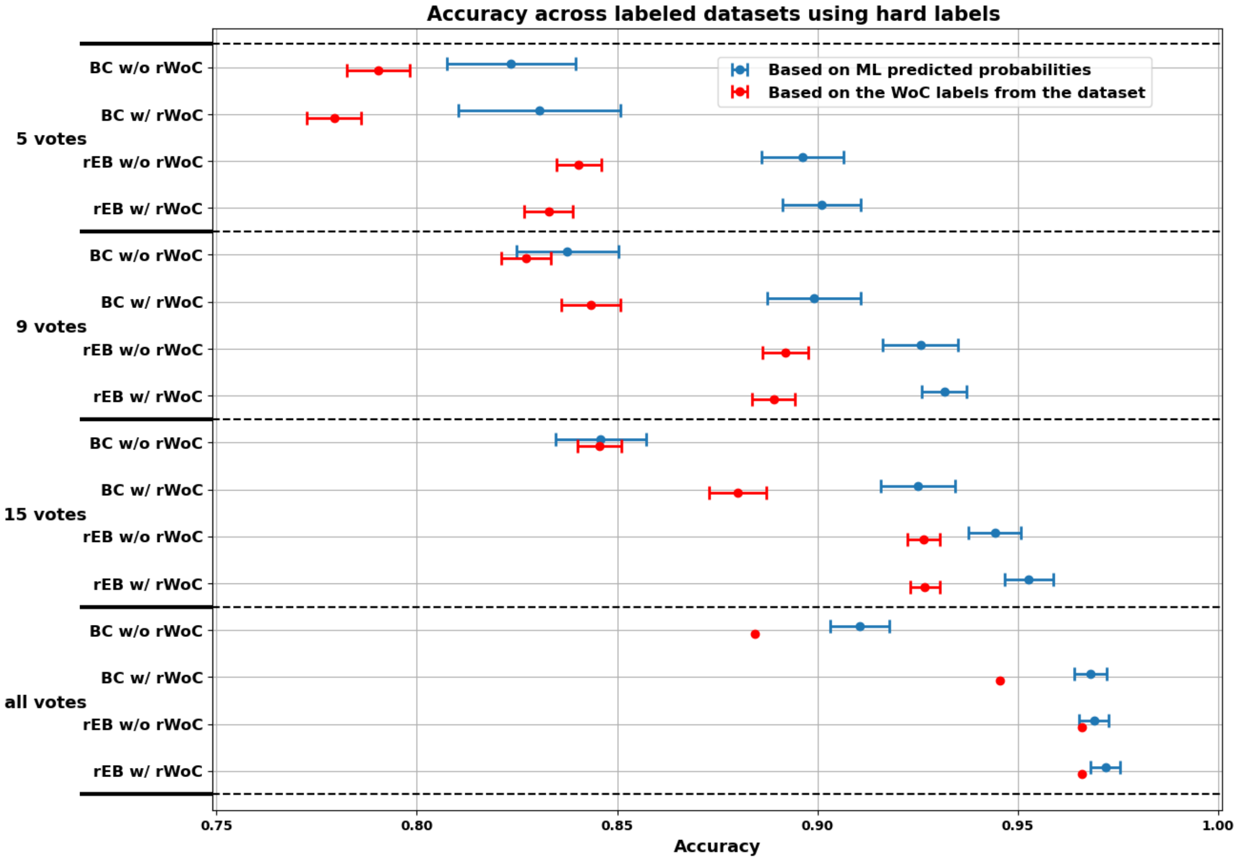


Figure 1: WoC (red) and classification model (blue) accuracies for different labeled datasets and crowd sizes (n = 5, 9, 15, and all). BC = binary choice, rEB = individual-level recalibrated elicited beliefs, rWoC = group-level recalibrated responses.

The impact of dose information on radiographers' subjective interpretation of image quality

Shauna Lane, BSc; Rena Young, MSc; Niamh Moore, MSc;
Andrew England, PhD; Mark F. McEntee, PhD.
Discipline of Medical Imaging & Radiation Therapy, University College Cork, Ireland

Rationale

The increased utilisation of computed tomography (CT) has raised concerns about the associated radiation doses. Optimisation of exposure factors helps to combat this increase and relies of subjective interpretation of image quality (IQ). The impact of prior knowledge of clinical information, for example radiation dose information, has raised questions regarding the possibility of confirmation bias. Such studies have not to the authors' knowledge evaluated confirmation bias in addition to the use of eye tracking. The purpose of this research is to investigate this concept evaluating the impact of displayed dose information on radiographers' subjective interpretation of CT IQ.

Methods

22 CT radiographers agreed to participate, all were attending an international radiology congress. Each observer evaluated 32 axial CT images acquired from an anthropomorphic neonatal phantom, scanned using both a standard dose (STD) and ultra-low dose (ULD) protocol. 16 images contained information on the radiation dose and 16 images contained no radiation dose information. The Tobii Pro Spark eye tracker was used to assess the visual search behaviour of the participants and this was assessed against the presence of dose information and the subsequent IQ evaluations (PGMI system).

Results

When the dose information was present, it was viewed. On average it took longer for the participants to find the dose information than the time spent visualising it. Statistical significance was found when the radiation dose information was present when compared to when it was absent (STD images only). Most IQ scores for STD CT images were scored as moderate while ULD images were scored inadequate.

Conclusions

The use of eye-tracking analysis found that when the dose information is present, it's viewed. Statistical significance was proven for the STD dose group, showing there is a possible effect of radiation dose information on IQ scoring.

WEDNESDAY AFTERNOON, August 7th
Session 3 – 1:00 pm to 2:20 pm

Trials and tribulations of implementing interpretive skills simulation system for resident breast imaging education

LR Salkowski MD PhD[1], AM Fowler PhD MD[1], EA Krupinski PhD[2], PJ Slanetz MD[3], MA Elezaby MD[1]
[1]University of Wisconsin-Madison SMPH, [2]Emory University, [3]Boston University Chobanian & Avedisian School of Medicine

Rationale

Residents have only 12-weeks of breast imaging to become proficient in all modalities and skills. There is no cross training in other rotations nor call experience to support their knowledge development in breast imaging. There is a need for effective methods to enhance trainee interpretive skills in breast imaging. This study explored the feasibility and challenges associated with the implementation of an interpretive skills simulation system for breast imaging into residency education.

Methods

This IRB-approved study used machine learning techniques to construct simulation case sets to be implemented in the 3rd or 4th week of a 4-week breast imaging rotation. Each simulation session had 50 screening mammography cases. Eye tracking tools were also used during the first 10 cases in the set. Psychometric data was collected on subjects (with paid stipend). Performance metrics included diagnostic accuracy, cancer detection rates, and interpretation time.
The six overarching implementation tasks included: developing the educational framework, obtaining funding, acquiring cases, developing solid collaborators, exploring and implementing new technology, acquiring a quiet space to administer simulation sessions.

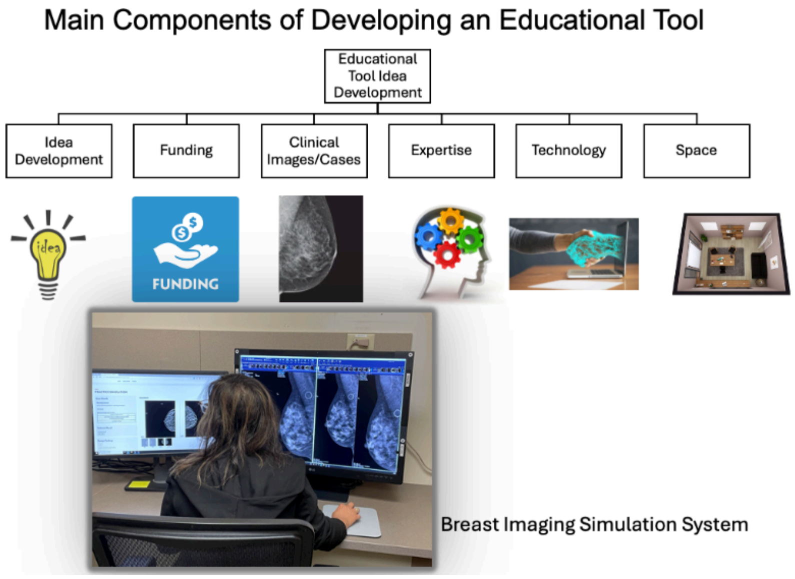
Results

Fifteen (15) of the 19 (79%) invited residents completed the 50-case set screening simulation. The average time to complete the simulation was 4.5 hours (SD=1 hr 6 min). The cancer detection accuracy was 64.71% (SD=7.30). The average sensitivity was 57.64% (SD=13.6) and average specificity was 75.16% (SD=13.0).

Some lessons learned include: keep the IRB up to date with project changes or pivots, use feedback from failed funding attempts to improve your next submission, keep funding office in the loop if there are setbacks, keep collaborators engaged to ensure progress is being made, allow adequate time for image processing, collaborators or other personnel may leave, be prepared for unexpected setbacks, be prepared for delays if hardware/software needs to be assessed for security risks, investigate vendors and their technology, and surround yourself with experts that you can call upon for advice.

Conclusions

The trials and tribulation faced during the implementation of the interpretive skills simulation system underscore the complexity of integrating technology-driven educational tools into medial education. Technical difficulties and onboarding of all stakeholders (including residents) arose. Positive outcomes emphasize the potential benefits of simulating real practice. Systems to enhance radiology education will ultimately improve patient care.


Main Components of Developing an Educational Tool
Educational Tool Idea Development
Idea Development
Funding
Clinical Images/Cases
Expertise
Technology
Space
idea
FUNDING
Breast Imaging Simulation System

Predicting Radiologists' Visual Attention for Mammograms

Jianxun Lou (MSc)[1], Richard White (PhD)[2], Susan Shelmerdine (PhD)[3],
and Hantao Liu (PhD)[1]
[1]School of Computer Science and Informatics, Cardiff University,
Cardiff, United Kingdom
[2]Department of Radiology, University Hospital of Wales,
Cardiff, United Kingdom
[3]Deptartment of Clinical Radiology, Great Ormond Street Hospital,
London, United Kingdom

Rationale

Previous research has demonstrated that medical diagnoses made by radiologists are associated with their visual attention during the interpretation of medical images. Visual saliency measures the degree to which visual content in an image attracts radiologists' attention when interpreting the image. The literature abounds with computational saliency models that are designed predominantly for predicting saliency of natural images under free-viewing conditions. Direct application of these models to medical images is impeded by significant structural disparities between natural and medical imagery, coupled with the differing viewing behaviour observed in free viewing and breast screening. The optimal use of visual saliency in medical imaging algorithms depends on the effectiveness and accuracy of the computational saliency for specific medical applications. Therefore, it is critical to create an eye-tracking dataset for mammograms scanning and develop application-specific saliency models that can reliably capture radiologists' visual attention on medical images. In this study, we have established a dependable eye-tracking dataset representative of radiologists' visual attention patterns during the scanning of mammograms; and constructed predictive computational models capable of automatically estimating radiologists' visual attention during the scanning of mammograms.

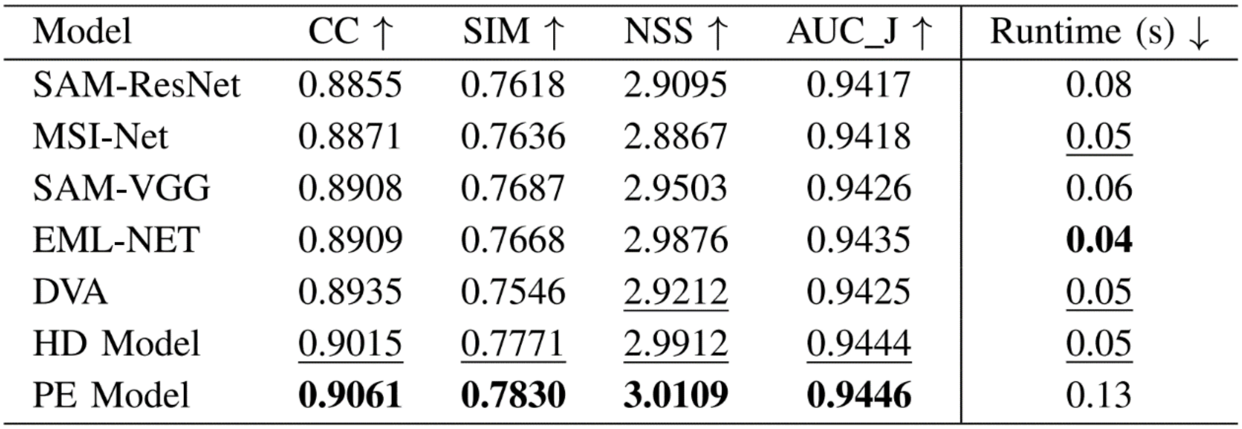
Methods

A large-scale mammogram eye-tracking dataset was established for the study of visual saliency in mammogram reading. The mammographic image set includes 196 mammograms from 98 cases. The eye-tracking data contains the gaze positions of 10 radiologists during the three of reading these mammograms.

We have developed two deep learning-based saliency prediction models, including *HD Model* and *PE Model*, for predicting mammograms' saliency. The *HD Model* utilises high resolution image representations to avoid potential losses of critical information. The *PE Model* adopts a parallel encoder to introduce stronger and more extended image representations for mammogram saliency prediction.

Results

The overall performance of the state-of-the-art saliency models and our models is shown in Fig. 1. The *PE Model* outperforms all other models across all saliency evaluation metrics and *HD Model* achieves the second-best performance. These findings reveal that saliency models specifically developed for mammograms more accurately predict the visual attention of radiologists during mammogram scans than those developed for natural images. The results also demonstrate that the *HD Model* not only outperforms other models (except the *PE Model*) in saliency prediction metrics, but also exhibits competitive inference speed.



Model	CC ↑	SIM ↑	NSS ↑	AUC_J ↑	Runtime (s) ↓
SAM-ResNet	0.8855	0.7618	2.9095	0.9417	0.08
MSI-Net	0.8871	0.7636	2.8867	0.9418	0.05
SAM-VGG	0.8908	0.7687	2.9503	0.9426	0.06
EML-NET	0.8909	0.7668	2.9876	0.9435	**0.04**
DVA	0.8935	0.7546	2.9212	0.9425	0.05
HD Model	0.9015	0.7771	2.9912	0.9444	0.05
PE Model	**0.9061**	**0.7830**	**3.0109**	**0.9446**	0.13

Fig. 1. Performance comparison of on the mammogram eye-tracking dataset. Bold font and underline indicate the best and 2nd best performance; "Runtime" indicates the seconds needed to infer a single mammogram on a NVIDIA GTX 1080 GPU.

Conclusions

In this study, we constructed an eye-tracking dataset of radiologists when scanning mammograms and developed visual saliency models aimed at predicting radiologists' visual attention during mammogram scanning. Experimental results demonstrate that our models, tailored specifically for this task, achieve state-of-the-art performance.

Breast Cancer Survivors' Perceptual Map of Breast Reconstruction Appearance Outcomes

Haoqi Wang,[a] Xiomara T. Gonzalez,[b] Gabriela A. Renta-Lopez,[a] Mary Catherine Bordes,[c]
Michael C. Hout,[d,] Seung Choi,[e] Gregory Reece,[c] Mia K. Markey[a,f,*]
[a]The University of Texas at Austin, Department of Biomedical Engineering, Austin, USA
[b]The University of Texas at Austin, Department of Electrical and Computer Engineering, Austin, USA
[c]The University of Texas MD Anderson Cancer Center, Plastic Surgery, Houston, USA
[d]New Mexico State University, Psychology Department, Las Cruces, USA
[e]The University of Texas at Austin, Department of Educational Psychology, Austin, USA
[f]The University of Texas MD Anderson Cancer Center, Imaging Physics, Houston, USA

Rationale

There is often a communication gap between healthcare providers and patients in consultations about reconstruction procedures to mitigate congenital or acquired appearance differences. It is hard for patients to articulate how they expect their appearance to be changed by reconstruction. There are prevalent misunderstandings regarding likely appearance outcomes and many patients develop unrealistic expectations of how they will look during and after reconstruction. Our overarching goal is to develop a tool to help breast cancer survivors visually express what they expect to look like after reconstruction. This study aims to comprehensively understand how breast cancer survivors perceive diverse breast appearance states by mapping them onto a low-dimensional, interpretable Euclidean space.

Methods

We conducted an observer study with breast cancer survivors (N = 25) using an incomplete block design to assess the pairwise similarities among clinical photographs of torsos (M = 100) depicting a range of torso appearances relevant to breast reconstruction. Observers' visual similarities were collected using the spatial arrangement method (SpAM) [1] with 10 photographs at a time. We used multidimensional scaling to visually represent the photographs in a latent space, referred to as a "perceptual map," in which the Euclidean distance between pairs of photographs indicates the perceived similarity as assessed by observers. The perceptual map was interpretated by identifying associations between latent dimensions in the survivors' perceptual map and the language of breast morphologies commonly employed by surgeons to describe their conscious assessments of breast reconstruction outcomes.

Results

Interpretation of the perceptual map (Figure 1A), constructed in two dimensions, revealed that observers' comparisons of breast appearance states was associated with the number of nipples present in the photographs (Figure 1B). Using a support vector machine (SVM), we partitioned the map into three regions representing 0 (blue), 1 (gray), and 2 (red) nipples with 85% accuracy. Linear regression was performed on all photographs together and separately within each of the three regions to investigate associations between latent dimensions and continuous breast morphology variables. For instance, in the region corresponding to photographs depicting two nipples, the ratio of breast size to body size (i.e., sum of breast volumes/BMI) increases from top to bottom (Figure 1C), explaining 23% of the variance ($R^2 = 0.23$) in our sample.

Conclusions

We conducted observer studies to establish a perceptual map of breast cancer survivors' perceptions of breast appearance states. Analysis of the perceptual map identified factors associated with survivors' perceptions of breast appearance states that should be emphasized in the appearance consultation process. In the future, the perceptual map can be used to identify photographs that visually represent what an individual patient expects to look like, thereby informing personalized consultation to address unrealistic expectations. This study also lays the groundwork for evaluating interventions intended to help patients form realistic expectations. Since a patient's appearance expectation can be quantified as coordinates defining a location in the perceptual map, the impact of an intervention can be quantified by the pre-post change in the location of the patient's expectation in the perceptual map.

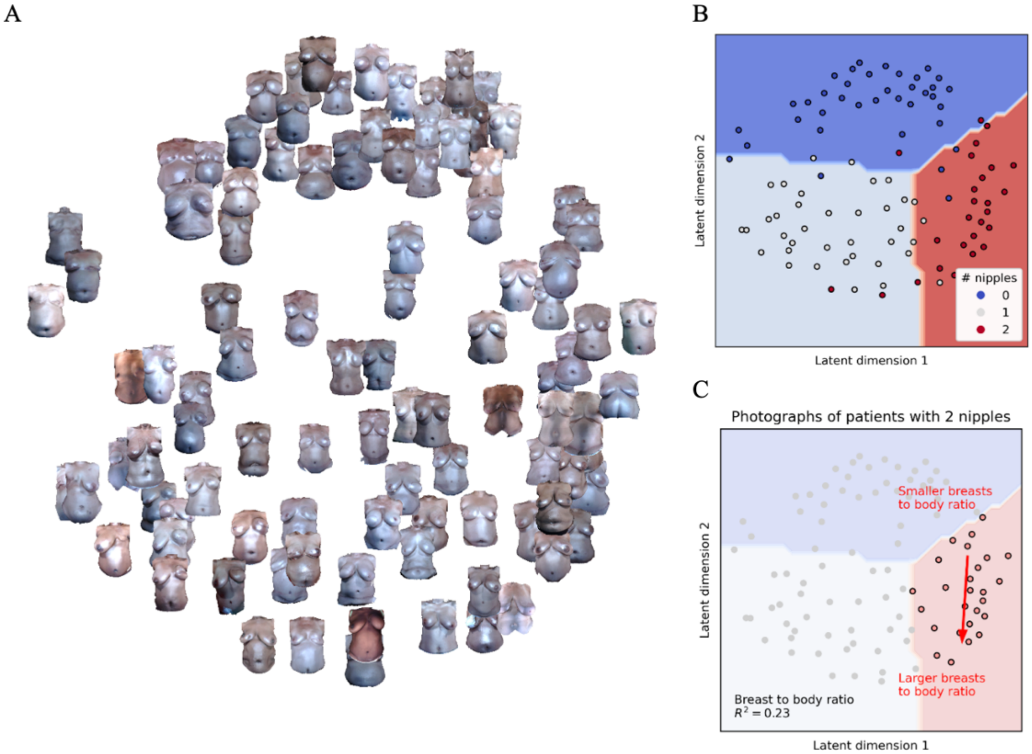


Figure 1. (A) We used multidimensional scaling to visually represent the photographs in a latent space, referred to as a "perceptual map," in which the Euclidean distance between pairs of photographs indicates the perceived similarity as assessed by observers. (B) Cluster analysis on the latent dimensions to investigate categorical variables. SVM partitioned the space into 3 regions (blue, gray, and red) based on the number of nipples that the patient had at the time the photograph was taken (0, 1, or 2). (C) Latent dimension analysis of breast size. Linear regression shows that for the map region corresponding to photographs depicting two nipples, the ratio of breast size to body size (i.e., sum of breast volumes/BMI) increases as one moves from the top of the map to the bottom of the map.

WEDNESDAY AFTERNOON, August 7th
Session 4 – 2:40 pm to 3:40 pm

Using a Limited Field of View to Improve Nodule Perception and Reduce Eye Strain

William F. Auffermann MD/PhD*[1], Rishabh Agarwal MD [1], Samual K. Zenger MD [2]
[1] Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, UT, USA.
*[2] Department of Radiology, University of Vermont, Burlington, VT, USA. *Corresponding Author*

Rationale

Medical images contain much information that may distract an observer from their perceptual task and search pattern. For example, when evaluating a chest radiograph (CXR) for pulmonary nodules, potentially distracting information may be present outside the lung and hinder identification of nodules. A prior pilot study sought to examine this by using CXR images with a limited field of view (LFoV) to focus the observer's attention on the lungs. Observers anecdotally commented that the LFOV images were overall less bright and easier to look at for longer periods of time. The goals of this study are: 1) Further examine the effects of presenting CXR images with a LFoV on perception of pulmonary nodules, and 2) Use a questionnaire to assess for changes in symptoms of eye strain when using standard versus LFoV images.

Methods

A total of 40 healthcare trainees (6 radiology residents and 34 medical students) participated in this IRB approved study. Subjects were randomly split into control and experimental groups, 20 subjects each. All subjects were introduced to the RadSimPE radiology simulator software, pulmonary nodules, and a lung search pattern. Both groups were then shown a set of 20 CXRs, half of CXRs contained a pulmonary nodule. Participants estimated the probability of a nodule being present using a 5-point receiver operating characteristic (ROC) scale, marked the nodule, and indicated their confidence in localization. Then the experimental group received training on identifying nodules on LFoV CXRs, while the control group received a journal article as an attentional control. For case-set 2, the control and experimental groups were shown similar images, the difference was that the experimental group LFoV CXRs were masked to exclude the chest wall and abdomen. Participants repeated the same image evaluation tasks. At the end of the study, participants were given two surveys: one contained 8 items to assess their thoughts about the perceptual education, the second contained 4 questions related to symptoms of eye strain.

Results

Both experimental and control groups showed an improved ability to identify pulmonary nodules on CXRs, but the experimental group performed better, Δ mean ROC area under curve (experimental-control) = 0.07, pooled standard deviation = 0.11, Bootstrap p-value = 0.02. P-values < 0.01 for all survey items, indicating that: 1) subjects preferred this perceptual education to conventional radiology education, and 2) participants felt the LFoV decreased symptoms related to eye strain.

Conclusions

Using limited field of view images may improve perceptual performance during focused high yield perceptual tasks and decrease symptoms of eye strain.

Assessing visual hindsight bias in radiologist eye tracking metrics

Jacky Chen, Ziba Gandomkar, PhD, Warren Reed, PhD

Discipline of Medical Imaging Science, The University of Sydney

Rationale

Approximately 74% of diagnostic errors stem from cognitive factors (1, 2) , leading to misdiagnoses and subsequent adverse patient outcomes that may trigger medicolegal litigation. Hindsight bias, also known as the "knew-it-all-along effect" , occurs where individuals with outcome knowledge underestimate the challenge of reaching the initial diagnosis (3). Clinicians, referencing prior imaging, may question oversights in earlier diagnoses. Awareness of the pitfalls of hindsight bias, addressed through education or peer auditing, can potentially mitigate this bias (3, 4). An investigation in 2020 (3), demonstrated the effect of visual hindsight bias on radiologist perception. This study will analyse the eye tracking data collected in the previous study, exploring the perceptual and cognitive processes exhibited by radiologists when encountering hindsight bias.

Methods

The methodology of the 2020 study (3) involved sixteen radiologists interpreting 15 postero-anterior chest images that contained solitary lung nodules with 25 incremental levels of blurring. Participants were asked to initially identify the nodules by decreasing the blurring and then were asked to increase the level of blurring until the nodule was de-identified. Participants then repeated the experiment but were provided with education on hindsight bias and were asked to counteract these effects. Eye tracking technology was used to capture radiologist eye movements and search patterns.

Results

The preliminary results, as depicted in Figure 1, show the eye tracking metrics, including search times, dwell times, and time to first fixation (TTF), across the different experimental phases.

Notably, there are discernible variations in these metrics between foresight and hindsight conditions, indicating the influence of hindsight bias on radiologist visual behaviour. For instance, the average time to first fixation during the foresight phase was substantially longer compared to the hindsight phase, suggesting a more deliberative and thorough examination process when hindsight bias is absent. Similarly, the total fixation duration was markedly different between foresight and hindsight conditions, underscoring the impact of bias on the duration and intensity of visual scrutiny. While further data analysis is ongoing, these preliminary findings underscore the importance of understanding radiologists' cognitive processes and decision-making in the context of hindsight bias. Moreover, the potential role of education in mitigating bias effects and enhancing diagnostic accuracy warrants further exploration, with implications for medical training and practice.

Conclusions

This study contributes to understanding radiologists' cognitive processes, peripheral processes and decision-making in the presence of hindsight bias, highlighting the potential impact of education on mitigating bias effects and improving diagnostic accuracy.

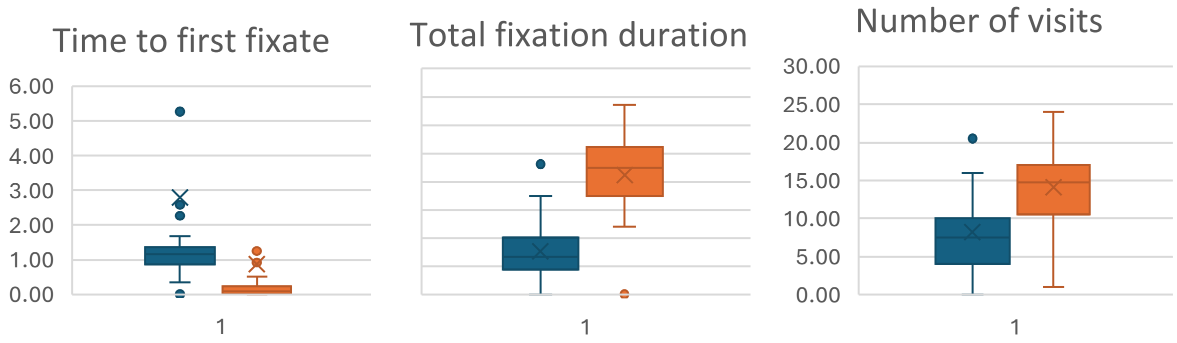


Figure 1- The blue and orange box plots depict the distribution of the "foresight" and "hindsight" conditions. For each metric, the median value from all readers was calculated, and the paired Wilcoxon signed rank test was utilized to explore if each metric differed between the two conditions. The p-values for Time to first fixate (p = 0.00002), Total fixation duration (p<0.00001), and Number of visits (p = 0.00013).

Mitigate, don't litigate: Dealing with "Look but Fail to See" errors in Radiology

Jeremy M. Wolfe, PhD, *Harvard Medical School, Brigham & Women's Hospital, Boston, Massachusetts*

John D. Banja, PhD, *Center for Ethics, Emory University, Atlanta, GA*

Stephen A. Waite, MD, *Department of Radiology, SUNY Downstate Medical Center, Brooklyn, NY*

Brian Sheppard, SJD, LLM, JD, *Seton Hall University School of Law Newark, NJ*

Elizabeth A. Krupinski, PhD, *Department of Radiology and Imaging Sciences, Emory University, Atlanta, GA*

Rolf Dieter Hollstein, MD, MPH, *Advanced Radiology Services, Grand Rapids, MI*

Michael A. Bruno, MD, MS, *Department of Radiology, Penn State Milton S. Hershey Medical Center, Hershey, Pennsylvania*

Radiologists, even expert radiologists, make false negative errors. Sometimes they miss targets because the image is poor or the diagnosis is unclear. However, not infrequently, the missed finding is "retrospectively visible", clearly identifiable once it is pointed out. Eye tracking can be used to categorize errors of this sort as "search errors", where the radiologist never fixated close enough to the target to detect it, and "recognition errors", where the eyes did land on or near the target but left again without leading to identification of the target. Together, these two types of error can be labeled as "perceptual errors" as distinguished from "decision errors" where the expert found the target but failed to correctly classify it. Perceptual errors are obviously not desirable. They can lead to harm to the patient and they can lead to legal consequences for the radiologist. Perceptual errors are thought to account for 60 to 80 percent of medical malpractice suits brought against radiologists. The logic of those cases is superficially appealing. If a clinician misses a clearly visible finding, surely that is evidence of negligence.

There is no doubt that radiologists can be negligent. However, we wish to push back against the argument that simply missing a retrospectively visible finding is evidence of negligence. Many of these errors are radiologic examples of a class of Look but Fail to See (LBFTS) errors so common that Wolfe, Kosovicheva, and Wolfe (2022) argued that they should be considered a form of "normal blindness". Outside of the radiology reading room, these errors manifest themselves as everything from typos that escape even vigilant proofreading to gorillas who wander undetected through ball games in Simons and Chabris' famous experiment. LBFTS errors have multiple causes. These include 1) capacity limitations on how many items can be processed with each fixation of the eyes, 2) "misguidance", where expectations or biases lead to selection of the wrong stimuli, and 3) incomplete processing, where an item may be successfully selected but abandoned before it can be successfully identified. All of these factors are biologically-based responses to the limited capacity of our visual system. They are not a matter of choice. Given those limitations, any “reasonable” doctor would be expected to miss things from time to time. If doctors exercise the care, skill, and judgment of “reasonable” doctors in their field they are not supposed to be at fault in a legal sense. The problem is that the legal profession and, even more so, the juries have not been persuaded that LBFTS errors can be reasonable.

In this talk, we will briefly review the roots of LBFTS errors and their application to medical image perception and we will discuss the changes that are needed in the medicolegal system.

THURSDAY MORNING, August 8th
9:00 – 10:00 am

Keynote Address: How domain-general object recognition may help us understand medical expertise

Isabel Gauthier, PhD

Department of Psychology, Vanderbilt University

THURSDAY MORNING, August 8th
Session 6 – 10:50 am to 12:10 pm

Gamification for Image Perception and Emergency Radiology Education

Soham Banerjee MD*[1], William F. Auffermann MD/PhD [2]
*[1] Department of Radiology, Baylor University, Houston, TX, USA. *Corresponding Author*
[2] Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, UT, USA.

Rationale

Radiology education outside the reading room has traditionally focused on learning via printed media, in-person lectures, and case conferences. There have been advances in medical education, especially using gamification. Gamification involves activities that are enjoyable and stimulate participants to engage with the activity, often used in education. Gamification has found increasing use in medicine, with relatively few applications in radiology to date. We have developed a general gamified framework intended to help healthcare trainees perceive abnormalities on medical images related to emergency radiology. The goal of this study is to determine if this gamified method of radiology education facilitates participants' perception of abnormalities on medical images containing urgent abnormalities.

Methods

Twenty medical students participated in this IRB approved study. Subjects were randomly assigned to control (n=10) and experimental (n=10) groups. All subjects were oriented to the

RadSimPE radiology workstation simulation software. Both groups were then shown the first set of 20 cases, a mixture of radiograph and CT cases, half contained an urgent medical finding. Participants estimated the probability of an acute abnormality being present using a 5-point receiver operating characteristic (ROC) scale, marked the location of the abnormality, and indicated their confidence in localization. The experimental group then received gamified perceptual education using our ‘Stab-the-Diagnosis’ program, hosted on the internet Unity gaming platform (unity.com). The control group received an attentional control journal article for the same period of time. Both groups were then shown a second set of 20 images and performed the same image evaluation tasks. At the conclusion of this study, subjects were given a questionnaire containing 8 items to examine how the gamified education compares with conventional radiology educational methods.

Results

The experimental group was more confident in their localization of abnormalities, mean difference in confidence between case sets 1 and 2 = 0.054, p-value = 0.008. The control group confidence in localization did not change significantly. There was no significant improvement in ability to identify abnormalities in either group. Survey results were generally positive with p-values < 0.05 for 6 of 8 questions.

Conclusions

These results suggest that gamification may be a useful adjunct to conventional methods of radiology education and improve perception of image abnormalities.

Global gist signal differs between specific genetic subtypes of breast cancer

Karla K. Evans. PhD, Roisin Bradley, MD
Psychology Department, University of York, England
York & Scarborough Teaching Hospitals NHS Foundations Trust, England

Rationale

An important component of expertise in radiography that allows for early non-invasive detection of cancer is the visual recognition skill of the radiologist. Studies have shown that this perceptual assessment of a mammogram also involves a rapid detection of a "global gist signal" that can indicate the presence of breast cancer even in the absence of a visibly actionable lesion. Detection of this signal relies on the characteristic image statistics. Genetic subtypes of cancer are treated differently and there is evidence to suggest that different subtypes might produce different image-based phenotypes. Bayesian Artificial Neural Network (BANN) algorithm can distinguish the appearance of the parenchyma in patients with or without BRCA1/2-related breast cancer. As with the global gist measure, these genetic effects are not only restricted to the breast that has overt signs of cancer. In the present study we test the hypothesis that the global gist signal is dependent on the genetic subtype of cancer a patient has and/or the possibility that all forms of cancer alter the texture of the cancerous breast.

Method

We used a rapid gist assessment paradigm conducted with two groups of expert medical observers testing mammograms from cases where the subtype of cancer was known and normal cases cancer free for the last 10 years. For the two experiments, we used a set of 30 cases from patients with each of the four main classifications of tumour: positive for oestrogen receptors (ER+), progesterone receptors (PR+), expression of the HER2 protein, or negative for all of these (triple negative). In addition, cases with BRCA1/2 mutations and a set of patients with microcalcifications proven by biopsy to be associated with cancer. After brief (500 ms) viewing of unilateral mammograms observers were asked to rate the likelihood of abnormality of the images on the 0-100 scale. Observers viewed and evaluated 180 cases in each experiment.

Results

Across two experiments we find expert medical observers can extract the global gist of abnormality from all six different genetic subtypes of cancer but to a varying degree of success (Figure 1a 7b). The differences in the strength of the signal are not due to breast density differences of the mammograms with different subtypes but signal the possibility that different cancer result in different image-based phenotypes present across the entire image.

Conclusions

Our findings indicate that all forms of cancer alter the texture of the cancerous breast but to a varying degree. Resulting in different subtypes of cancer being associated with phenotypes that can be perceived during the early extraction of the global gist of the abnormal. Early knowledge of the risk of the possible subtype of cancer if fully differentiated could allow for better understanding of the global gist signal and its feasibility in using it for targeted testing and treatment increasing the likelihood of survival.

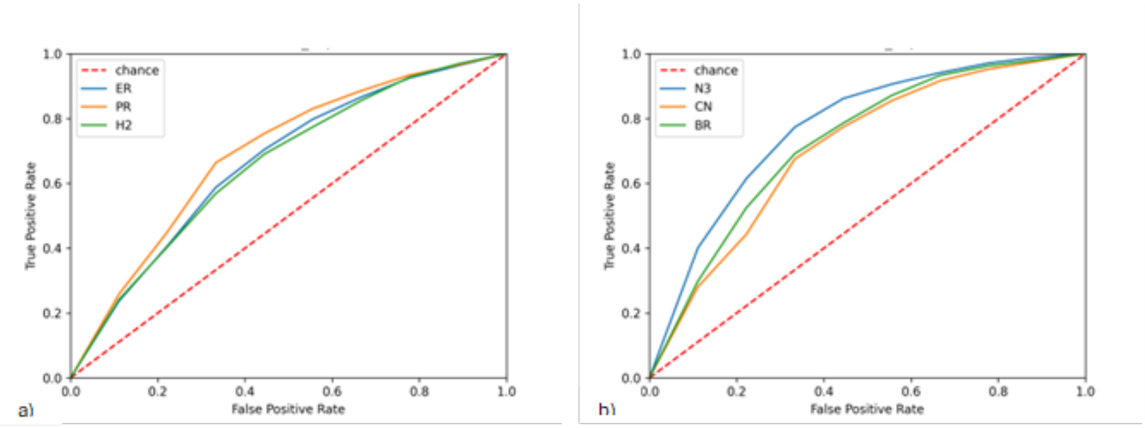


Figure 1. *Roc curves showing performance on rapid gist assessment experiment* ***a)*** *with normal and mammograms with cancers being positive only for oestrogen receptors (ER+), progesterone receptors (PR+), or expression of the HER2 protein; b) with normal and mammograms with cancer negative for all receptors (triple negative N3), BRCA1/2 mutations and a set with microcalcifications proven (CN) by biopsy to be associated with cancer.*

The impact of radiation dose information on the subjective evaluation of CT image quality

Conor Lee, MSc; Niamh Moore, MSc; Rena Young, MSc, Lisa Kingston, MSc; Andrew England, PhD; Mark F. McEntee, PhD.
Discipline of Medical Imaging & Radiation Therapy, University College Cork, Cork, IE

Rationale

The use of computed tomography (CT) continues to increase substantially. Concerns exist regarding exposure to ionising radiation from CT. Ultra-low dose (ULD) CT represents one of the solutions to this issue. Adoption of ULD protocols has the disadvantage of increased image noise, diminished image quality (IQ) and potential impact on diagnostic yield. ULD CT protocols, therefore, are, most appropriate in anatomic areas with high inherent contrast such as the lungs or in clinical settings where IQ reduction will not decrease diagnostic yield, for example in monitoring selected disease states. Formal IQ comparisons between CT protocols are not new; however, such studies report high inter-observer variability. Confirmation bias may be one cause of such variability. This study investigates, if the addition of radiation dose information to CT images results in increased magnitude of confirmation bias.

Methods

Two CT datasets were created by scanning a paediatric anthropomorphic phantom twice. Scanning included a standard (STD) and ULD CT protocols (DLP 22.92mGy and 1.49mGy, respectively). A group of radiographers were individually invited to score the IQ from a range of images within each dataset using the ‘PGMI’ scoring system. Images were presented to participating radiographers either with or without the dose information present ($CTDI_{vol}$ / DLP). Data were analysed using MS Excel and SPSS.

Results

Fifty-three participants were included. Twenty-five evaluated the dataset with the radiation dose information visible, and 28 evaluated the dataset without the dose information. Review of STD CT images with dose information not visible resulted in 23% and 22% higher IQ scores for cortical and trabecular bone, respectively (**Figure 1**). When ULD CT images were reviewed in the absence of dose information, 7% and 4% higher IQ scores for cortical and trabecular bone were observed (**Figure 1**).

Conclusions

Study data suggests that the addition of radiation dose information created a confirmation bias in datasets containing STD and ULD images. A larger impact was seen for the STD images. Observers rate IQ lower when radiation dose information is visible than when they were absent. The current work hypothesises this is due to confirmation bias. Seeing low exposure factors confirms for the user that the image is low IQ. Further studies evaluating IQ should consider the impact of radiation dose information on perceptual scores. Studies which report subjective IQ scores should report whether reviewers were blinded to radiation exposure indices and acquisition parameters during image review.

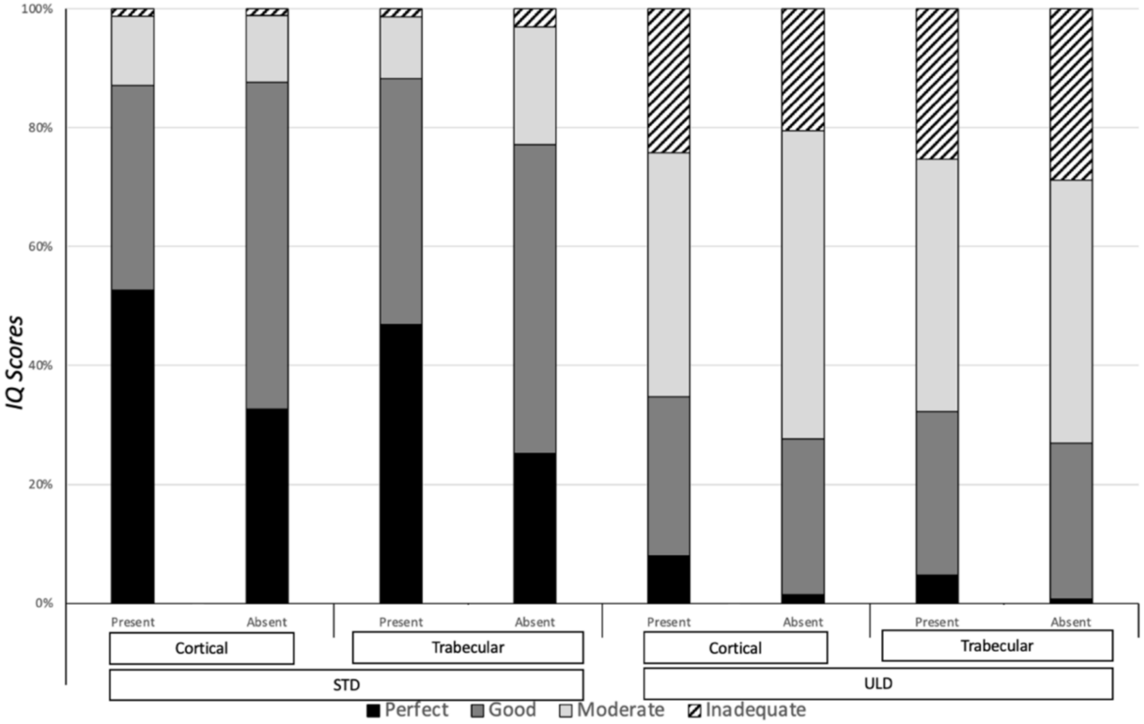


Figure 1. Bar chart summarising the categorical characterisation of IQ across the two CT datasets. STD – standard CT protocol (22.92mGy); ULD – ultra-low dose CT protocol (1.49mGy).

Identifying the Textural Components of the Global Gist Signal of the Abnormal

Cameron Kyle-Davidson, PhD, Lyndon L. Rakusen MSc, Karla K. Evans, PhD
Department of Psychology, University of York

Rationale

Expert radiologists can detect the presence of abnormality in mammograms even when those mammograms are only presented for fractions of a second, far faster than the time it would take to detect and localise a lesion. This ability extends to contralateral mammograms, mammograms acquired years before any visible lesion, and even textural patches from the abnormal breast. This implies that radiologists are extracting the “gist of the abnormal” from the mammogram; a signal composed of global, low-level image statistics which informs them of the presence of abnormalities. This signal most likely originates from the textural make-up of the parenchyma. We investigated whether we can identify the textural components which constitute this gist signal, and which radiologists are using to inform their judgements of abnormality, by computationally analysing tens of thousands of mammograms.

Methods

We first preprocess the OMIDB2 dataset to remove unsuitable mammograms. This resulted in 66264 usable mammograms from 8368 patients. We divide these mammograms into obvious abnormality, subtle abnormality, normal, and prior groups then compute three distinct textural metrics for each group. We compute: 1.) Frequency-power spectra, to identify frequency differences 2.) Radiomic features 3.) Simoncelli steerable-pyramid decompositions to account for differences at different scales. We then identify how these metrics can be used to classify normal/abnormal mammograms via Support Vector Machines (SVMs).

Results

There is no difference in average frequency spectra between the normal and abnormal groups. However, trained SVMs are able to identify which group a mammogram belongs to at an above-chance level. We find the most “useful” frequencies lay in the mid-band, and within a region that has been shown to reduce radiologist performance if removed from the mammogram (Figure 1). With a similar process we identify the most important radiomic features, and find that maximising said features leads to the appearance of calcification-like features. Steerable pyramid analysis reveals that significant differences in textural composition arise in the mid-size structures, rather than in the very fine structure of the parenchyma.

Conclusions

We analysed the most significant textural differences that arise between normal and abnormal mammogram parenchymas, finding that the most important features contributing to the difference tend to be mid-size components of the texture, suggesting global differences arise neither at the very fine, or very coarse scale. This is reinforced by evidence from steerable pyramid analysis, which shows a similar pattern, while radiomic features appear to be highly sensitive to the presence of calcifications, also a mid-level textural feature. Together, this narrows the likely driving components behind the gist signal to within the mid-size textural structures of the parenchyma.

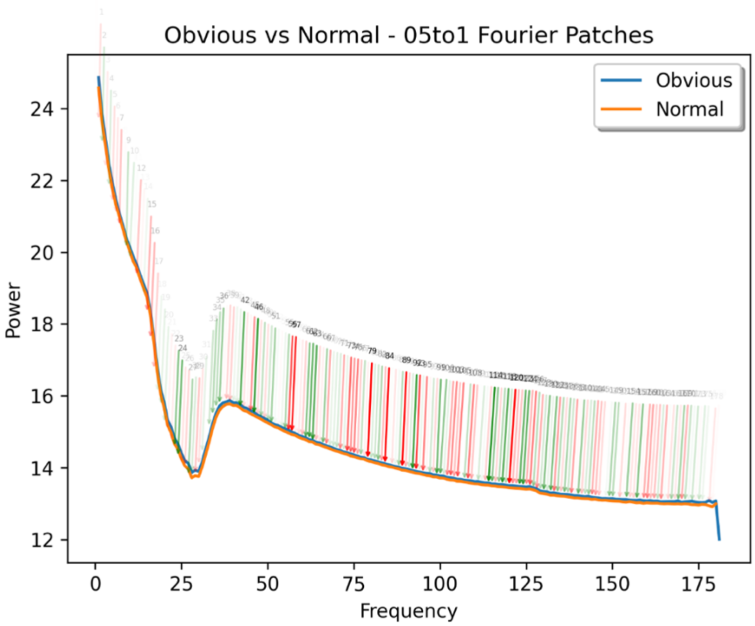


Figure 1: *Frequency-power spectra for a bandpassed mammogram with frequency power artificially decreased. This decrease leads to poorer performance in radiologists. The importance of each frequency to the SVM is shown by the opacity of each arrow. Arrow colour indicates positive or negative coefficient.*

THURSDAY AFTERNOON, August 8th
Session 7 – 1:20 pm to 2:40 pm

Prediction and generalization performance for ramp-spectrum forced-localization tasks

Craig K. Abbey[1], Frank W. Samuelson[2], Rongping Zeng[2], Kyle J. Myers[3], John M. Boone[4], and Miguel P. Eckstein[1]
[1]*Dept. of Psychological and Brain Sciences, University of California Santa Barbara, CA, USA*
[2]*Division of Imaging Diagnostics and Software Reliability, United States Food and Drug Administration, Silver Spring, MD, USA*
[3]*Puente Solutions, LLC, Phoenix, AZ, USA*
[4]*Departments of Radiology and Biomedical Engineering, University of California, Davis, CA, USA*

Rationale

Design of medical imaging systems requires important decisions to be made in the workings of scanner hardware or processing of acquired data. A longstanding goal of model observer studies is to provide guidance in these decisions using a relatively simple computation that is readily implemented in early stages of system development. These models typically involve the detection or discrimination of a "signal" profile that is considered non-random (i.e. fixed). It remains an open question whether these simple models are sufficient for predicting performance in far more complex clinical tasks performed by human observers in which the appearance of an abnormality can assume a variety of forms, referred to generically as signal uncertainty.

Thus, the enduring question confronting model observers is whether they generalize to more complex tasks. In this work we evaluate this generalization in terms of location uncertainty in the context of ramp-spectrum noise that mimics the acquisition noise in CT systems.

Methods

We re-analyze data from a publication (1) that explored human-observer performance in ramp-spectrum noise using forced-localization tasks. The studies evaluated 24 conditions that encompassed 3 factors: signal size, background variability, and apodization. Human-observer performance ranged in efficiency (30% to 80%) with respect to the ideal observer, indicating suboptimal performance of human observers that is dependent on imaging conditions. Our technical goal is to assess suboptimal model observers to see if they can accommodate observer performance across the various factors in the forced localization data.

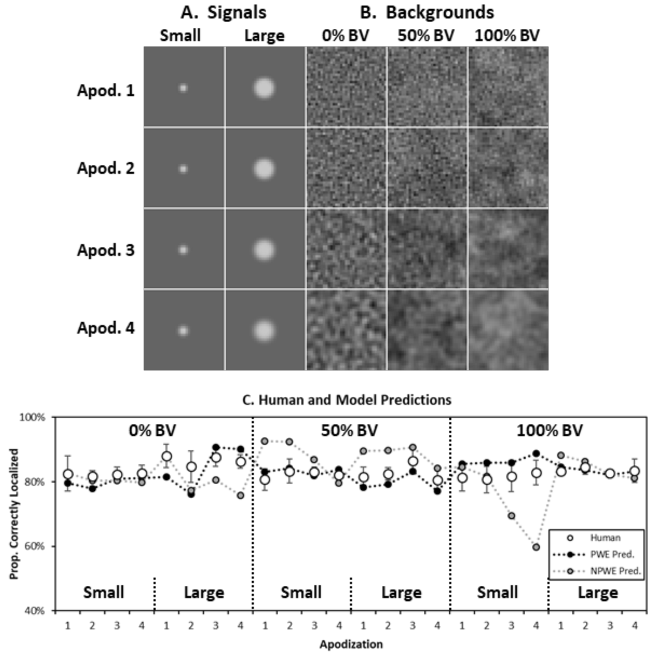
To accomplish this, we parameterize model observers based on eye-filtered prewhitening (PWE) and non-prewhitening (NPWE) matched filters. The NPWE model (2) includes an eye-filter with a frequency profile that has peak sensitivity as a free parameter. The model includes a free parameter, β, for a proportional internal-noise term. The PWE model (3) also contains a proportional noise term, with an additional free parameter for an intrinsic internal noise term that has a $1/f^2$ spectrum. The proportion correct in the localization task is estimated using the M^* approach of Burgess (4), in which signal uncertainty is characterized by an approximate number of independent locations. Prediction performance for these models is estimated using a leave-one-out methodology.

Results

Figure 1 summarizes results to be described in the presentation. It gives examples of the signals and noise textures magnified and cropped for display (Fig. 1A), shows prediction performance across conditions (Fig. 1B), and tabulates overall performance. We find notably better performance with the PWE model, consistent with previous results (1) demonstrating adaptation to background statistics in these tasks.

Conclusions

A simple and traditional approach to model observers appears to generalize reasonably well to localization tasks that involve search. We find that the PWE model best predicts localization performance in these studies with a deviation that is roughly equivalent to the standard deviation across human subjects.



D. Overall Performance

Err.	NPWE	PWE
Dev.	5.4%	3.07%
Dev./SD	188.1%	103.6%

Figure 1. Signal (A) and noise samples (B) show the target and backgrounds used in the forced-localization tasks. Human observer performance (C) is plotted across the 24 experimental conditions along with leave-one-out predictions of the models.

Multi-reader multi-case AUC analysis methodology for studies involving Artificial Intelligence Applications

Stephen L. Hillis, PhD
Departments of Radiology and Biostatistics, University of Iowa

Rationale

As artificial intelligence (AI) applications become more frequently employed, it is important to be able to statistically evaluate the performance of such systems when used by themselves compared to the performance of human readers using the AI system as an aid, as well as with the performance of unaided human readers.

Methods

In this talk I discuss how the Obuchowski-Rockette (OR) method, which treats both cases and human readers as random samples, can be easily adapted for comparing the usefulness of the following three modalities: (1) AI standalone; (2) AI-unaided human readers; and (3) AI-aided human readers. Simulation results are presented.

Results

Comparison of the performance of human readers with the performance of AI-aided human readers can be accomplished by a straightforward application of the OR method using a paired design where the same readers assess images with and without using AI as an aid. In contrast, comparison of AI standalone with AI-aided human readers, or with AI-unaided readers, can be accomplished by treating the difference of the reader and AI standalone performance outcomes as single treatment outcomes, and then computing a confidence interval for these differences using the OR method appropriate for one treatment, as described by Hillis (2014, Statistics in Medicine). Chakraborty (2018, *Observer Performance Method for Diagnostic Imaging,* Chapter 22) describes how to do this, gives a real-data example, and provides R software for computing the confidence interval.

However, Chakraborty’s software is the only software that I know of that is specifically designed to compare AI standalone with AI-aided or AI-unaided human readers. In this talk I describe a “workaround” involving a straightforward rearrangment of the data that allows the confidence interval for the difference (discussed above) to be easily computed using standard OR software, thus eliminating the need for specialized OR software. A real-data example is presented to illustrate this method.

Validation of the OR method for comparing stand-alone AI with aided or unaided human readaers is provided through a simulation study where ratings are simulated from an appropriate adaption of the Roe and Metz simulation model. We find that for typical situations, the OR method shows acceptable performance in terms of type I error.

Conclusions

Although comparing AI-aided human reader performance with AI-unaided human reader performance can be accomplished by a straightforward application of conventional OR software, there is little OR software designed for comparing AI standalone performance with that of either AI-aided or AI-unaided human readers. I have discussed and illustrated use of a workaround that makes it easy to use conventional OR software for these comparisons, and have validated the approach with a Roe and Metz type simulation study.

A Dual-Branch Deep Learning Model for MRI Image Quality Assessment

Yueran Ma[1] (MSc), Jean-Yves Tanguy[2] (MD), Hantao Liu[1] (PhD)
[1]School of Computer Science and Informatics, Cardiff University, United Kingdom
[2]Department of Radiology, Angers University Hospital, France

Rationale

Magnetic Resonance Imaging (MRI) is pivotal in medical diagnostics, offering non-invasive, high-contrast imaging of the body's internal structures, crucial for diagnosing various conditions, especially in soft tissues. Unlike X-rays and CT scans, MRI's non-radiographic nature and detailed imaging capability present unique challenges for image quality assessment (IQA), including noise, motion blur, and artifacts, which can compromise diagnostic accuracy. Although IQA algorithms have excelled in natural image processing, their adoption in medical imaging remains limited due to the complexity of medical images, data sensitivity, and the need for clinical validation. This paper introduces a deep learning-based IQA model, evaluated on the RADIQMRI database, showcasing its efficacy and potential in enhancing MRI image clarity, optimising scan parameters, and potentially improving patient comfort and diagnostic quality in healthcare.

Methods

The RADIQMRI database, developed at the University Hospital of Angers, France, encompasses 112 medical images derived from eight sets spanning six anatomical areas, each with two energy levels of artifacts. This database categorises common artifacts into four groups based on their structure and spectral density, serving as a foundation for our research in image quality assessment (IQA) of medical imaging.

CNNs and Transformers, pivotal in natural image processing, bring complementary strengths to this task. CNNs are adept at extracting local features, essential for analysing detailed medical imagery, while Transformers excel in modelling global image contexts through long-range dependencies.

To harness these strengths, we propose a dual-branch model for medical IQA, merging CNNs for local detail with the global context modelling capabilities of Swin Transformers. The Swin Transformer is notable for its hierarchical, scalable architecture, facilitating efficient processing of detailed medical images without the computational burdens of traditional Transformers. Feature fusion in our model is achieved through deformable convolution, which aligns and integrates features from both branches, further enhanced by an adaptive parameter adjusting the contribution ratio between CNN and Swin Transformer features. This ensures a balanced, comprehensive feature set for IQA. Our prediction module employs a refined two-branch patch-wise mechanism, incorporating direct scoring and spatial attention to calculate a final weighted score.

Results

We present an evaluation of several state-of-the-art deep learning IQA models from the natural image domain, alongside our proposed model, using the RADIQMRI database. Additionally, we conduct ablation studies to explore the impact of incorporating or omitting the CNN branch, including a scenario where only the CNN branch is considered, as well as the effect of substituting different CNN backbone networks. The experimental results are shown in Table 1, demonstrating the robustness and efficacy of our proposed dual-branch model with adaptive branch feature proportion parameter.

Table 1. Performance Evaluation and Ablation Studies of Deep Learning Models on the RADIQMRI Database for Image Quality Assessment.

	PLCC	SROCC
simpleCNN	0.5979	0.3148
VGG19	0.6909	0.0989
ResNet34	0.7048	0.4544
swin-T	0.9014	0.8885
MANIQA	0.9147	0.9012
simpleCNN+swin-T+AP (ours)	0.9208	0.9294
VGG19+swin-T+AP (ours)	0.9233	0.9206
ResNet34+swin-T+AP (ours)	**0.9320**	**0.9289**

Results include evaluations of single and dual-branch models with and without the adaptive parameter (AP) for feature scaling. AP is utilised in our dual-branch models to dynamically balance the contribution of features from the CNN and Swin Transformer branches.

Conclusions

In this study, we have introduced a novel dual-branch deep learning model for image quality assessment (IQA) that leverages the strengths of both CNNs and Swin Transformers, augmented by an adaptive branch feature proportion parameter. Our comprehensive evaluation, including a series of ablation studies conducted on the RADIQMRI database, has demonstrated the superior performance of our model compared to existing deep learning-based approaches in the natural image IQA domain.

Exploring the Explainability of a Machine Learning Model for Prostate Cancer: Do Lesions Localize with the Most Important Feature Maps?

Destie Provenzano MS[1], Murray Loew PhD[1], Shawn Haji-Momenian MD[2]
1. George Washington University School of Engineering and Applied Science
2. George Washington University School of Medicine and Health Sciences

Rationale

As the popularity of artificial intelligence in medical imaging has grown, so has the need for better explainability metrics. Many explainability methods exist, but debate persists as to how to best evaluate them. This study evaluated a metric for model explainability in prostate cancer by assessing whether the most important feature maps generated from a CNN classifier (ResNet) localized to the same region of abnormality on the image as identified by a diagnostic radiologist.

Methods

The ProstateX dataset from The Cancer Imaging Archive was used to build a series of ResNet models that classify prostate magnetic resonance (MR) images (axial T2 sequence) as containing either clinically significant (CS) or not clinically significant (NCS) cancer. All 63 images with CS lesions were used, and 63 images with NCS lesions were randomly selected to create a balanced dataset of 126 lesions. Two image sets were created: 1) the full axial image of the pelvis which includes the prostate ("Full"), and 2) segmented image of the prostate only ("Seg") by a radiologist. Two types of predictive models: (1) All layers ("From Scratch") and (2) only the last layer ("Transfer Learning") were then trained and tested (80/20 split) on both datasets to generate four total models and eight total test sets. A shuffle test and 5-fold cross- validation was used to assess statistical significance. The central coordinates of each lesion were then compared to the locations of pixels in the highest-weighted feature maps. Final models were assessed for classification accuracy and lesion localization.

Results

All models were able to statistically significantly initially classify CS vs NCS. The experiments were: (A) From Scratch Full/Full: 89%, (B) Transfer Full/Full: 90%, (C) From Scratch Seg/Seg: 97%, and (D) Transfer Seg/Seg: 92%. Performance dropped, however, when models were tested on the respective other datasets: (A) From Scratch Full/Seg: 52%, (B) Transfer Full/Seg: 56%, (C) From Scratch Seg/Full: 46%, and (D) Transfer Seg/Full: 52%. Across all training/ testing combinations, 98% of lesions localized with the feature maps generated from the "Transfer

Learning" models. The model trained "From Scratch" on the full dataset localized with the lesion 21% of the time whereas the model trained on the segmented dataset localized 86% of the time. Validation for any "From Scratch" model on the Full dataset tended to not localize (15%) whereas validation on segmented data tended to localize regardless of training type (91%).

Conclusions

Transfer Learning models and models trained or tested on segmented data better localized overall on this limited dataset. This methodology presents a potential way to determine whether a classification model is considering the appropriate anatomical regions in prostate cancer, which could better explain, and inspire trust in, these "black-box" models.

THURSDAY AFTERNOON, August 8th
Session 8 – 3:00 pm to 4:00 pm

The Perceptual Bias Cascade in the Medical-AI Information Value Chain

Jennifer S. Trueblood, PhD; Andrew Caplin, PhD; Gunnar Epping; William R. Holmes, PhD; Daniel Martin, PhD

Psychological and Brain Sciences, Indiana University; Department of Economics, New York University; Psychological and Brain Sciences, Indiana University; Mathematics and Cognitive Science, Indiana University; Department of Economics, University of California Santa Barbara

Rationale

The integration of AI systems into the medical image interpretation process holds great promise, but these systems reside within a human-centric information value chain. This chain starts with humans annotating biomedical images. Because expert annotators are costly, companies aggregate the decisions of multiple non-experts to generate image labels (harnessing a phenomenon known as Wisdom of the Crowds, WoC). The resulting labeled image sets are used to train machine learning (ML) models, which are ultimately used by diagnosticians in clinical settings. We investigate a phenomenon we term the "perceptual bias cascade”, occurring when cognitive and perceptual biases propagate through the medical-AI information value chain.

Methods

The key to using WoC for image annotation is that individual errors cancel when aggregated. Challenges arise due to correlated errors stemming from shared biases among annotators, such as prevalence-induced biases (specifically, the low prevalence effect). This study (N = 600 MTurk participants) investigates the cascade of individual biases through crowdsourced labels in a white blood cell annotation task. We used a 2 prevalence rate (50%, 20%) x 2 response mode (binary choice, BC; elicited beliefs, EB) between-subjects design. After a learning stage, the main task consisted of 5 blocks of 20 trials where participants decided whether a cell image was a “blast” or a “non-blast”. The prevalence rate was stated at the start of each block. EB responses were on a 0 (non-blast) to 100 (blast) scale.

Results

In BC, the miss rate at 20% prevalence (M= 51.5%) was higher than at 50% prevalence (M = 35.4%). Likewise in EB (binarized so that responses > 50 were classified as blasts), the miss rate at 20% prevalence (M = 41.6%) was higher than at 50% prevalence (M = 28.8%). We also observed a WoC low prevalence effect, with elevated miss rates in the 20% prevalence conditions (M = 46.0% in BC, M = 43.9% in EB) as compared to the 50% prevalence conditions (M = 15.2% in BC, M = 11.1% in EB).

Using the 50% prevalence conditions, we generated four sets of crowdsourced labels: 2 response modes (BC, EB) x 2 label types (hard, soft). Using the 20% prevalence conditions, we generated eight sets of crowdsourced labels: 2 response modes (BC, EB) x 2 label types (hard, soft) x 2 bias levels (corrected, not corrected). Bias correction involved changing the WoC classification threshold. We then separately trained a convolutional neural network (CNN) on these 12 labeled datasets. CNNs trained on labels without bias correction had much lower test accuracy than those that had been corrected, due to elevated misses (see Figure 1).

Conclusions

Results illustrate how biases in crowdsourced image labels can propagate through ML outputs.

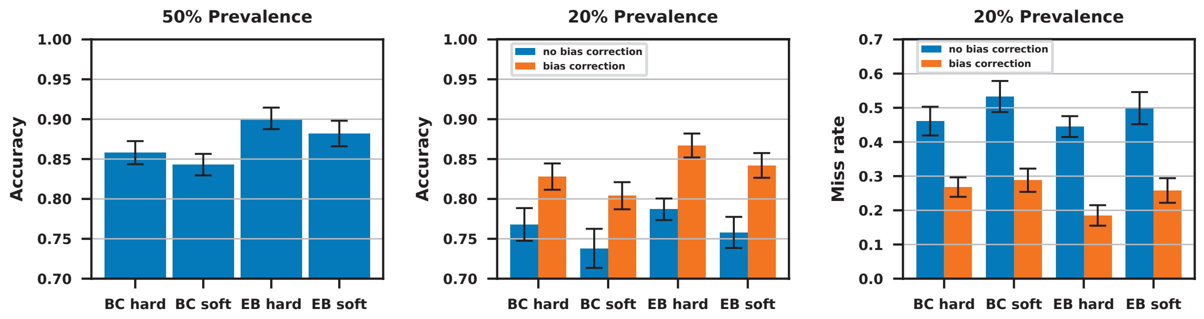


Figure 1: Test accuracy for CNNs trained with different types of crowdsourced labels (BC = binary choice, EB = elicited

Medical Image Processing by Artificial Intelligence Software: Outcome Quality Vetting by Human Observers

Jay Hegdé, PhD [1]; Nicholas J. Tustison, PhD [2]; William T. Parker, MD [3]
[1] Department of Neuroscience and Regenerative Medicine, Medical College of Georgia, Augusta University, Augusta, GA 30912, USA, [2] Department of Radiology and Medical Imaging, School of Medicine, University of Virginia, Charlottesville, VA 22903, USA; [3] Department of Radiology and Imaging, Medical College of Georgia, Augusta University, Augusta, GA 30912

Rationale

Artificial intelligence (AI) software, especially deep neural network (DNN) software, can be highly effective in processing medical images to enhance the usefulness of the images for clinical practitioners. While a variety of metrics have been developed to evaluate the physical quality of the processed images, the extent to which these metrics comport with the subjective percepts of the human observers remains largely unclear. Here we describe a suite of methods to address this issue using super-resolution (SR) images of the adult human hand as an illustrative example.

Methods

Software for creating SR images of the hand, in which each output image has a substantially higher resolution that the corresponding input image, are not hitherto unavailable. We used the powerful ANTsX software platform to create, train, and test a Residual Neural Network (ResNet), a DNN architecture that has proven effective in creating SR images in other contexts. We used the fully trained ResNet to generate SR versions of hitherto unseen, clinical-quality magnetic resonance (MR) input images of the adult human hand. To compare the physical quality of the input *vs*. output images, we used two different established metrics, PSNR (peak signal-to-noise ratio) and SSIM (structural similarity index measure). We also studied the subjective perception of image quality by human observers. To do this, we asked naïve, non-professional subjects ($N = 12$) to view each given pair of input *vs.* output images, and report the perceived difference of image quality using an on-screen slider. We systematically compared the physical *vs.* subjective metrics of image quality using the representational similarity analysis (RSA) within and across subjects.

Results

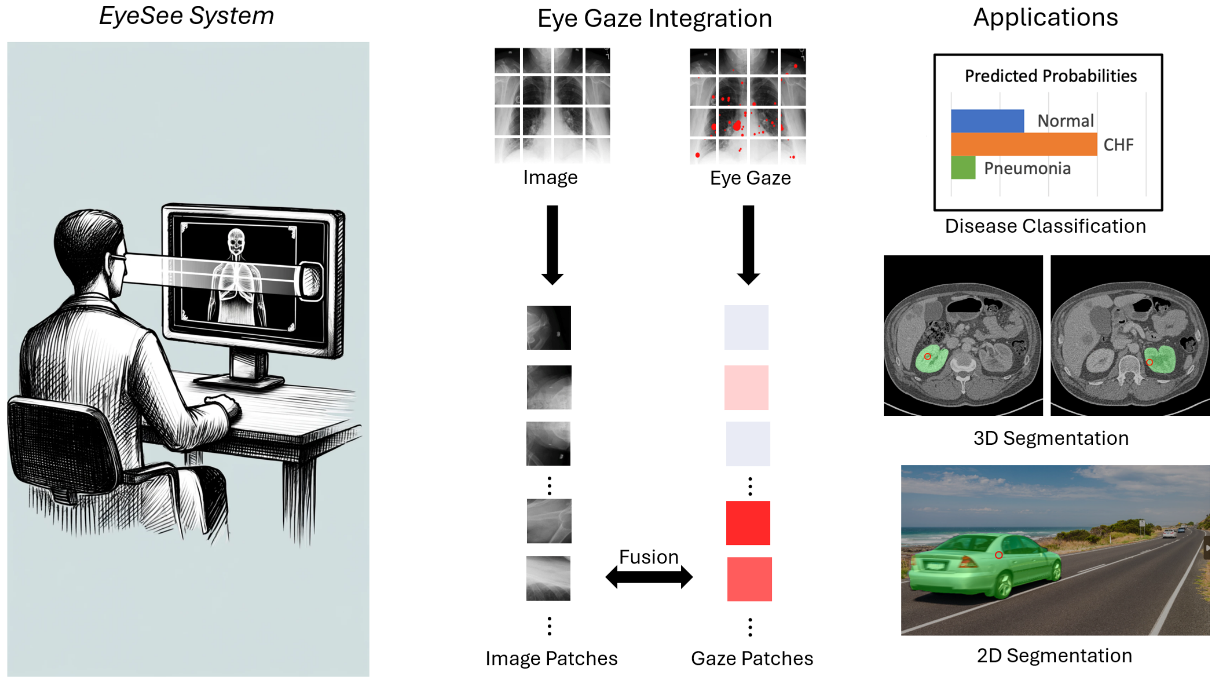
As expected, the resolution of the output images was significantly higher than the corresponding input images by both PSNR and SSIM for each individual subject (paired one-tailed *t*-tests, $p <$ 0.05, corrected for multiple comparisons) indicating that our software was highly effective in producing SR images of the hand. The results of within-subject RSA revealed that PSNR and SSIM were each correlated significantly with the corresponding subjective percepts ($p < 0.05$, randomization analysis, corrected). Similar results were obtained using RSA across all subjects, although SSIM was significantly better correlated with the reported percepts across the subject sample than PSNR (p < 0.05, randomization analysis, corrected).

Conclusions

The perceived quality of SR images depends more on global image features such as object similarity than more local, pixel-level comparisons at least for non-professional subjects. More generally, our methods represent a set of principled, quantitative tools for measuring the effectiveness of AI tools in processing medical images.

EyeSee: Integrating Eye Tracking with Deep Learning for Improved Medical Image Analysis

Bin Wang and Ulas Bagci, Ph.D.
Machine and Hybrid Intelligence Lab, Departments of Radiology, Biomedical Engineering, and Electrical and Computer Engineering, Northwestern University



Rationale

Traditionally, Eye Tracking technology can be used to record radiologists' eye movements to identify the regions they are looking at during the reading. Integrating this technology with real-world clinical workflows still presents several challenges: real-world screening environments, absence of 3D analysis capabilities, and difficulties in seamlessly integrating deep learning-based medical image analysis. We address these challenges by providing a real radiology room screening experience with minimal to no distraction for radiologists and integrating eye-gaze patterns with cutting-edge deep learning solutions for radiology applications. We exemplify the proposed end-to-end system with classification and segmentation tasks in medical image analysis: 1) real-time auto-segmentation of organs and tumors with gaze prompts and SAM (segment anything model), and 2) newly developed graph convolution deep learning with gaze-embedded attention mechanisms to improve diagnostic decisions from Chest X-Ray images.

Methods

In this study, we developed ***EyeSee***, an eye tracking-based medical viewer system with automatic diagnosis capabilities. The *EyeSee* viewer allows us to capture the eye gaze data of radiologists during scan reading. It integrates key functionalities of the PACS viewers including zooming, scrolling, contrast adjustment, measurement tools, and support for different medical imaging formats. After the reading is finished, the collected eye gaze information is processed and input into the trained deep learning algorithms for (1) diagnosis or (2) image analysis. *EyeSee* supports both 2D and 3D scans (ultrasound, X-ray, CT, PET, MRI, etc.). For diagnostic enhancement, *EyeSee* automatically crops the images into patches and extracts the deep image features of each patch, meanwhile, we apply the proposed time-aggregation method to calculate each patch's eye gaze time duration feature. Then, *EyeSee* combines two features together to conduct medical image classification or segmentation. A public Chest X-ray dataset with eye gaze was used in the experiments, which included 1083 cases with three different labels (Normal, Congestive Heart Failure, and Pneumonia). We also used CT, MRI, and X-Ray, as well as computer vision images for real-time segmentation purposes.

Results

We employed 10-fold cross-validation to validate our diagnostic results in X-Ray disease classification, achieving an accuracy of 83.18% with an inference time of just 0.353 seconds per patient. This demonstrates that our method not only boosts diagnostic accuracy but also minimizes time consumption, enhancing the efficiency of integrating eye tracking into diagnostic readings. Furthermore, we evaluated various deep learning algorithms and found that graph neural networks outperform both convolutional neural networks and transformer-based models in this application. Our segmentation strategy is based on combination of gaze prompts with SAM for real-time segmentation of objects, videos are attached for visual examples.

Conclusions

In this study, we built ***EyeSee*** system that integrates eye tracking technology into medical image analysis smoothly to simulate a real radiology room experience in real time. By utilizing the expert knowledge hidden in the eye gaze collected, our system enhances diagnostic accuracy with impressive efficiency. ***EyeSee*** viewer supports multiple modalities from 2D to 3D images. In the future, we aim to expand the application of eye tracking across all medical modalities generally.

FRIDAY MORNING, August 9th
Session 9 – 9:00 am to 10:20 am

Training ophthalmologists to recognize novel retinal markers identified using artificial intelligence

Ipek Oruc, PhD, Gulcenur Ozturan, MD, Lei Yuan, BSc
Ophthalmology and Visual Sciences, University of British Columbia
Neuroscience, University of British Columbia

Rationale

Retinal images are commonly used to diagnose and manage ocular diseases. Retina can also reveal clues regarding the status of cardiovascular, neurological and systemic health, though many often overlooked, or perhaps presently unknown. In recent work, we proposed a methodology to extract novel retinal characteristics from a deep learning model trained to classify fundus images. Using this methodology, we uncovered previously unidentified retinal features that show differences between females and males, a patient trait that is not currently recognized by ophthalmologists in this modality. Here, we examine whether human observers can learn to recognize patient sex in fundoscopic images.

Methods

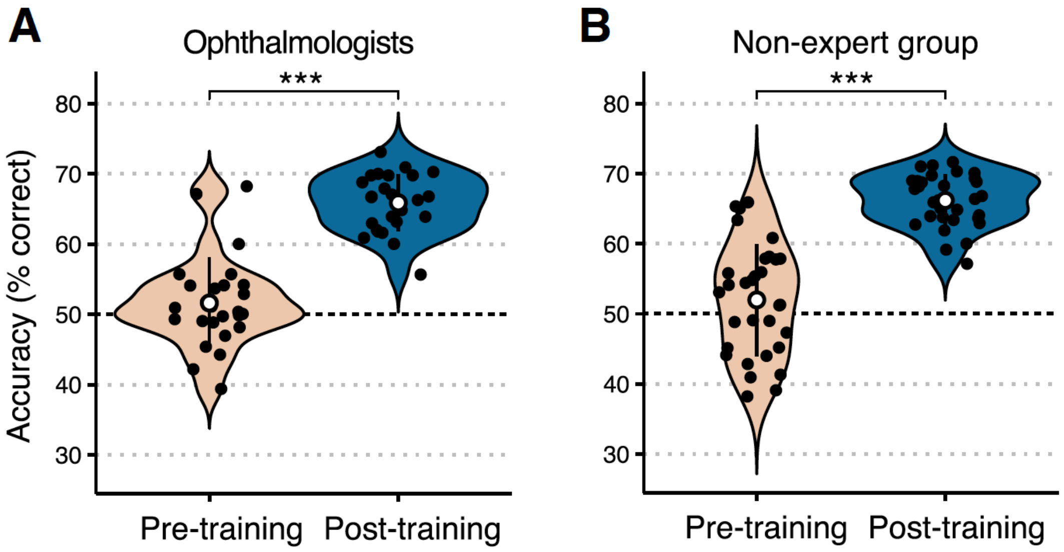
We developed a training paradigm that consisted of didactic and practical components. In the didactic component, participants viewed a short presentation of descriptions of retinal characteristics that have been found to differ between males and females (e.g., brighter peripapillary region in females, greater vascular prominence in the superior temporal quadrant in males) as well as visual illustrations of how these might present in fundoscopic images. In the practical component, participants completed 50-trials of a sex-recognition task in which they chose the male image among a male-female pairs in a two-alternative forced-choice (2-AFC) paradigm. Feedback was provided which highlighted the correct choice. A separate block of 200 2-AFC trials without feedback was used to assess sex-recognition performance. This latter test did not use any images that were seen in the practice component. Participants also completed a novel object memory test (NOMT) to assess general object recognition ability. Twenty-six participants in the expert group (all ophthalmologists; 16 females; age: *M*=34.85) and 31 naive participants with no experience regarding fundus images (18 females; age: *M*=34.00 years) took part in the study.

Results

Pretraining sex-recognition accuracy was *M*=51.62% for the expert ophthalmologist group, and *M*=51.97% for the nonexpert group, which did not differ from chance level (50%) in the expert ophthalmologist group (*p*=0.22) and in the nonexpert group (*p*=0.18), and performance did not differ between the two groups (*p*=0.86). Post-training, there was a significant increase in sex recognition accuracy in the expert ophthalmologist group (M = 65.89%, $p \ll 0.001$, $d = 2.64$), and in the nonexpert group ($M = 66.16\%$, $p \ll 0.001$, $d = 2.28$). Performance in the NOMT was not correlated with post-training improvement in sex recognition accuracy in the expert group (*r*=−0.08, *p*=0.7) and in the nonexpert group (*r*=0.1, *p*=0.58). However, NOMT performance was correlated with performance during the training block for the expert group (*r* = 0.6, *p*=0.001) and the nonexpert group (*r*=0.39, *p*=0.028)

Conclusion

These results empirically show that ophthalmologists do not recognize patient sex in fundus images, yet they can be trained to do so. Future work with this AI-powered approach can be extended to discover and add novel signs of systemic and neurodegenerative disease in retinal images to the toolkit of ophthalmologists.



Accuracy in the 2-AFC sex-recognition task is shown for the pretraining and post-training blocks for the expert ophthalmologist group A) and the nonexpert group B).

Using Expert Gaze for Self-Supervised and Supervised Contrastive Learning of Glaucoma from OCT Data

Wai Tak Lau, MS, Ye Tian, MS, Roshan Kenia, BS, Saanvi Aima, Kaveri A. Thakoor, PhD

Departments of Computer Science, Biomedical Engineering, and Ophthalmology,

Columbia University

Rationale

Machine Learning has grown in importance in clinical workflow; however, one major challenge is that it is hard to acquire a large quantity of high-quality labeled data. We address this challenge of limited data availability common in healthcare settings by using clinician (ophthalmologist) gaze data on optical coherence tomography (OCT) report images as clinicians diagnose glaucoma, a top cause of irreversible blindness world-wide. Gaze data offers a wealth of information regarding the focus of attention and the expertise level of individuals examining medical reports. In addition, gaze data from experts is especially abundant in medicine. In this work, gaze data is used to extract information about the OCT report being viewed via our Transformer-based (deep learning architecture widely employed in natural language processing and Generative AI) model we call 'GazeFormerMD'. We use the extracted information - pseudo-labels generated using gaze representations - to inform an image classification model to aid learning when labeled data is scarce.

Methods

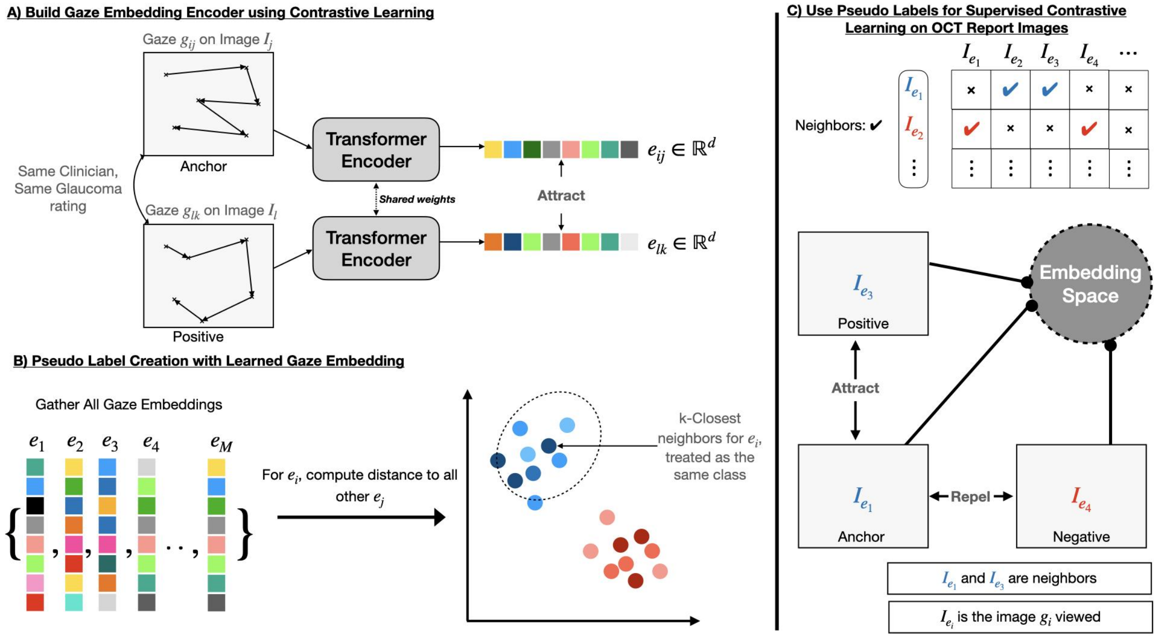
We directly learn gaze representations with our 'GazeFormerMD' model to generate pseudo-labels using a novel multi-task objective, combining triplet and cross-entropy losses. The pseudo-labels encode class information of the OCT report being looked at by a given ophthalmologist, and the triplet loss enables generalizability when combined with the cross-entropy loss. We use these pseudo-labels for weakly-supervised contrastive learning (WSupCon) to detect glaucoma from a partially-labeled dataset of OCT report images using ResNet50 as our backbone. We apply our method on our dataset of 177 OCT reports and ophthalmologist 467 gazes sequence, on an unseen dataset of OCT reports without gaze, and on the EGD-CXR (chest x-ray) dataset with 1083 radiologist gaze-image pairs [1].

Results

Our natural language-inspired region-based-encoding GazeFormerMD model pseudo-labels, trained using our multi-task objective, enables downstream glaucoma detection accuracy via WSupCon exceeding 91% even with only 70% labeled training data. Furthermore, a model pretrained with GazeFormerMD generated pseudo-labels and used for linear evaluation on an unseen OCT-report dataset with 6941 images achieved comparable performance to a fully-supervised, trained-from scratch model while using only 25% labeled data. Similarly, GazeFormerMD achieves 81.57% test accuracy at classifying the presence of pathologies, performing competitively with only 50% of data from the EGD-CXR dataset.

Conclusions

We conclude that gaze data contains valuable information that can be used to enhance disease classification, especially in a setting where labeled data is lacking. Future work will explore ways to pretrain on larger datasets of gaze data, as well as integrating gaze and images with other modalities (e.g., such as text).


A) Build Gaze Embedding Encoder using Contrastive Learning
Gaze g_{ij} on Image I_j
Anchor
Same Clinician, Same Glaucoma rating
Gaze g_{lk} on Image I_l
Positive
Transformer Encoder
Shared weights
Transformer Encoder
$e_{ij} \in \mathbb{R}^d$
Attract
$e_{lk} \in \mathbb{R}^d$
B) Pseudo Label Creation with Learned Gaze Embedding
Gather All Gaze Embeddings
e_1 e_2 e_3 e_4 e_M
For e_i, compute distance to all other e_j
k-Closest neighbors for e_i, treated as the same class
C) Use Pseudo Labels for Supervised Contrastive Learning on OCT Report Images
I_{e_1} I_{e_2} I_{e_3} I_{e_4} ...
Neighbors: ✓
I_{e_1}
I_{e_2}
Embedding Space
I_{e_3}
Positive
Attract
I_{e_1}
Anchor
Repel
I_{e_4}
Negative
I_{e_1} and I_{e_3} are neighbors
I_{e_i} is the image g_i viewed

	I_{e_1}	I_{e_2}	I_{e_3}	I_{e_4}	...
I_{e_1}	×	✓	✓	×	×
I_{e_2}	✓	×	×	✓	×
⋮	⋮	⋮	⋮	⋮	⋮

Figure: Schematic of GazeFormerMD pseudo-label generation pipeline.

Eye-Tracking to Evaluate AI Use

Elizabeth A. Krupinski, PhD
Department of Radiology & Imaging Sciences Emory University

Rationale

AI can work synergistically with humans to improve efficacy and efficiency of data analysis and interpretation by human decision makers. However, we know little about optimal ways to present AI output.

Methods

5 radiologists and 3 residents read chest images with eye-tracking to decide COVID present or absent and rate severity as mild, moderate, or severe in 5 conditions: no AI, single Word noting severity of COVID or normal, Heat map, probability Graph, or Heat map + Graph.

Results

There was no difference in the distribution of decisions ($X^2 = 11.90$, $p = 0.7511$) made on the first (no AI provided) viewing of the images no matter what type of AI after the first decision. There was an overall significant difference ($X^2 = 11.60$, $p = 0.0206$) between radiologists and residents, with radiologists having lower FP and higher TN rates than the residents. Overall, there was a significant difference ($X^2 = 289.74$, $p < 0.0001$) in the distribution of confidence levels for each decision type with both radiologists and residents having more probable confidence levels for TP, TPW (TP wrong severity), FP, and FN decisions, and more definite confidence levels for TNs. The data were analyzed for significant differences in changes in decisions between first no AI view vs with AI: a) no change, b) positive change due to AI, and c) negative change due to AI. Overall, there was no significant difference in the no change group as a function of AI type ($X^2 = 0.80$, $p = 0.8489$). There was a significant difference overall in the positive change group ($X^2 = 21.78$, $p < 0.0001$). The FN-TP group approached significance (X2 = 7.58, p = 0.0555). There was a significant difference ($X^2 = 25.06$, $p < 0.0001$) in the negative change group. If TP-TPW is not considered negative (as COVID was detected correctly but severity incorrect), the negative group is not significantly different ($X^2 = 3.32$, $p = 0.3449$). There were no significant differences between radiologists and residents for no change ($X^2 = 0.27$, $p = 0.9655$) or positive change ($X^2 = 3.93$, $p = 0.2690$), but there was for negative change ($X^2 = 12.52$, $p = 0.0058$). Overall, the was no significant difference for the no change group between AI types ($X^2 = 1.03$, $p = 0.7928$) or the positive change group ($X^2 = 0.20$, $p = 0.9777$). There was a significant difference ($X^2 = 8.19$, $p = 0.0423$) for the negative change group. There were no significant differences overall between radiologists and residents for no change in confidence ($X^2 = 2.33$, $p = 0.5073$) or negative change ($X^2 = 0.83$, $p = 0.8413$) but there was for positive change ($X^2 = 14.78$, $p = 0.0020$).

Conclusions

Our results demonstrate that the form of the AI output is important as it can impact clinical decision making and efficiency. Perhaps of equal importance is that it may impact different users in different ways, especially trainees versus those with more experience.

False-Color Images to Facilitate Image Perception and Radiology Education

William F. Auffermann MD/PhD *[1], Soham Banerjee MD [2]
[1] Department of Radiology and Imaging Sciences, University of Utah, Salt Lake City, UT, USA.
*[2] Department of Radiology, Baylor University, Houston, TX, USA. *Corresponding Author*

Rationale

A subset of medical image abnormalities may not be perceived due to suboptimal differences in contrast between target and background. Most radiology images are saved and viewed in grayscale. The human eye is able to perceive approximately 64 grayscale levels. This limitation in grayscale perception may contribute to a subset of perceptual errors in radiology. However, the human eye is able to differentiate over 1 million different colors. Consequently, converting grayscale images to false-color may assist with perception and learning for low contrast lesions, such as ground-glass nodules in the lung on chest computed tomography (CT). The goal of this study is to determine if using false-color chest CT images improves participants' perception of ground-glass nodules.

Methods

A total of 19 medical students participated in this IRB approved study. Subjects were randomized to control (n=10) and experimental (n=9) groups. Control and experimental subjects were oriented to the RadSimPE radiology workstation simulation software. All participants were then shown the first set of 20 chest CTs, half contained a ground-glass nodule. Participants marked the location of the ground-glass nodule if present, indicated their confidence in localization, and rated the probability of a ground-glass nodule being present using a 5-point receiver operating characteristic (ROC) scale. The experimental group then received perceptual education on how to use false-color CT images to identify ground-glass nodules, while the control group received an attentional control journal article. Subsequently, both groups viewed a second set of 20 images (false-color for experimental group, grayscale for control group) and performed the same image evaluation tasks. After completing the educational session, subjects were given a questionnaire containing 8 items to determine how the false-color images and perceptual education compared with standard educational methods for radiology.

Results

Participants had a highly positive opinion about their training, p-values for all 8 questions < 0.001. There was no statistically significant improvement in performance for nodule identification, localization, or confidence. Subjects anecdotally stated that they felt like they 'saw more things' on the false-color images, but did not know how to interpret them.

Conclusions

While there was no quantitative improvement in perceptual performance, subjects had an overall very positive feedback about the educational intervention and the use of false-color images. Subjects anecdotally stated that they felt like they ‘saw more things’ on the false-color images, but did not know how to interpret them. The use of false-color images to improve medical image perception and education may benefit from further study.

FRIDAY MORNING, August 9th
Session 10 – 10:40 am to 11:40 am

Using Deep Learning to Predict Radiologists' Decisions when Reading Mammograms

Karthika Kelat[1], MS, Sarah E. Gerard[1], PhD, Bulat Ibragimov[3], PhD,
Claudia Mello-Thoms[1,2], PhD
[1] Roy J. Carver Department of Biomedical Engineering, University of Iowa, Iowa City, IA, USA
[2] Department of Radiology, University of Iowa, Iowa City, IA, USA
[3] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

Rationale

This study aims to develop a decision prediction model for radiologists and residents in reading mammograms using deep learning methods and Gabor filters. Our goal is to build individualized decision prediction models for radiologists.

Methods

In our study, we conducted two experiments to assess the effectiveness of incorporating Gabor filters into deep learning models for decision prediction in breast cancer screening. We utilized a dataset consisting of 120 mammogram cases, each containing both cranio-caudal (CC) and medio-lateral oblique (MLO) views. Among these cases, 59 mammograms exhibited malignant masses, while the remaining 61 were categorized as normal and had exhibited stability for a minimum duration of 2 years. These mammograms underwent independent interpretation by four breast radiologists, who held Mammography Quality Standards Act (MQSA) certification and possessed expertise in breast imaging, and by four radiology residents, that had undergone their breast imaging rotations. During the reading sessions, the observers were tasked with identifying and reporting malignant lesions. The observers wore a head-mounted eye tracker (ASL H6, Applied Sciences Laboratories, Bedford, MA), enabling the precise tracking of eye gaze coordinates relative to the display. Subsequently, the radiologists provided diagnostic decisions for each case, which were categorized as True Positives (TP), False Negatives (FN), and False Positives (FP), following a predefined truth table. Preprocessing steps included windowing, leveling, and normalization to standardize mammographic image intensity levels. Additionally, regions corresponding to decisions (TPs, FNs, and FPs) were cropped based on observers' field of view, approximating a circle with a radius of 2.5 degrees of visual angle (dva). These preprocessing steps were applied across both experiments.

In Experiment 1, we evaluated two base models, VGG19 and ResNet50, pretrained with ImageNet, both with and without the addition of a Gabor filter layer on top of each base model. Gabor filters were selected for their resemblance to the response of simple cells in the human visual system to visual stimuli, making them a suitable choice for feature extraction. Previous studies have also utilized Gabor filters to investigate the correlation between the performance of model observers and human observers. The Gabor filter layer, implemented as a 2D convolutional layer with fixed weights corresponding to Gabor filters, aimed to extract Gabor features from the input images. For Experiment 2, we used two inputs for the Deep Learning network. Input A represented the output of the Gabor filters after passing through the fixated areas, while Input B consisted of the fixated areas input directly to the convolutional layers. The outputs of these layers were concatenated and passed to the output layer for prediction, enabling the handling of multiple inputs. Experiment 2 is currently underway and will soon be completed. Figure 1 illustrates the setup and workflow of both experiments, providing a visual representation of the methodology employed in our study.

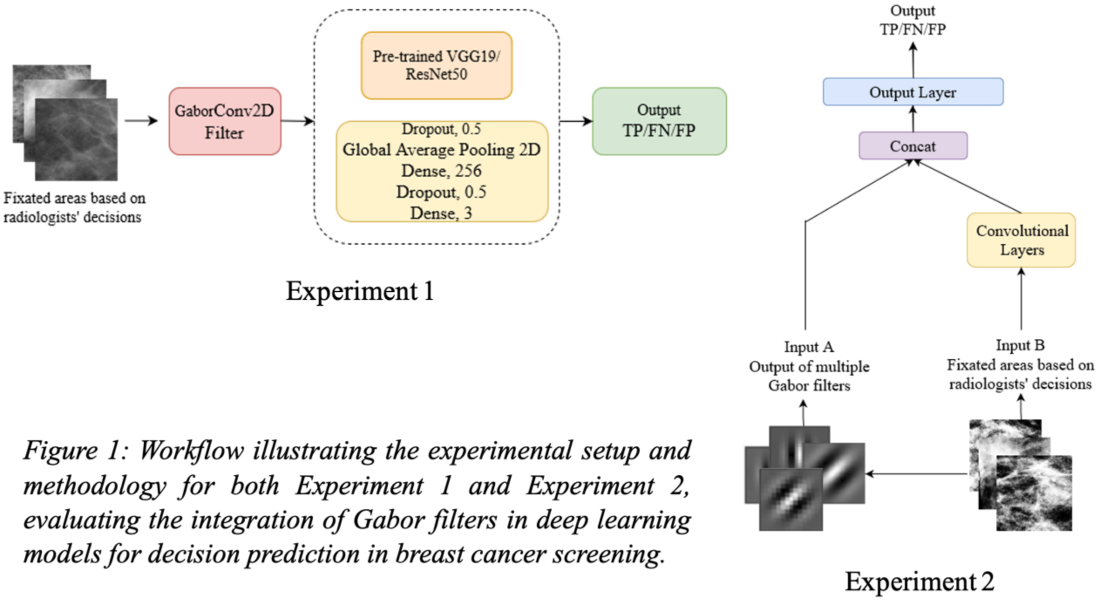

Fixated areas based on radiologists' decisions
GaborConv2D Filter
Pre-trained VGG19/ ResNet50
Dropout, 0.5
Global Average Pooling 2D
Dense, 256
Dropout, 0.5
Dense, 3
Output TP/FN/FP
Experiment 1
Output TP/FN/FP
Output Layer
Concat
Convolutional Layers
Input A Output of multiple Gabor filters
Input B Fixated areas based on radiologists' decisions
Experiment 2

Figure 1: Workflow illustrating the experimental setup and methodology for both Experiment 1 and Experiment 2, evaluating the integration of Gabor filters in deep learning models for decision prediction in breast cancer screening.

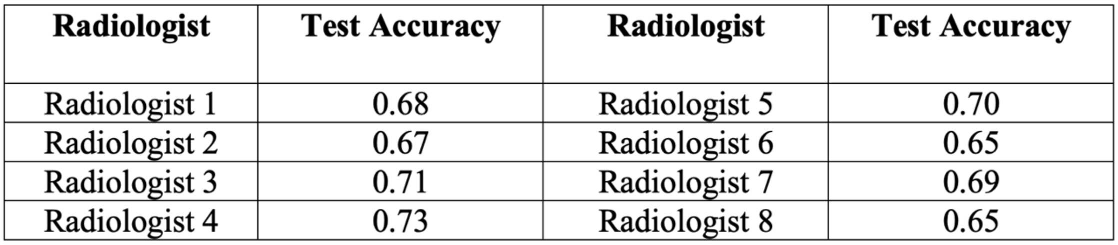
Results

In Experiment 1, we conducted 10-fold cross-validation to evaluate the performance of our decision prediction models. The mean test accuracy across all 10 folds is summarized in Table 1. Notably, each of the four decision prediction models yielded varying results across radiologists, with Radiologist 4 consistently achieving higher accuracy compared to others. However, we observed fluctuations in test accuracy among the models, indicating potential influences of individual radiologist data on model performance. Further analysis of the confusion matrix (not shown) revealed misclassifications primarily between True Positive (TP) and False Negative (FN) classes. Interestingly, models without Gabor filter layers exhibited better decision prediction capabilities. In Experiment 2, preliminary results did not surpass the performance of Experiment 1. However, it's important to note that Experiment 2 is still in its preliminary stage, and optimization efforts are underway. Future enhancements will involve integrating additional visual search features such as dwell time and mean pupil diameter into the network architecture. Table 2 summarizes the test accuracy of three radiologists' performance in Experiment 2.

Table 1: Four models used for decision prediction and corresponding mean test accuracy across all 10 folds of eight different radiologists.

Radiologist	**Mean Test Accuracy across all 10 Folds**			
	VGG19	**VGG19 with GaborConv2D**	**ResNet50**	**ResNet50 with GaborConv2D**
Radiologist 1	0.75	0.73	0.76	0.68
Radiologist 2	0.69	0.65	0.66	0.63
Radiologist 3	0.70	0.68	0.72	0.66
Radiologist 4	**0.84**	**0.78**	**0.83**	**0.81**
Radiologist 5	0.72	0.66	0.69	0.67
Radiologist 6	0.66	0.65	0.68	0.70
Radiologist 7	0.73	0.67	0.74	0.67
Radiologist 8	0.70	0.69	0.67	0.63

Table 2: Test accuracy of Experiment 2.



Radiologist	**Test Accuracy**	**Radiologist**	**Test Accuracy**
Radiologist 1	0.68	Radiologist 5	0.70
Radiologist 2	0.67	Radiologist 6	0.65
Radiologist 3	0.71	Radiologist 7	0.69
Radiologist 4	0.73	Radiologist 8	0.65

Conclusion

In conclusion, our study investigated the integration of Gabor filters into deep learning models for decision prediction in reading mammograms. Experiment 1 demonstrated consistent performance across different folds of 10-fold cross-validation, with slight variations in accuracy observed. While individual radiologists yielded varying results, models without Gabor filter layers showed superior prediction capabilities. However, preliminary results from Experiment 2 did not surpass Experiment 1's performance. Despite this, ongoing efforts aim to optimize Experiment 2 by integrating additional visual search features into the network architecture. Overall, our findings emphasize the complexity of decision prediction in reading mammograms and underscore the need for further research to enhance the accuracy and reliability of decision support systems in clinical practice.

A phantom image quality study comparing ultra-low and standard dose computed tomography protocols for the investigation of non-accidental injury

Lisa Kingston, MSc; Niamh Moore, MSc; Rena Young, MSc, Conor Lee, MSc; Andrew England, PhD; Mark F. McEntee, PhD.
Discipline of Medical Imaging & Radiation Therapy, University College Cork, Cork, IE

Rationale

Early identified of paediatric non-accidental injury (NAI) is of paramount importance. Whilst successful diagnosis is key, steps are needed to maintain radiation dose levels that are low as reasonably practicable (ALARP). The aim of this study is to determine whether there is potential for the development and utilisation of low dose computed tomography (CT) protocols for the diagnosis of suspected non-accidental injury (NAI) and to compare whether am ultra-low dose (ULD) CT protocol could deliver image quality (IQ) comparable to that of a standard dose (STD) NAI CT protocol.

Methods

CT images were acquired on a Revolution Apex™ scanner, using a new-born whole body phantom, and reconstructed using deep learning software. Two protocols were compared in terms of IQ, an ULD protocol, with a dose-length product (DLP) of 1.49mGy and a STD protocol, with a DLP of 22.92mGy. Participants graded IQ using a modified 4-point Likert scale (Perfect, Good, Moderate, Inadequate). Mann-Whitney and Chi-Squared tests were carried out to determine if there was a significant difference in IQ between the protocols. A t-test was performed to determine if there was a significant difference in participants level confidence in the IQ between the protocols for the investigation of suspected NAI.

Results

27 of the 31 responses (87%) were included in the study due to the application of the inclusion and exclusion criteria. Both the Mann Whitney and Chi-squared test demonstrated significant difference ($P \leq 0.05$) in the IQ between the protocols (**Figure 1**), for both the evaluation of cortical and trabecular bone. T-test results demonstrated a significant difference ($P \leq 0.05$) in the level of confidence amongst participants between the protocols.

Conclusions

This phantom based study suggests that the IQ of the STD protocol was significantly better when compared to the ULD protocol, for the investigation of suspected NAI. Further work is required to fully optimise an ULD-CT protocol for the investigation of NAI and to confirm its clinical potential.

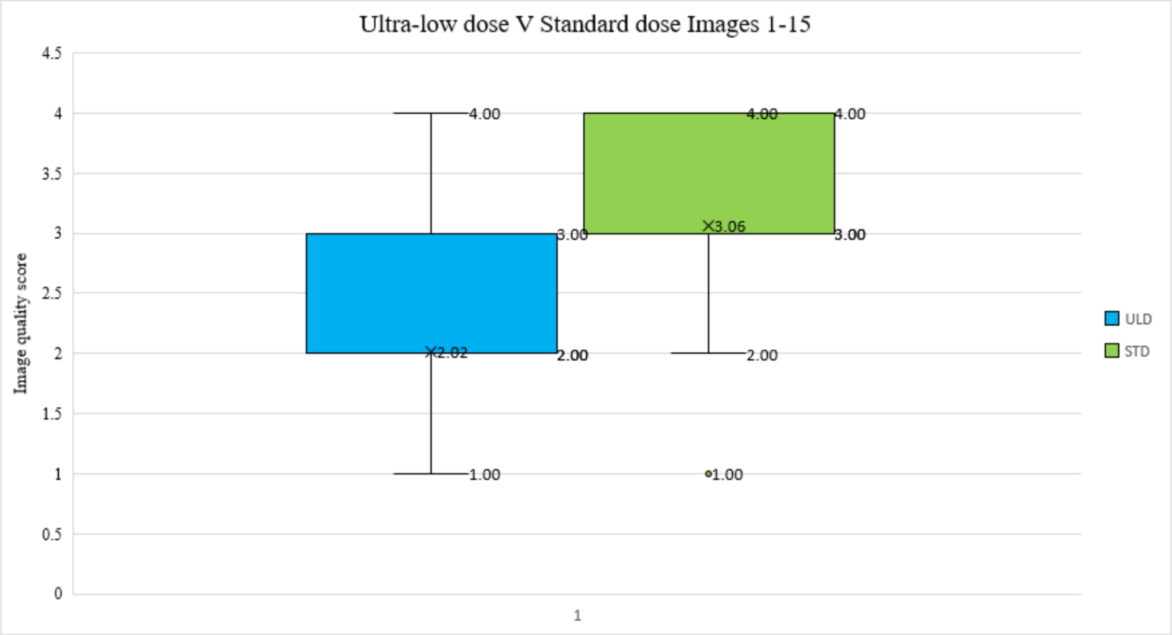


Figure 1. Box plot summarising the characterisation of IQ across the two CT datasets. STD – standard CT protocol (22.92mGy); ULD – ultra-low dose CT protocol (1.49mGy).

Convolutional Neural Network Model Observer During Search in Virtual Digital Breast Tomosynthesis

Miguel P. Eckstein[1,2], Aditya Jonnalagadda[2], and Craig K. Abbey[1]
[1]*Department of Psychological and Brain Sciences*
[2]*Department of Electrical and Computer Engineering*
Santa Barbara, USA

Rationale

Model observers are computational tools to evaluate and optimize task-based medical image quality. Linear model observers, such as Channelized Hotelling Observer, predict human accuracy in detection tasks with a few possible signal locations in clinical phantoms or real anatomic backgrounds. In recent years, Convolutional Neural Networks (CNNs) have been proposed as a new type of model observer for task-based assessment of medical image quality. What is not well understood is what CNNs add over the more common linear model observer approaches. Here, we compare the detection accuracy of the Channelized Hotelling Observer (CHO) and a Convolution Neural Network (CNN) to the radiologists' accuracy in searching for two types of signals embedded in 2D/3D breast tomosynthesis phantoms (DBT).

Methods

First, we explored the relationship among the accuracies of two classic linear model observers (Channelized Hotelling Observer models, CHO, and Filtered Channel Observer, FCO), and a Convolutional Neural Network (CNN) for the commonly used Location Known Exactly task (LKE) in which the target can appear in M specified locations. The task was yes/no with the target being present in 50 % of the trials. Second, we compared the accuracies of CHO, FCO, CNN, and radiologists for a yes/no search task in which the target could appear at any location. The comparisons were conducted for two 2D images and 3D image stacks using digital breast tomosynthesis phantoms and for two target types (a mass-like target and a micro-calcification type target).

Results

We show that for the LKE detection task, the CHO model's accuracy is comparable to the CNN's performance. However, for the search task with 2D/3D DBT phantoms, the Channelized Hotelling models' detection accuracy was significantly lower than the CNN accuracy. A comparison to radiologists' accuracy showed that the CNN but not the Channelized Hotelling could match or exceed radiologist accuracy for the 2D microcalcification and 3D mass search task. An analysis of the eye position of radiologists showed that they fixated more often and longer at the locations corresponding to CNN false positives. Most of the CHO false positives related to the phantoms' normal anatomy were not fixated by radiologists.

Conclusion

We show that CNNs can be used as an anthropomorphic model observer for the search task for which traditional linear model observers fail. The linear model observers show inability to discount false positives arising from the anatomical backgrounds and result lower search accuracy than radiologists.