10.1148/ryai.2020190208

Appendix E1: Description of the AI System

The used AI system (MammoScreen V1, Therapixel) uses two groups of deep convolutional neural networks (CNN) combined together with an aggregation module.

The image-wise group, that takes as input entire images and outputs a prediction score for each individual image. This group was trained to predict the benign (absence of cancer) or malignant (presence of cancer) status of an image. We reused the architecture we submitted to the Digital Mammography DREAM Challenge (34) (called the dream-net), and we further extended it with a symmetrized version taking a decision from a pair of left and/or right craniocaudal or mediolateral oblique images instead of a single image. The symmetric dream-net is able to learn meaningful features from the difference (or similarity) between left-right images for the classification task at stake. Both dream-net and its symmetric version have the interesting property of being end-to-end (from images to prediction in one go), mitigating the risk of cumulating errors in-between several disjoint steps. However, image resolution had to be lowered down to 1152 × 832 pixels to allow efficient CNN training on current hardware. This loss of resolution may negatively impact performance as subtle details required for this classification task could be wiped out in the process. Nevertheless, efficiency of such algorithm was demonstrated when winning the DREAM competition (34). Finally, each mammogram image is inputted to both dream-net and its symmetric version, and the two results are eventually averaged to form the final prediction for this image.

The lesion-wise group, where detection of regions-of-interest and their characterization is done in two steps. First, a detection CNN (35) identifies image findings regardless of their level of suspicion and classifies them either as calcifications or soft-tissue lesion. It was trained on a dataset of 13,666 fully annotated images to exhibit a high level of sensitivity on malignant findings. Each image was carefully reviewed by a trained expert and any identifiable calcification cluster and soft-tissue lesion (including mass, asymmetry, distortion, lymph node, and opacity in general) was contoured with a bounding box delimiting its extent. Second, another CNN characterizes the level of suspicion of each finding outputted at the previous step. This characterizing CNN takes as input high-resolution patches centered on image coordinates and outputs a prediction score. It was trained to predict the benign (benign finding) or malignant (cancerous finding) status of the patch. Like image-wise group, we developed a symmetric version of this model, which outputs a prediction from a patch and its corresponding patch in the symmetric image, to allow CNN to learn meaningful symmetric features for this characterization task. For each detected finding, two predictions are eventually formed: one from the characterizing CNN and one from its symmetric version. The predictions are averaged to form the final prediction of this finding.

The aggregation module combines image-and lesion-wise predictions. An image prediction is averaged with its highest finding (ie, the highest score of all findings within that image) to form the final image prediction. Then, the new value is assigned to the highest finding, and all other findings are scaled accordingly. This method ensures adequacy between image and lesion-wise predictions: strong positive or negative predictions are only possible when both groups agree. On the contrary, a disagreement between the two would produce a result close to 0.5, mitigating the negative impact of an erroneous decision of one of the groups. By combining

conceptually different approaches, this was found capable of better generalizing to unseen data than models using a single family of CNN.

Finally, prediction scores lying in the continuous range (0.0–1.0) are discretized on a (1–10) integer scale to ease their interpretation by physicians. Discretization for each score (1–10) was determined as follows: (score 1) absence of findings (the detection CNN did not reveal anything abnormal), (score 2) the likelihood of malignancy was less than 0.5%, (score 3) LOM is between 0.5%–1%, (score 4) LOM is between 1%–2.5%, (score 5) LOM is between 2.5%–5%, (score 6) LOM is between 5%–25%, (score 7) LOM is between 25%–50%, (score 8) LOM is between 50%–75%, (score 9) LOM is between 75%–99%, and (score 10) LOM is greater than 99%.

AI System Training

A total of 60 000 screening mammograms (240 000 images), among which 5600 were positive cases confirmed either by biopsy or surgery, were collected from four different European centers for algorithm training. Mammograms originated from the following vendors: 59% were from Fuji, 24% from Hologic, 15% from GE, and 2% from Philips. A total of 25 000 image findings (14 400 negative and 10 600 positive) were made by breast radiologists who had access to radiologic and biopsy reports. The (symmetric) dream-nets were trained directly from images and their status (negative or positive). The detection CNN was trained from images and their annotations regardless of the statuses, while the (symmetric) characterization CNN were trained from image annotations and their status. For each CNN, a 10-fold cross-validation scheme was used: mammograms were equally split in 10 folds by patient (data of the same patient could not belong to more than one fold), and each model was trained on the data from 9 folds while performances were measured on the left-out fold. By permuting the left-out fold, 10 instances per model were obtained making a total of 50 instances (5 model families times 10) for the final algorithm. Therefore, each model prediction was obtained from an average of 10 model predictions instead of just one. This further improved the generalizability of our algorithm by reducing overfit as shown in Dietterich et al (36).

AI User Interface

Algorithm results were presented to physicians in a dedicated interface on a separate monitor. Unlike conventional computer-aided detection that overlay algorithm findings directly on top of images in the mammography review software, we deliberately chose to leave the review software unaltered. Doing so offers several advantages: *(a)* it allows a finer control about what is presented to users as we are not limited by the mammography review software capacity (and limitations); *(b)* it offers the same level of service to everyone without altering their preferred review software; and *(c)* it makes concurrent reading de facto possible by synchronizing the case opening in the mammography review software with its opening in MammoScreen user interface. Therefore, users were trained how to interpret information presented by the dedicated interface on a second monitor to guide their image review. The user interface consists in a representation of the four views composing the mammogram with marks placed on each finding showing their type and LOM. The maximum LOM per breast (ie, maximum LOM encountered in all findings of the two views composing a breast) was reported on a 1–10 scale to quickly indicate to readers which breast was given the highest LOM. Similarly, the entire mammogram was given a LOM equal to the maximum LOM of both breasts. A consistent color scheme was used to outline

findings and indicate breast and mammogram LOM, from green (LOM of 1) to red (LOM of 10). The combination of per-case, per-breast and per-finding information allows for a fast coarse-to-fine interpretation of the algorithm results by answering three questions: *(a)* Is the case deemed suspicious? *(b)* If yes, which breast (s)? and *(c)* If yes, where in the breast (s)?

Appendix E2–Reading Time Analysis

The goal of the reading time analysis was to learn how reading time changes when using the AI-system and whether it changed between the first and the second reading session to evaluate the learning effect between the first and the second time that readers used the AI system.

All values beyond 10 minutes were considered as outliers and excluded from this analysis because considered as not representative of the real clinical practice. Table E1 summarizes the characteristics of the excluded values.

Table E1: Outliers values characteristics.

Nr of removed values	Average value, s	Median value, s	SD	Min, s	Max, s	Lower 95% CI bound, s	Upper 95% CI bound, s
51	1721.40	1274.36	1014.16	610.26	4430.72	1436.16	2006.63

To model the reading time as a function of the covariates, a Poisson generalized linear model was used. Fixed covariates are *reading session* (two levels) and *reading condition* (two levels). The interaction term was *reading session* x *reading condition*. The 'stats' package in the software R (37) was used to fit the model. Table E2 reports the obtained regression parameters.

Table E2: Estimated regression parameters, standard errors, z-values and *P* values for the presented Poisson GLM.

	Estimate	Std. error	z-value	*P* value
Intercept	4.14	0.003	1339.34	<2e-16
Reading condition Assisted	0.14	0.004	32.10	<2e-16
Reading session Second	–0.09	0.004	–20.78	<2e-16
Reading condition Assisted: Reading session Second	–0.05	0.006	–8.58	<2e-16

Both considered variables affect significantly the reading time individually as well as the interaction of reading method and session. Tables E3 and E4 report the mean reading time the standard deviation and the number of the included values by reading session, reading condition and readers.

Table E3: Average reading time and standard deviation for each of the possible combination of reading condition and reading session.

		Mean, s	SD, s	Included values
All		63.52	45.51	6669
1st reading session		67.35	46.41	3332
2nd reading session		59.69	44.28	3337
Unaided		60.00	43.15	3338
Assisted		67.04	47.51	3331
1st reading session	Unaided	62.79	41.94	1667
	Assisted	71.93	50.08	1665
2nd reading session	Unaided	57.22	44.16	1671
	Assisted	62.16	44.27	1666

The last column reports the number of the considered reading times for each category after the exclusion of outlier values.

Table E4: Average reading time, standard deviation and number of reading times across readers for both reading conditions and reading sessions.

Reader	Years of Experience	First reading session [s]						Second reading session [s]					
		Unaided			With AI			Unaided			With AI		
		Mean	SD	#	Mean	SD	#	Mean	SD	#	Mean	SD	#
1	8	63.18	41.92	120	83.22	56.25	119	66.12	60.35	120	55.14	19.30	119
2	13	66.68	52.39	119	64.88	46.10	119	61.19	42.42	119	76.51	69.62	119
3	12	63.65	50.54	118	82.64	39.22	120	54.67	32.31	120	50.66	28.57	120
4	5	77.25	35.06	118	70.87	35.84	119	61.33	45.25	120	86.59	56.13	119
5	25	74.29	42.39	120	88.20	58.75	119	85.70	44.00	119	82.16	31.96	120
6	23	62.09	50.16	119	61.27	35.34	118	44.96	30.63	119	57.25	27.31	119
7	5	68.22	37.13	120	83.36	58.92	119	43.43	24.66	119	49.91	29.73	119
8	3	30.46	24.50	119	64.92	76.39	118	47.02	60.63	119	53.36	79.65	118
9	6	68.89	32.71	118	78.23	28.66	119	66.01	42.07	120	54.69	45.41	119
10	0	68.39	60.66	119	59.79	57.39	119	55.58	31.35	119	76.00	41.41	119
11	21	53.21	23.64	119	70.96	38.41	118	61.90	51.16	119	57.68	30.11	119
12	7	49.29	26.77	120	52.42	41.98	119	39.75	28.16	120	42.45	32.19	119
13	10	53.67	25.05	119	61.73	31.02	120	50.64	39.57	120	59.99	25.25	118
14	9	79.93	37.75	119	84.30	57.39	119	62.81	47.17	118	67.74	28.49	119

References

34. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. JAMA Netw Open 2020;3(3):e200265.

35. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. arXiv:170802002 [cs]. [preprint]. http://arxiv.org/abs/1708.02002. Posted August 7, 2017. Accessed May 22, 2019.

36. Dietterich TG. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Berlin, Germany: Springer, 2000; 1–15 https://doi.org/10.1007/3-540-45014-9_1.

37. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2019. https://www.R-project.org. Accessed May 2020.