High Performance of FDA-Cleared Platform for Mammography Triage

Tara A. Retson, MD/PhD[1], Vivian Lim, MD[1], Alyssa T. Watanabe, MD[2,3]

1. UCSD School of Medicine, La Jolla, CA, 2. USC Keck School of Medicine, Los Angeles, CA, 3. CureMetrix, Inc., La Jolla, CA

BACKGROUND

- Screening mammography saves lives with early detection of breast cancer.
- Double reading increases cancer detection and decreases recall, but is often impractical.
- Compared to computer aided detection (CAD) programs which highlight individual imaging features, <u>triage programs prioritize or flag exams within a radiology worklist.</u>
- Recent studies suggest that artificial intelligence (AI) based triage programs could improve cancer detection and expedite radiologist workflow.
- We sought to evaluate the performance of a commercial AI-based triage algorithm on exams with varying breast densities and lesion types.

METHODS – Test Set

Retrospective study of patient exams with IRB approval and wavier of formal patient consent. All mammograms were anonymized using a HIPAA-compliant protocol. Images were analyzed by a commercially available AI algorithm, cmTriage, CureMetrix.

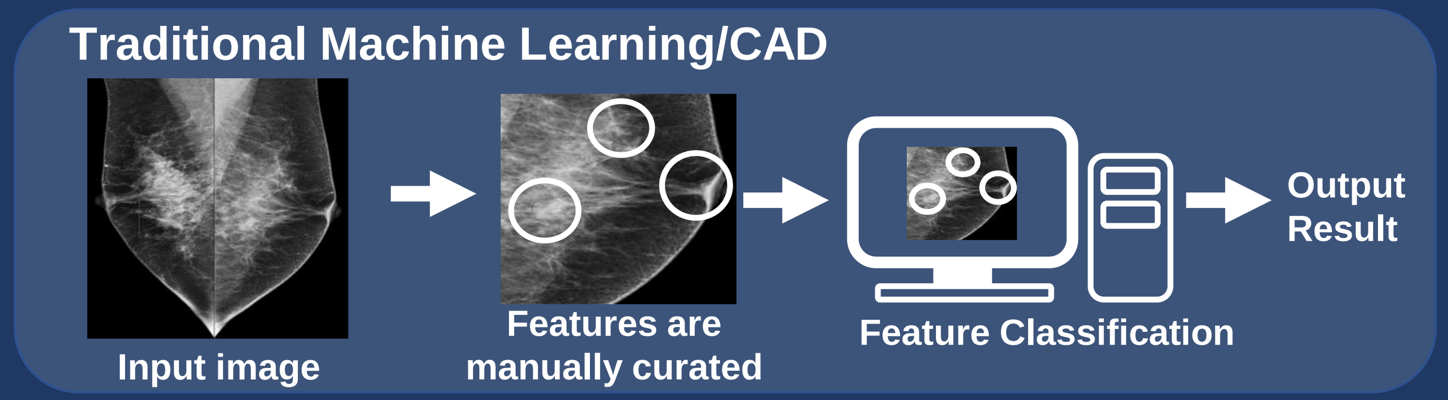
<u>Enriched, multi-institutional</u>

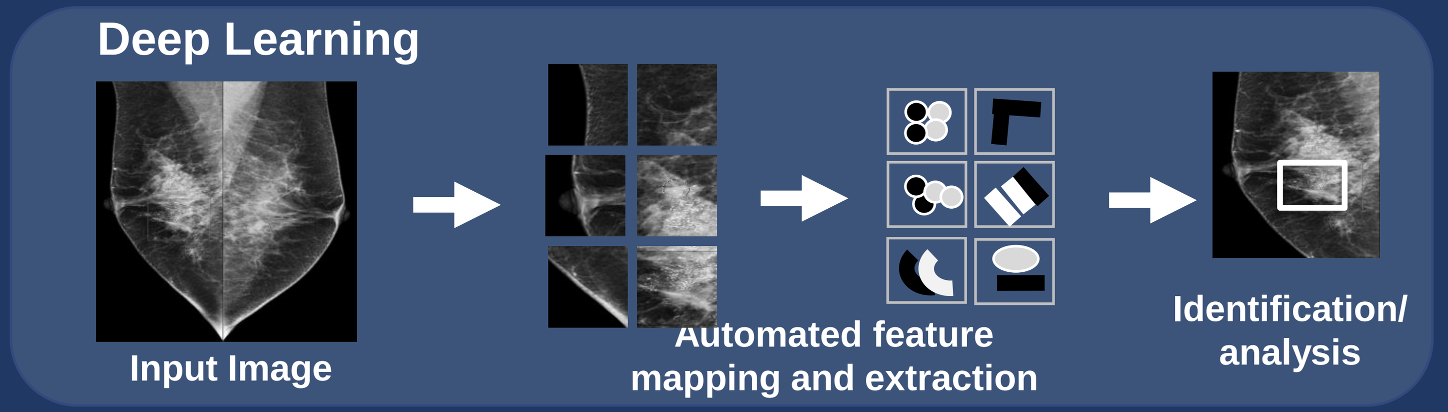
- 1255 screening 2D digital mammograms
- 4 view screening mammograms (LCC, LMLO, RCC, RMLO)
- 400 biopsy proven cancers
- 855 negative
- 31.9% cancer prevalence
- 3 different imaging facilities, multiple equipment vendors

Patient demographics:

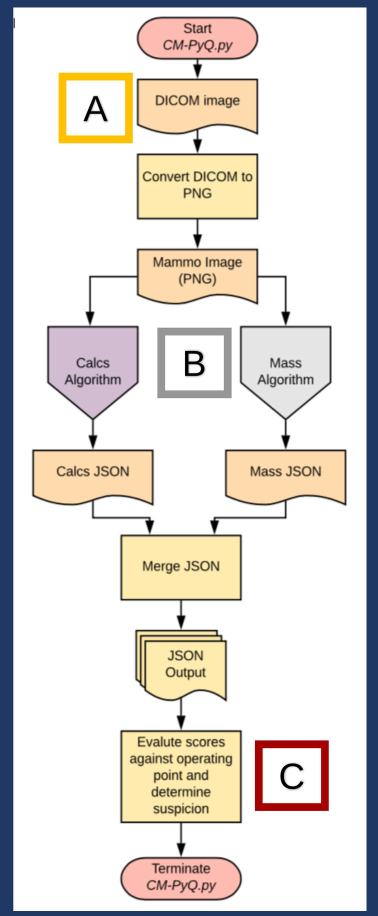
		Cancer	Normal	Total	% Test Set	% Population
	Patients	400	855	1255		
Density	Fatty	32	107	139	11%	14%
	Scattered fibroglandular	124	297	421	34%	45%
	Heterogeneously dense	177	366	543	43%	34%
	Extremely dense	67	85	152	12%	7%
Lesion Type	Mass	278			69.5%	69%
	Calcification	122			30.5%	31%
Lesion Size (Masses)	1-5mm	9			3.2%	12.7%
	5-10mm	68			24.5%	25.6%
	10-15mm	76			27.3%	25.5%
	15-20mm	60			21.6%	14.7%
	> 20mm	65			23.4%	21.5%
Age	18-39			11	1%	3%
	40-44			92	7%	12%
	45-49			135	11%	14%
	50-54			160	13%	15%
	55-59			164	13%	15%
	60-64			176	14%	13%
	65-69			182	14%	10%
	70-74			124	10%	7%
	75-79			101	8%	5%
	80+			109	9%	5%
	Mean				61.7	57.5
	Median				61	56

METHODS – AI analysis

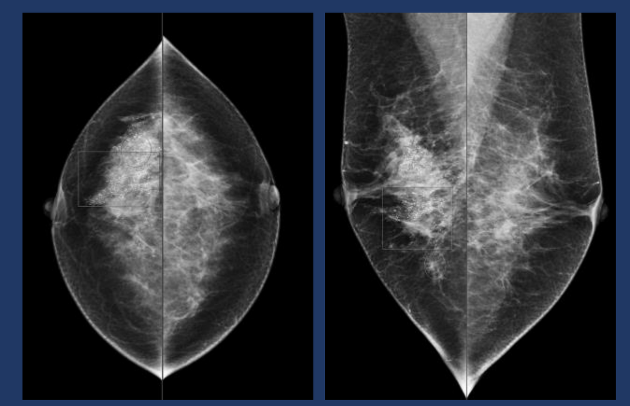




Algorithm Overview

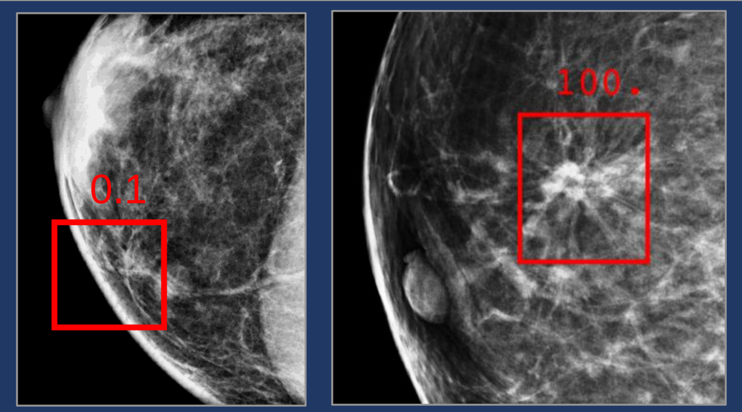


A



4 view screening mammograms (in DICOM format) are loaded into the algorithm for analysis

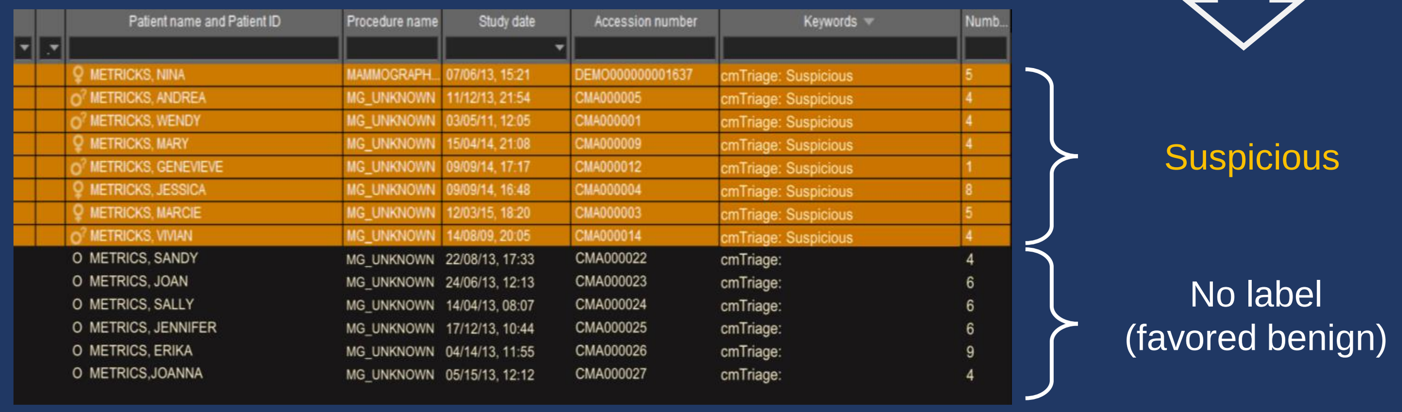
B



Low score: **favored benign**

High score: **suspicious**

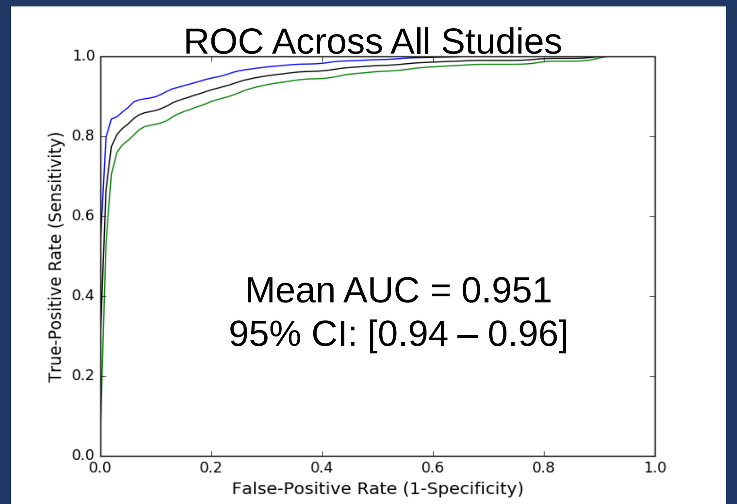
C AI analysis of mammograms generates case-based, quantitative scores (compiled from AI-based, pixel-wise lesion scoring). If the overall exam score meets a software-defined threshold, it is labeled as "Suspicious," and placed at the top of a worklist.



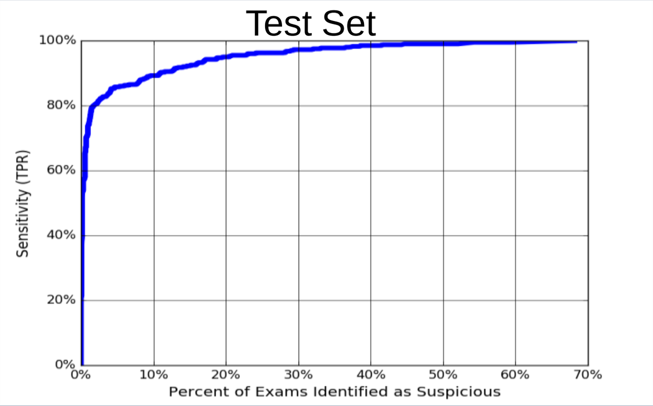
Example worklist shown above. In the practical implementation of this software, no diagnostic information is given beyond "Suspicious" or unlabeled.

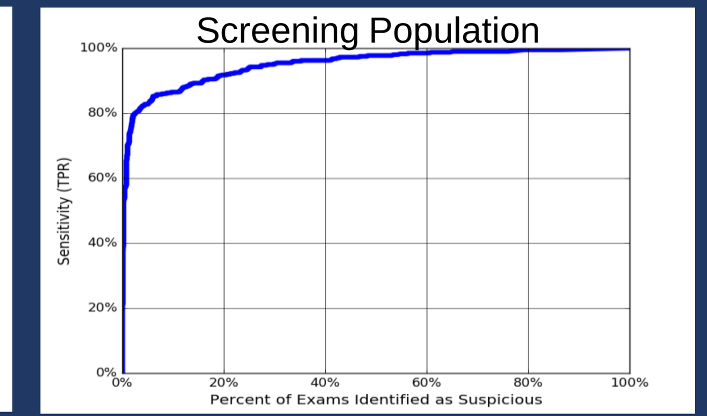
RESULTS

Algorithm Performance - Receiver Operating Characteristic Curves / Area Under the Curve (AUC)



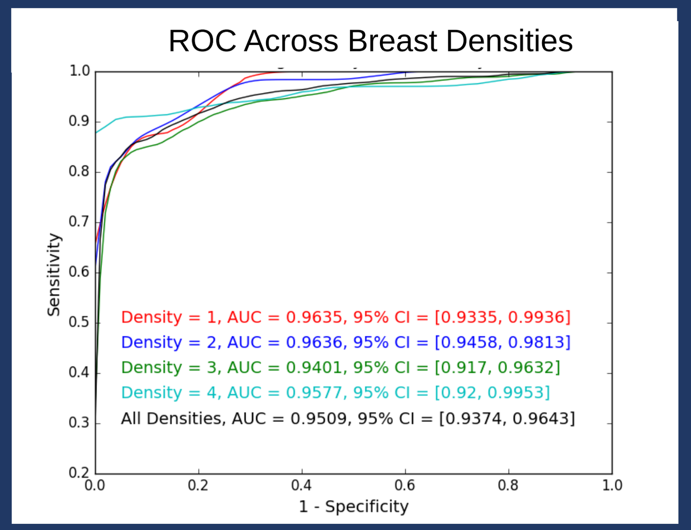
Across all lesion types, sizes and breast densities, the AUC was 0.951 (black line) with 95% confidence interval (blue and green lines)





Above, curves indicate algorithm performance across a range of sensitivities in relation to the percentage of exams that would be identified as suspicious. Left indicates performance on this cancer enriched test set, while the graph at right is extrapolated to performance on a screening population with a 0.5% cancer rate.

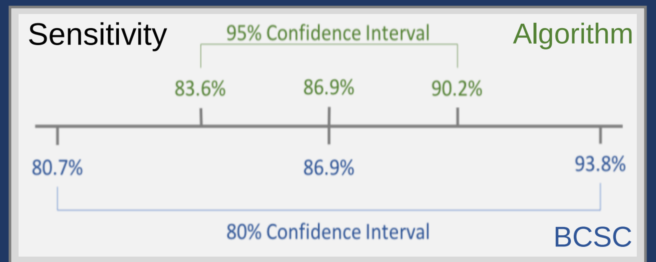
At the default sensitivity of 93% (specificity = 76.3%), the algorithm will label 41.4% of exams as suspicious (compared to 32% true positives) in this enriched test set. Adjusting for a 0.5% cancer rate, at 93% sensitivity, it would indicate 24% of exams as suspicious (compared to real-world callback rate of 11.6%).

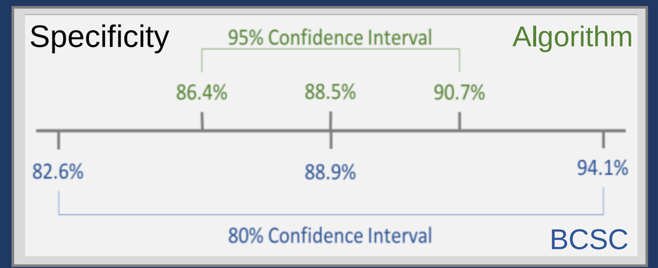


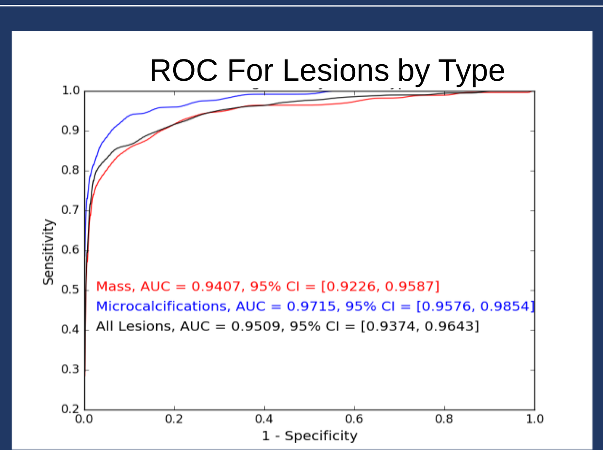
Almost half of patients have extremely dense or heterogeneously dense breasts, carrying an increase in cancer risk both from primary causes and a masking effect. In contrast to several previous works, the algorithm tested here shows similar performance across densities.

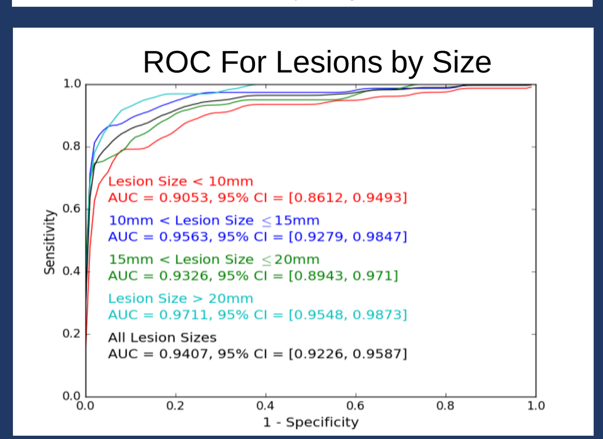
Density 1/A = Fatty
Density 2/B = Scattered Fibroglandular
Density 3/C = Heterogeneously Dense
Density 4/D = Extremely Dense

The Breast Cancer Surveillance Consortium (BCSC) study is a US-based multicenter study with data from over 1.6 million screening mammograms. It describes a real-world imager sensitivity of 86.9% and specificity of 88.9%. Testing this algorithm at the BCSC clinical sensitivity of 86.9% yields a similar specificity of 88.5%. The low end of the algorithm 95% CI for sensitivity and specificity (83.5% and 86.3%, respectively), exceeded BCSC's low end of their 80% CIs (80.7% and 82.6%). This may suggest algorithm performance in line with imagers in a clinical setting.









Algorithm performance was evaluated on two lesion types (top chart) and performed slightly better at detection of microcalcifications (AUC 0.97), compared to masses (AUC 0.94).

When subdividing masses by size (bottom chart), performance on detection of masses was similar between, 10mm to >20mm (lesions measuring 10-15mm had an AUC of 0.95, 15-20mm an AUC of 0.93, and >20mm an AUC of 0.94). Performance was comparatively decreased on small lesions measuring <10mm (AUC of 0.90).

Recent Studies	AUC	AUC For Dense Breasts (densities 3 and 4)
Tested algorithm	0.95	0.94, 0.96
Salim et al. (3 commercial algorithms evaluated)	0.96 0.92 0.92	0.94 0.90 0.90 (density was divided by low vs high)
Yala et al.	0.82	0.85, 0.71
McKinney et al. (2 populations tested)	0.89 0.81	
Schaffter et al. (top 2 performers in a challenge)	0.90 0.86	

Above, comparisons between the performance of the algorithm tested here and recently published works describing other mammography algorithms. This algorithm is at the top of the performance range and is notable for high performance on dense breasts.

CONCLUSION

The commercially available algorithm tested here is capable of functioning at the level of practicing radiologists, making it an attractive candidate for a digital second reader. By drawing attention to suspicious exams rather than offering a diagnosis, AI-based triage may provide positive reader bias to improve accuracy. Indeed, there is increasing evidence of the combined improvement in performance when radiologists work with AI, paving the way for AI to assist with increasing workloads and potentially eliminating obviously negative exams, while enhancing patient care.

REFERENCES

- Benchmarks for Screening Sensitivity & Specificity :: BCSC. Available at: https://www.bcsc-research.org/statistics/screening-performance-benchmarks/Benchmarks-sens-spec. (Accessed: 12th January 2021)
- Lehman, C. D. et al. National performance benchmarks for modern screening digital mammography: Update from the Breast Cancer Surveillance Consortium. Radiology 283, 49–58 (2017).
- McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. Nature 577, 89–94 (2020).
- Salim, M. et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol. 6, 1581–1588 (2020).
- Schaffter, T. et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. JAMA Netw. open 3, e200265 (2020).
- Yala, A., Schuster, T., Miles, R., Barzilay, R. & Lehman, C. A Deep Learning Model to Triage Screening Mammograms: A Simulation Study. Radiology 293, 38–46 (2019).