Decision Quality Support in Diagnostic Breast Ultrasound through Artificial Intelligence

Lev Barinov [1,2,3], Ajit Jairaj [1], Lina Paster [4], William Hulbert[1], Richard Mammone[5], Christine Podilchuk[1]

[1] ClearView Diagnostics Inc., Piscataway, NJ 08854
[2] Princeton University, Department of Molecular Biology, Princeton NJ 08540
[3] Rutgers University, Robert Wood Johnson Medical School, Piscataway, NJ 08854
[4] Robert Wood Johnson University Hospital, Department of Radiology, New Brunswick, NJ 08901
[5] Rutgers University, Department of Electrical and Computer Engineering, Piscataway, NJ 08854

Abstract— **Medical Ultrasonography is a valuable imaging technology for medical diagnostics and, more recently, as a screening alternative to mammography for women with dense breasts. However, ultrasound imaging within the contexts of both diagnostic and screening mammography suffers from inter-operator and intra-operator variability. Consequently, there is a broad distribution of performance profiles, even for radiologists of similar training. Typically, these profiles tend to err on the side of caution, preferring false positive errors to false negative errors. While this approach may lead to a higher Cancer Detection Rate (CDR), it also lowers the Positive Predictive Value (PPV_3) of performed biopsies. A lower PPV3 translates to an increase in benign biopsies, the annual cost of which are estimated to be on the order of $1 - $3 billion USD (not including pathological workups). And, of course, there is the immeasurable cost of pain, worry, and suffering borne by women undergoing these potentially unnecessary procedures. In this paper, we evaluate the ability of the ClearView cCAD algorithms to increase overall performance and reduce the inter-operator variance on a set of imaged lesions. The cCAD system provides an automated assessment of some ACR BI-RADs criteria and calculates a preliminary BI-RADs assessment, given as BI-RADS categorical bucket (1-3) or (4-5). Through the evaluation of 1300 breast lesion images, 3 MQSA certified radiologists were asked to determine both a Likelihood of Malignancy (LoM) and a BI-RADs assessment, from which their ROC curve AUC as well as PPV_3 could be calculated. The cCAD system was also evaluated, on the same set of lesions, by a similar set of metrics. From this analysis we have been able to show that the cCAD system outperforms radiologists at all operating points within the scope of this study design. Furthermore, we've shown that through simple fusion schemes we are able to increase performance beyond that of either the cCAD system or the radiologist alone by all typically tracked quality metrics, and significantly reduce inter-operator variance.**

Keywords—ultrasound imaging; computer aided diagnosis; inter-operator variability;

I. Introduction

Breast cancer screening and, subsequently, diagnostic workflows generally exhaust the initial modality and eventually move onto to an ultrasound evaluation of the area in question [1]. This workflow is followed when there are suspicious non-imaging findings or the initial screening mammogram shows that the breast tissue is dense. It has been shown that typical screening mammography can fail to detect 20-30% of breast lesions, 60-70% of which are visible on retrospective evaluation. These results are particularly striking as most clinicians err on the side of caution, leading to a much higher benign biopsy rate and consequently lower Positive Predictive Value (PPV_3) for biopsies. According to the National Mammography Database, over 70% of all biopsies come back benign [2] with an estimated annual utilization of 984,000 biopsies per year. [3] According to Burkhardt and Sunshine [4], the average billed costs of a surgical and core biopsy are $3764 and $1496, respectively. The resulting annual cost ranges from $1.47 to $3.7 billion, if you don't include the necessary pathological workup. In addition, a recent study [5] has found that on average breast biopsies burden patients with an additional $310 in out-of-pocket costs. This costly combination of false positive and false negative error rates has motivated researchers to develop computer aided detection (CAD_e) and computer aided diagnosis (CAD_x) algorithms [6-11]. Although several CAD tools are available for clinical use, their benefits have been called into question, with some studies suggesting they offer no benefit whatsoever [12].

The current state of CAD technology, offering dubious benefit to mammography, and offering no benefit at all to ultrasound, seems to indicate that a paradigm shift in the utilization of intelligent medical algorithms is necessary. Recently, ClearView Diagnostics Inc. (CDI) released a new product, cCAD (awaiting FDA clearance: K161959), aimed at improving the overall diagnostic quality of breast lesion analysis through automated reporting of lesion parameters. In this paper, we explore a novel application of their underlying algorithms as a decision support system aimed at providing support feedback to the clinician, instead of an independent CAD assessment. We will explore raw performance metrics of the cCAD system from a traditional CAD standpoint, as well as various fusion schemes with clinician grading for utilization in a decision support system. This paper will examine if any improvement in performance or reduction in inter-operator variability is possible through this new application paradigm.

II. CAD Versus Decision Suppport Systems

It is important to illuminate the distinction between Computer-Aided Detection/Computer-Aided Diagnosis and Decision Support Systems. Although the underlying image analysis approaches share some methods and similarities, the intent and utilization of these two systems draws a clear distinction between them. CAD systems are traditionally independent image analysis pipelines aimed at providing an independent assessment of an image, lesion, or region of

This study was sponsored by ClearView Diagnostics Inc.

interest. In effect, they are used as a second opinion in place of, or as an adjunct to, a second reader. This type of utilization is useful when there is significant inter-operator variability or a relative dearth of trained personnel that can perform and interpret examinations.

In contrast, decision support systems aim to establish a human-machine interface that benefits both the Artificial Intelligence (AI) platform and the trained reader through a process termed symbiotic learning. In this framework, recommendations are passed between the two systems in a bidirectional fashion, in order to facilitate optimal performance by the joint human-AI system. The system can be viewed as coupling the human's decision making pipeline to the AI's, so as to make a single accurate and consistent decision.

III. Study Design

To evaluate the performance of the cCAD system as a traditional CAD system, as well as its utility in a Decision Support System, a large database of lesions was aggregated. These lesions were, in part, made of ACRIN 6666 study lesions [13] in addition to internal studies conducted by CDI. This database consisted of over 1300 individual images and included 680 individual lesions. This dataset contained 600 cancers and 700 benign examples when counted by image, or 298 cancers and 382 benign examples when counted by lesion. The ground truth for all these lesions was established by biopsy or 1-year follow-up with initial BI-RADs (Breast Imaging and Reporting Data System) assessments distributed as seen in Figure 1.

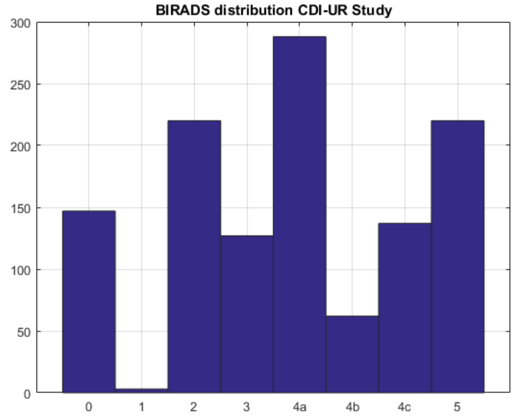


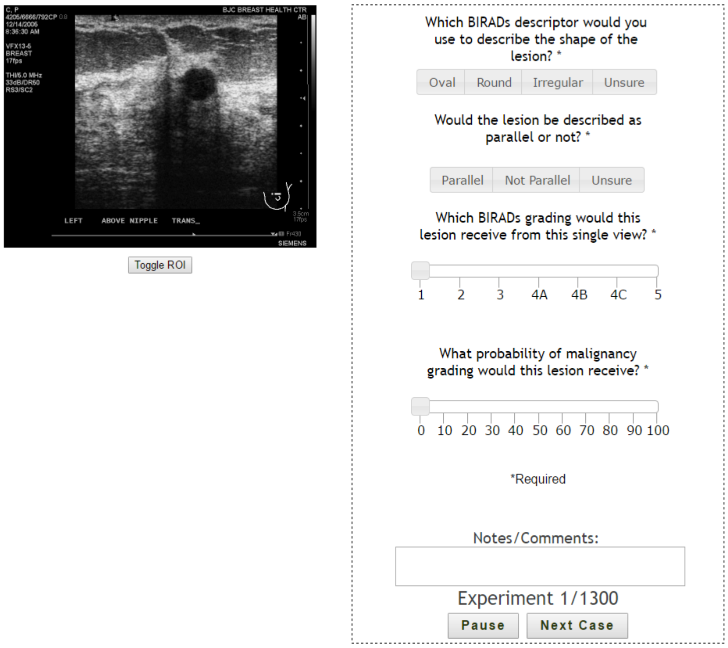
Figure 1 - BI-RADS Distribution by image for data included in study. BI-RADS grading of 0 denotes that explicit BI-RADS score was not known.

Three MQSA certified radiologists, as described in Table 1, were asked to evaluate all 1300 images on an image by image basis and assess both the Likelihood of Malignancy (LoM) and the preliminary BI-RADs assessment. The reading radiologists would typically be privy to a host of other information, including patient history and previous imaging studies. For the purposes of this study, that information was not presented. This decision was made in order to isolate and compare the image reading capacity of the cCAD system and the radiologist while avoiding potential confounding from external information.

Table 1 Radiologists included in study.

Radiologist	MQSA Certified	Years of Experience	Annual Cases Read
Rad 1	X	20+	18605
Rad 2	X	20+	8060
Rad 3	X	10	7201

Each radiologist was given a unique login and asked to assess each of the 1300 images per the interface seen in Figure 2.



Which BIRADs descriptor would you use to describe the shape of the lesion? *

Would the lesion be described as parallel or not? *

Which BIRADs grading would this lesion receive from this single view? *

What probability of malignancy grading would this lesion receive? *

*Required

Notes/Comments:

Experiment 1/1300

Figure 2 Study assessment tool

The radiologist was presented with the full ultrasound image. The viewer included controls to adjust zoom, gain, and contrast. Additionally, they were given the option to toggle an ROI overlay on the image. This was done so that the assessment was purely of their diagnostic ability, and unaffected by potential difficulties in locating the lesion within an image. When presented with a lesion, the radiologist filled out and submitted the accompanying form describing their assessment of its shape, orientation, likelihood of malignancy, and BIRADS category. Upon submitting a lesion, the radiologist was no longer allowed to return to or edit their interpretation of it. This process was continued until each radiologist finished evaluating each of the 1300 lesions. So as to ensure that neither fatigue nor urgency played a role in their decision making process, the radiologists were given a period of weeks over which to complete these assessments.

Upon the completion of the radiologists' assessments, the same protocol is followed by the cCAD system; each ROI in each image is fed through the system and a BI-RADS bucket assessment as well as its confidence in that assessment is recorded. To facilitate the comparison, in this paper the confidence for a BI-RADs bucket is interpreted as a proxy measure of Likelihood of Malignancy (LoM) for the cCAD system. It is important to note, however, that the cCAD system has made no claims as to the validity of this interpretation, and its indications for use are explicitly reserved for aiding in the compliance of the BI-RADS ultrasound lexicon form.

IV. Results

For the initial analysis, we simply compare the results of the three radiologists' LoM, evaluated using Receiver Operating Characteristic (ROC) curves. These curves compared to the output of the LoM equivalent provided by the cCAD system. Area Under the Curve (AUC) is calculated for each radiologist and the cCAD system using a simple trapezoidal rule approximation. It is interesting to note that, although the radiologists were not given specific criteria for how to grade LoM, the likelihood estimates they gave were still coarsely quantized. This is evident from the linear regions of the ROC curve, which are separated by coarse jumps between likelihood thresholds. The curves and their respective AUCs are shown in Figure 3.

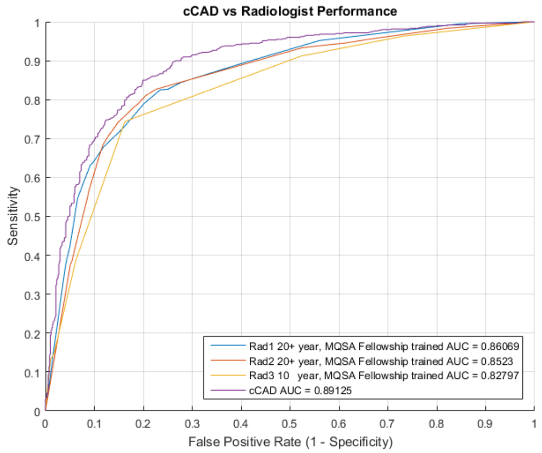


Figure 3 ROC Analysis comparing radiologists to raw cCAD

From the results in Figure 3, it is clear that cCAD independently outperforms the radiologists in the assessment of LoM within the study protocol and its parameters.

Since the goal of this paper was to establish whether or not AI, statistical methodologies, and machine learning can be used to augment the decision making pipeline of the physician we have analyzed the results of the cCAD and radiologist system as if they were a single pipeline. In its most simple rendition, we evaluate a derivative score as one where both the cCAD result and the clinician's evaluation are treated as equal weighted estimates of the LoM, offering support to one another. To generate the final output one would simply define the score as the equal weighted mean of the *N* outputs where these outputs can be defined as independent clinical evaluations, the cCAD evaluation, or any other method of evaluating the LoM for a particular lesion

$$\boldsymbol{S} = \omega_1 s_1 + \omega_2 s_2 + \omega_3 s_3 \ldots + \omega_n s_n \qquad (1)$$
$$\omega_n = 1/N$$

Taking this approach and analyzing the statistics of aggregate performance parameters for each radiologists, we can assess the benefit of the fusion scheme on clinical parameters. Since the operating point of the cCAD system, the radiologists, and their fused scores lie on a continuum, there are many possible operating points to analyze. Here, we offer MQSA tracked statistics for two such points:

1. Fixed Sensitivity to the Radiologists performance prior to fusion (Table 2)
2. Fixed Specificity to the Radiologists performance prior to fusion (Table 3)

In both Table 2 and Table 3 we reference a new statistic, called the Benign Biopsy Reduction (BBR), which is defined as the percent reduction in number of benign lesions that were originally sent to biopsy. For example, if four lesions were originally sent to biopsy and one of these lesions is eliminated by the new fusion technique, we would report a BBR of 25%. These values are obviously sensitivity dependent and, as such, should always be analyzed with an equivalent sensitivity and specificity (or derivative) metric.

Table 2 Performance shifts when sensitivity is fixed to original radiologists' sensitivity

Radiologist	Benign Biopsy Reduction	ΔPPV$_3$	Δsensitivity
Rad 1	34.09%	7.31%	0%
Rad 2	47.60%	19.54%	0%
Rad 3	55.40%	20.26%	0%

Table 3 Performance shifts when sensitivity is fixed to original radiologists' specificity

Radiologist	Benign Biopsy Reduction	ΔPPV$_3$	Δsensitivity
Rad 1	25.00%	0.57%	1.33%
Rad 2	25.48%	1.41%	3.17%
Rad 3	31.02%	2.04 %	4.83%

Furthermore, in Figure 4, a clear and emergent benefit to both the sensitivity and specificity metrics is visible. This figure, which compares the sensitivity and specificity of each radiologist before and after fusion, brings to light another benefit of the decision support system. In addition to the consistent improvements in sensitivity, specificity, and MQSA tracked metrics, the fused results demonstrate an overall reduction in inter-operator variability.

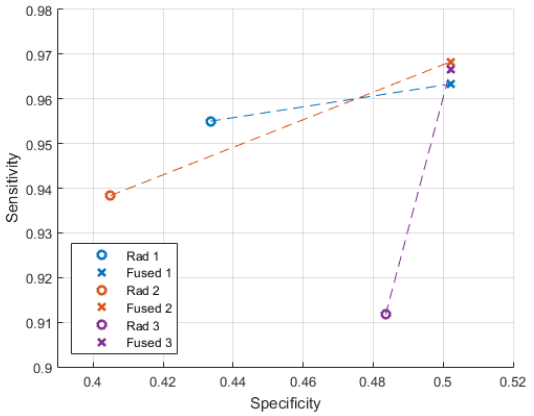


Figure 4 Operating point shifts for a fixed specificity system.

In order to verify this observation, we perform an unweighted Cohen Kappa [14-16] analysis to each pair of radiologists with and without decision support from the cCAD software. The results are summarized in Tables 4 and 5 below:

Table 4 Unweighted Cohen Kappa analysis for original radiologist LoM (p = .05)

Unweighted Cohen Kappa	Without CDI Decision Support		
	Rad 1	Rad 2	Rad 3
Rad 1	1.00	0.67 [.62 - .72]	0.61 [.55 - .66]
Rad 2	0.67 [.62 - .72]	1.00	0.61 [.56 - .66]
Rad 3	0.61 [.55 - .66]	0.61 [.56 - .66]	1.00

Table 5 Unweighted Cohen Kappa analysis for LoM with cCAD Decision Support (p = .05)

Unweighted Cohen Kappa	With CDI Decision Support		
	Rad 1	Rad 2	Rad 3
Rad 1	1.00	0.89 [.86 - .92]	0.90 [.88 - .93]
Rad 2	0.89 [.86 - .92]	1.00	0.91 [.87 -. 94]
Rad 3	0.90 [.88 - .93]	0.91 [.87 -. 94]	1.00

There is a clear and statistically significant improvement in Kappa that occurs alongside the improvements in quality metrics. This suggests that the fusion system is, in fact, decreasing inter-operator variability while simultaneously improving diagnostic quality.

CONCLUSION

We have investigated the efficacy of adjoining clinical workflows and CDI's cCAD platform into a decision support system. This framework has measurably increased MQSA tracked quality metrics such as sensitivity and PPV$_3$. Concurrent with the performance improvement, the cCAD's influence on the combined decision has the added effect of reducing inter-operator variability, which suggests the potential for boosting both reliability and consistency in those decisions.

REFERENCES

[1] Madjar, H. (2010). Role of breast ultrasound for the detection and differentiation of breast lesions. Breast Care, 5(2), 109-114.

[2] American College of Radiology. National Mammography Database (NMD) [September 2016]. Available at: http://www.acr.org/Quality-Safety/National-Radiology-Data-Registry/National-Mammography-DB

[3] Ghosh K, Melton L, III, Suman VJ, et al. Breast Biopsy Utilization: A Population-Based Study. Arch Intern Med. 2005;165(14):1593-1598. doi:10.1001/archinte.165.14.1593.

[4] Burkhardt, Jeffrey H., and Jonathan H. Sunshine. "Core-Needle and Surgical Breast Biopsy: Comparison of Three Methods of Assessing Cost 1." Radiology 212.1 (1999): 181-188.

[5] Alcusky M, Philpotts L, Bonafede M, Clarke J, and Skoufalos A. Journal of Women's Health. September 2014, 23(S1): S-11-S-19. doi:10.1089/jwh.2014.1511.

[6] Chou, Y. H., Tiu, C. M., Hung, G. S., Wu, S. C., Chang, T. Y., & Chiang, H. K. (2001). Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis. Ultrasound in medicine & biology, 27(11), 1493-1498.

[7] Yap, M. H. (2008). A novel algorithm for initial lesion detection in ultrasound breast images. Journal of Applied Clinical Medical Physics, 9(4).

[8] Fujita, H., Uchiyama, Y., Nakagawa, T., Fukuoka, D., Hatanaka, Y., Hara, T., ... & Zhou, X. (2008). Computer-aided diagnosis: The emerging of three CAD systems induced by Japanese health care needs. Computer methods and programs in biomedicine, 92(3), 238-248.

[9] Chang, R. F., Wu, W. J., Moon, W. K., & Chen, D. R. (2005). Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors. Breast Cancer Research and Treatment, 89(2), 179-185.

[10] Shen, W. C., Chang, R. F., Moon, W. K., Chou, Y. H., & Huang, C. S. (2007). Breast ultrasound computer-aided diagnosis using BI-RADS features. Academic radiology, 14(8), 928-939.

[11] Drukker, K., Giger, M. L., Horsch, K., Kupinski, M. A., Vyborny, C. J., & Mendelson, E. B. (2002). Computerized lesion detection on breast ultrasound. Medical physics, 29(7), 1438-1446.

[12] Lehman, Constance D., et al. "Diagnostic accuracy of digital screening mammography with and without computer-aided detection." JAMA internal medicine 175.11 (2015): 1828-1837.

[13] Berg, W. A., Blume, J. D., Cormack, J. B., Mendelson, E. B., Lehrer, D., Böhm-Vélez, M., ... & Mahoney, M. C. (2008). Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. Jama, 299(18), 2151-2163.

[14] Cohen, Jacob. "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." Psychological bulletin 70.4 (1968): 213.

[15] Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72(5), 323.

[16] Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy, 85(3), 257-268.