Radiology

ORIGINAL RESEARCH • BREAST IMAGING

Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms

Ayelet Akselrod-Ballin, PhD • Michal Chorev, PhD* • Yoel Shoshan, BSc • Adam Spiro, PhD • Alon Hazan, MSc • Roie Melamed, PhD • Ella Barkan, MSc • Esma Herzel, MSc • Shaked Naor, BSc • Ehud Karavani, BSc • Gideon Koren, MD • Yaara Goldschmidt, PhD • Varda Shalev, MD, MPH • Michal Rosen-Zvi, PhD • Michal Guindy, MD, MPH*

From the Department of Healthcare Informatics, IBM Research, IBM R&D Labs, University of Haifa Campus, Mount Carmel, Haifa 3498825, Israel (A.A.B., M.C., Y.S., A.S., A.H., R.M., E.B., S.N., E.K., Y.G., M.R.Z.); MaccabiTech, MKM, Maccabi Healthcare Services, Tel Aviv, Israel (E.H., G.K., V.S.); and Department of Imaging, Assuta Medical Centers, Tel Aviv, Israel (M.G.). Received November 19, 2018; revision requested January 16, 2019; revision received April 14; accepted April 29. **Address correspondence to** M.C. (e-mail: *michalc@il.ibm.com*).

* A.A.B. and M.C. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

Radiology 2019; 00:1–12 • https://doi.org/10.1148/radiol.2019182622 • Content codes: BR IN

Background: Computational models on the basis of deep neural networks are increasingly used to analyze health care data. However, the efficacy of traditional computational models in radiology is a matter of debate.

Purpose: To evaluate the accuracy and efficiency of a combined machine and deep learning approach for early breast cancer detection applied to a linked set of digital mammography images and electronic health records.

Materials and Methods: In this retrospective study, 52 936 images were collected in 13 234 women who underwent at least one mammogram between 2013 and 2017, and who had health records for at least 1 year before undergoing mammography. The algorithm was trained on 9611 mammograms and health records of women to make two breast cancer predictions: to predict biopsy malignancy and to differentiate normal from abnormal screening examinations. The study estimated the association of features with outcomes by using *t* test and Fisher exact test. The model comparisons were performed with a 95% confidence interval (CI) or by using the DeLong test.

Results: The resulting algorithm was validated in 1055 women and tested in 2548 women (mean age, 55 years ± 10 [standard deviation]). In the test set, the algorithm identified 34 of 71 (48%) false-negative findings on mammograms. For the malignancy prediction objective, the algorithm obtained an area under the receiver operating characteristic curve (AUC) of 0.91 (95% CI: 0.89, 0.93), with specificity of 77.3% (95% CI: 69.2%, 85.4%) at a sensitivity of 87%. When trained on clinical data alone, the model performed significantly better than the Gail model (AUC, 0.78 vs 0.54, respectively; $P < .004$).

Conclusion: The algorithm, which combined machine-learning and deep-learning approaches, can be applied to assess breast cancer at a level comparable to radiologists and has the potential to substantially reduce missed diagnoses of breast cancer.

Online supplemental material is available for this article.

Breast cancer is the second leading cause of cancer-related deaths and the most commonly diagnosed cancer in women across the world (1). Digital mammography (DM) is the primary imaging modality of breast cancer screening in women who are asymptomatic. In a diagnostic workup setting (2), DM has been shown to reduce breast cancer mortality (3). In standard clinical practice, a radiologist reads mammograms and classifies the findings according to the American College of Radiology (4) Breast Imaging Reporting and Data System (BI-RADS) lexicon. An abnormal finding depicted at DM typically requires a diagnostic workup, which may include additional mammographic views or possibly additional imaging modalities. If a lesion is suspicious for cancer, further evaluation with a biopsy is recommended. Analyzing these images is challenging because of the subtle differences between lesions and background fibroglandular tissue, different lesion types, the nonrigid nature of the breast, and the relatively small proportion of cancers in a screening population of women at average risk (2). This leads to substantial intraobserver and interobserver variability (5). The average performance measures for screening mammography by a radiologist was reported by Lehman et al (6) to be 86.9% sensitivity and 88.9% specificity.

Breast cancer risk prediction models on the basis of clinical features can help physicians estimate the probability of an individual or population to develop breast cancer within certain time frames. As a result, they are often used to recommend an individual screening plan. In a systematic survey of risk prediction models, Meads et al (7) reported a limited performance when applied to general populations (area under the receiver operating characteristic curve [AUC], ≤0.67; 95% confidence interval [CI]: 0.65, 0.68), and showed improved results when applied to high-risk populations (AUC, 0.76; 95% CI: 0.70, 0.82).

Abbreviations

AUC = area under the receiver operating characteristic curve, BI-RADS = Breast Imaging Reporting and Data System, CI = confidence interval, DL = deep learning, DM = digital mammography, ML = machine learning

Summary

A deep learning algorithm trained on a linked data set of mammograms and electronic health records achieved breast cancer detection accuracy comparable to radiologists as defined by the Breast Cancer Surveillance Consortium benchmark for screening digital mammography and revealed additional clinical risk features.

Key Points

- A deep learning algorithm predicted breast malignancy detected within 12 months from the index examination (area under the receiver operating curve [AUC], 0.91; specificity of 77.3% at a sensitivity of 87%).
- The algorithm identified breast cancer in 34 of 71 (48%) women in whom the initial radiologist interpretation was negative for cancer but in whom cancer was detected within a year.
- The deep learning algorithm improved risk prediction over the Gail model (AUC, 0.78 vs 0.54, respectively; $P < .004$).
- The deep learning algorithm identified white blood cell profiles and thyroid function tests as associated with breast cancer despite that these factors are not currently integrated in published risk scores.

Machine learning (ML) and its subdiscipline, deep learning (DL), have recently obtained good results in the health care domain (8–13). Although DL-based models are increasingly used to analyze health care data, the efficacy of traditional computer-aided detection systems is still controversial (14–16).

We hypothesized that a model combining ML and DL (hereafter, ML-DL model) can be applied to assess breast cancer at a level comparable to radiologists and therefore be accepted in clinical practice as a second reader. The purpose of our study was to evaluate the performance of an ML-DL model for early breast cancer prediction when applied to a large linked data set of detailed electronic health records and digital mammography.

Materials and Methods

This retrospective study was approved by the research ethics review board of Assuta Medical Centers and they waived need to obtain written informed consent. The data were collected and managed by Maccabi Health Services. The authors from Maccabi Health Services and Assuta Medical Centers (E.H., G.K., V.S., and M.G.) obtained the approval for the retrospective study and the anonymization process. The analysis was conducted by all other authors who were employees at IBM at the time of the study. The IBM code is provided at *https://www.research.ibm.com/haifa/dept/imt/dl-breastcancer/dl-breastcancer-login.shtml*.

Study Design and Patient Eligibility

The cohort was composed of women who underwent at least one DM examination between 2013 and 2017 in one of the five Assuta Medical Centers imaging facilities and had at least 1 year of clinical history in Maccabi Health Services before undergoing DM. We did not exclude images on which there were common foreign bodies (eg, clips, markers, and pacemakers), and theses constituted 14.0% of our cohort (1848 of 13 214). We excluded women with a history of breast cancer, previous breast operations (eg, lumpectomy and mammoplasty), previous radiation therapy in the breast, chemotherapy, implants, and mammograms on which the biopsy side was undetermined. Studies that were BI-RADS category 1–2 without 2 years of normal follow-ups were also excluded (Fig 1).

For each woman, we considered the first DM examination during our study period as the index examination if it met the inclusion and exclusion criteria. The model used clinical data before the index mammogram (Appendix E1 [online]).

We split the data set into three nonoverlapping sets: 73% training (9611 women with 38 444 images), 8% validation (1055 women with 4220 images), and 19% test (2548 women with 10 192 images). The breakdown into subcohorts is summarized in Table E1 (online). A false-negative interpretation by a radiologist was defined as an index mammogram that was read as either BI-RADS categories 1 or 2 and subsequently found to have a malignant breast lesion within 12 months from the index mammogram.

Outcome Definitions

Our study focused on evaluating two clinical objectives for DM screening by using an ML-DL model:

Objective 1, prediction of malignancy (evidenced by biopsy positive for cancer): Women were considered to have cancer if the Maccabi Health Services registry and pathologic database indicated the diagnosis of breast cancer by biopsy within 12 months from the index examination. Examinations that were considered positive for cancer comprised any pathologic specimens for cancer including ductal carcinoma in situ. All other examinations were considered negative for cancer including biopsy examinations negative for cancer.

Objective 2, identification of normal DM examinations: A normal DM examination was defined as a BI-RADS category 1–2 on a mammogram with normal follow-up examinations for at least 2 years after the index examination. Benign and malignant biopsies and BI-RADS category 3 were not considered normal examinations. BI-RADS category 1–2 on mammograms with insufficient follow-up period were excluded.

Of note, the two objectives are not the exact inverse of each other because biopsies negative for cancer and BI-RADS category 3 examinations were always part of the complementary outcome.

The results were reported for each breast to better facilitate a comparison with other models that evaluated their performance per breast. The results were also determined at the individual woman level to allow a comparison with the clinical setting.

Development of the ML-DL Model

We created an ML-DL model that combined a set of algorithms to achieve the two listed prediction objectives. For each woman, the input was the DM standard four-view images and the detailed clinical histories (Fig 2, *A*).

First, we used XGBoost (17) (version 0.81; *https://xgboost.ai*), an open-source Python implementation of gradient boosting

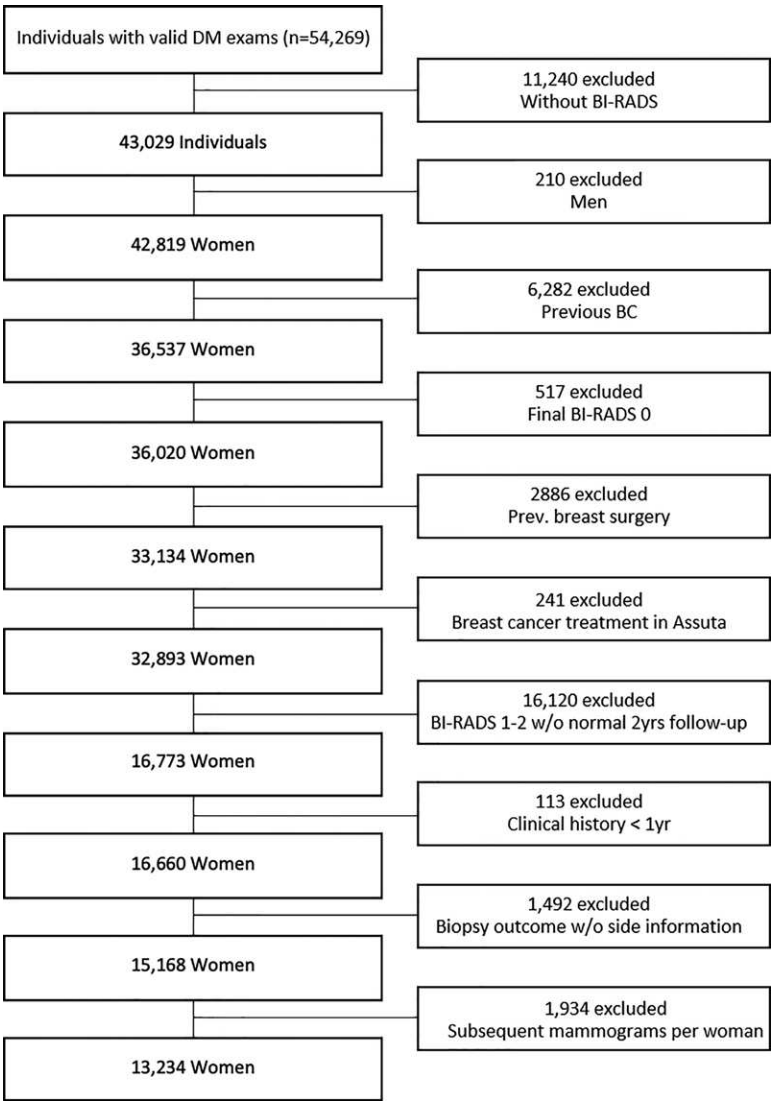

Figure 1: Flowchart of study inclusion and exclusion on the basis of the Strengthening Reporting of Observational Studies in Epidemiology (known as STROBE). BC = breast cancer, BI-RADS = Breast Imaging Reporting and Data System, DM = digital mammography, w/o = without.

machines classifier, to identify a subset of the clinical features showing the greatest contribution to prediction of a biopsy positive for cancer (Fig 2, *B*). These were fused into a deep neural network and trained on each DM image for each of the prediction objectives (Appendix E1; Tables E2, E3 [online]). For robustness, we trained the deep neural network algorithm on three random 80% partitions of the training set, with and without the subset of clinical features; this resulted in six algorithms for each objective (Fig 2, *C*), which were then used as an ensemble average for each objective. After this step, we extracted features from the last fully connected layer and the estimated probability for both objectives from the deep neural network (18) ensemble for each image. We then combined those imaging features with the entire set of clinical features. Therefore, the probability of cancer at the breast level was acquired from a feature set composed of imaging features obtained from both views of the breast (craniocaudal and mediolateral oblique) for both prediction objectives and the entire set of clinical features (Fig 2, *D*). The final probability for either a biopsy positive for cancer or so-called normal identification was estimated by using XGBoost (Fig 2, *E*). When analyzed at the individual woman level, we assigned the higher probability of the two values obtained by the ML-DL model for the breast level, similar to clinical practice. To view the areas on the image that were suspicious for cancer, we used the technique developed by Fong and Vedaldi (19), which identified the smallest continuous areas in the input image that contributed the most to the malignancy prediction by the ML-DL model.

Statistical Analysis

We used Fisher exact test and Student *t* test to estimate the univariate association of features with the cancer-positive biopsy outcome. We corrected for multiple hypotheses by employing the Bonferroni correction. The significance of the differences between AUCs was estimated by using a 95% CI or DeLong test. A *P* value less than .05 was considered to indicate a statistically significant difference.

The percentage of women in our cohort who had subsequent biopsy procedures was higher than the actual percentage of women who underwent the procedures at Assuta Medical Centers. This occurred because the data initially transferred to us covered their more severe cases, which could bias the number of cancerous specimens in the data. This reflects neither their distributions at Assuta facilities nor those previously reported in the literature, specifically by the Breast Cancer Surveillance Consortium (6) (0.58% biopsy positive for cancer, 2% biopsy negative for cancer). We used a bootstrapping approach to counteract the potential effect of this bias on the model's performance. This adjustment was essential to maintain real-world distributions for the data, especially when the mammographic examinations with benign results are considered part of the opposite label in both prediction objectives. We sampled with replacement a proportional number of women with normal, benign, and malignant results on mammograms according to the proportions of individuals in the Breast Cancer Surveillance Consortium and calculated the corresponding AUC. We repeated this process 1000 times to obtain a mean AUC and a 95% CI (20). The percentages of malignant and benign mammography for individuals in the literature are reported per population. We estimated that the prevalence of biopsies positive and negative for cancer per breast is roughly half of the one reported for individuals (6).

Results

Study Population

The training set consisted of data for 9611 women (mean age, 56 years ± 10 [standard deviation]; body mass index of 26.9 kg/m^2 ± 5.4 (Table 1). In the training set, 1049 women (10.9%)

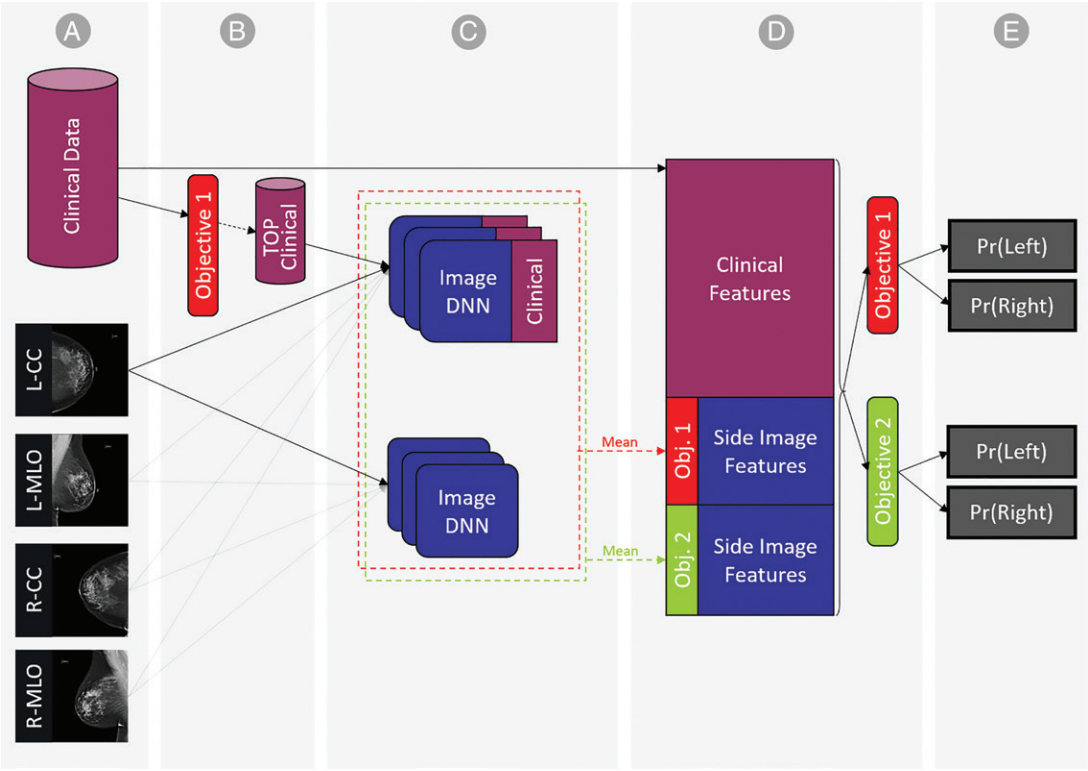

Figure 2: Outline of machine- and deep-learning model. *A*, Input is standard four-view mammograms and clinical data (1343 features). *B*, Selection of top clinical features for deep neural network by using the XGBoost algorithm. *C*, Ensemble of deep neural network (DNN) algorithms trained separately for each prediction objective (objective 1, biopsy positive for cancer; objective 2, normal examination identification) per image either with or without the subset of clinical features. *D*, A woman's side is represented by a feature set composed of the imaging features obtained from both views of the breast (craniocaudal [CC] and mediolateral oblique [MLO]) for both prediction objectives (Obj) and joined with the entire set of clinical features. *E*, Predicted probability of outcome (Pr) per objective and per side is obtained by training an XGBoost algorithm on the vector formed in *D*. L = left, R = right.

had a biopsy positive for cancer within 1 year from their index examination, 1903 (19.8%) had a biopsy negative for cancer, 247 (2.6%) were assigned BI-RADS category 3 without a subsequent biopsy, and 6412 (66.7%) had consistently normal (BI-RADS category 1–2) mammography for at least 2 years. The data set put aside for testing consisted of 2548 women (mean age, 55 years ± 10; body mass index, 26.8 kg/m^2 ± 5.3). Of this set, 289 women (11.3%) had a biopsy positive for cancer, 501 (19.7%) had a biopsy negative for cancer, 70 (2.7%) were assigned BI-RADS category 3 without a subsequent biopsy, and 1688 (66.2%) had normal examinations (validation set is in Table 1; statistics by BI-RADS are in Tables E4, E5 [online]).

A total of 102 of 13 234 women (0.8%) had false-negative findings on mammograms as read by radiologists. To analyze our model's success in identifying those examinations, we eliminated them from the training set. Instead, we inserted these mammograms into the validation and test sets: 31 of 102 were added to the validation set (to determine the sensitivity operation point threshold) and 71 of 102 were added to the test set (to reflect their existence in real-world settings).

We integrated each woman's clinical data with the image information (data source in Appendix E1 [online]). For each woman in the linked data set, we extracted all of the 1343 available clinical features, including clinical features previously recognized as risk factors for breast cancer (7).

Women with biopsies positive for cancer tended to be older than those without biopsies positive for cancer (mean age, 59 vs 55 years, respectively; Bonferroni adjusted $P < .001$; Table 2). They had higher body mass index (last measured mean, 27.5 kg/m^2 vs 26.8 kg/m^2, respectively; $P < .003$), and had more indications of symptoms (lump, nipple retraction, or discharge; 516 of 1449 with indications vs 1532 of 11 765, respectively; $P < .004$). Women with a biopsy positive for cancer also tended to have a lower number of relatives with breast cancer in general (463 of 1149 vs 4630 of 11 765, respectively; $P < .003$) and first-degree relatives in particular (281 of 1149 vs 3329 of 11 765, respectively; $P < .001$). This is opposite of what is generally expected, and serves to further indicate the need to adjust for bias in the data. Please see a complete list of features association with the outcome of biopsy positive for cancer in Table E6 (online). Table E7 (online) lists the same association for first-examination individuals.

Testing of the ML-DL Model

We evaluated the two prediction objectives in the test sample of the following four cohorts: *(a)* the general cohort (the entire test sample); *(b)* exclusion of findings suspicious for cancer that only appeared on US images: a subcohort of the test sample, in which women with findings suspicious for cancer that were detected only on US images with no evidence on mammograms were excluded (final BI-RADS cat-

Table 1: Women's Characteristics in Training, Validation, and Test Sets

Characteristic	Training Set	Validation Set	Test Set
No. of women	9611	1055	2548
Age (y)*	56 ± 10	55 ± 10	55 ± 10
Body mass index*	26.9 ± 5.4	27 ± 5.5	26.8 ± 5.3
Age first menstruation*	13 ± 1	13 ± 1	13 ± 1
Postmenopause	3109 (32.4)	348 (33.0)	766 (30.0)
1-year outcome			
Biopsy positive for cancer	1049 (10.9)	111 (10.5)	289 (11.3)
Biopsy negative for cancer	1903 (19.8)	192 (18.2)	501 (19.7)
BI-RADS category 3†	247 (2.6)	25 (2.4)	70 (2.8)
Normal examination‡	6412 (66.7)	727 (68.9)	1688 (66.3)

Note.—Unless otherwise indicated, data are numbers of women and data in parentheses are percentages. Because the machine learning–deep learning model made use of the standard four-view images, we used four images per woman. BI-RADS = Breast Imaging and Data Reporting System.

* Data are means ± standard deviation.

† BI-RADS category 3 found at examination and no subsequent biopsy procedure within 1 year.

‡ Normal examinations are index test examinations with final BI-RADS category of 1–2 with at least 1 years of normal follow-up examinations.

Table 2: Association of Features of Interest with Biopsy Positive for Cancer

Parameter	No. of Women	Women with Biopsy Positive for Cancer	Women with Normal or Negative Biopsy	Adjusted *P* Value
Age	13214 (100)	59 ± 13	55 ± 10	<.001
Most recent body mass index	12839 (97.2)	27.5 ± 5.5	26.8 ± 5.4	<.003
Age at first menstruation	10524 (79.6)	13 ± 1	13 ± 1	>.99
Postmenopause	2885 (21.8)	405 ± 28	2480 ± 21	<.004
Breast radiology history				
Previous BI-RADS breast density > 2	2699 (20.4)	211 ± 15	2488 ± 21	<.001
Previous benign breast disease	1611 (12.2)	107 ± 7	1504 ± 13	<.003
No. of previous breast imaging examinations	6735 (51.0)	0.55 ± 0.71	0.63 ± 0.69	<.005
Family history				
First-degree family member with breast cancer	3610 (27.3)	281 ± 19	3329 ± 28	<.001
Any family member with breast cancer	5093 (38.5)	463 ± 32	4630 ± 39	<.003
Any family member with ovarian or breast cancer	5279 (40.0)	488 ± 34	4791 ± 41	<.001
Medications				
Past use of fertility hormones	1228 (9.3)	100 ± 7	1128 ± 10	<.05
Past or present use of progesterone	4089 (31.0)	359 ± 25	3730 ± 32	<.003
Symptoms				
Current symptom (ie, palpable lump, nipple retraction, or discharge)	2048 (15.5)	516 ± 36	1532 ± 13	<.004
Past symptom	2109 (16.0)	115 ± 8	1994 ± 17	<.001
Current lump detected by doctor	1821 (13.8)	451 ± 31	1370 ± 12	<.009
Past lump detected by doctor	1943 (14.7)	101 ± 7	1842 ± 16	<.001

Note.—Data in parentheses are percentages. Mean data are ± standard deviation. Complete list of features for the general cohort may be found in Table E6, for first-examination individuals in Table E7 (online). *P* value is Bonferroni adjusted by the number of features. *P* values less than .05 are considered to indicate statistical significance. BI-RADS = Breast Imaging Reporting and Data System.

egories: DM, 1–2; US, ≥3). Because our model was trained on DM images alone, testing our model on this subcohort seemed appropriate; *(c)* first examination only, a subcohort limited to the first DM in a woman; and *(d)* first examination and by excluding findings suspicious for cancer that only appeared at US.

The results were analyzed at the breast level (Table 3, Fig 3) and individual level (Table E8 [online]). The following results refer to the breast level.

Overall, the ML-DL models that combined information from both images and clinical data performed better than the ML-DL models trained by using images or clinical data alone.

Table 3: Results of the Prediction Objectives Compared with Deep-Learning Models on the Breast Level

Prediction Objective	AUC*	Specificity with Sensitivity of 87%†	Specificity with Sensitivity of 80%†
Objective 1, prediction of malignancy			
General cohort			
All features	0.91 (0.89, 0.93)	3139/4061 (77.3)[69.2, 85.4]	3537/4061 (87.1) [81.5, 92.7]
DM images only	0.88 (0.86, 0.90)	2835/4061 (69.8) [59.3, 80.3]	3314/4061 (81.6) [74.5, 88.7]
Clinical only	0.78 (0.75, 0.81)	1888/4061 (46.5) [38.5, 54.5]	2392/4061 (58.9) [50.9, 66.9]
Excluding US-only suspicious findings subcohort			
All features	0.94 (0.93, 0.95)	3196/3691 (86.6) [80.7, 92.5]	3444/3691 (93.3) [89.3, 97.3]
DM images only	0.91 (0.89, 0.93)	2942/3691 (79.7) [78.7, 80.8]	3296/3691 (89.3) [85.9, 92.7]
Clinical only	0.81 (0.78, 0.84)	1923/3691 (52.1) [45.4, 58.8]	2399/3691 (65) [55.4, 74.6]
First examination subcohort			
All features	0.94 (0.93, 0.95)	1055/1213 (87) [82.7, 91.3]	1132/1213 (93.3) [90.0, 96.6]
DM images only	0.93 (0.91, 0.95)	1008/1213 (83.1) [76.8, 89.4]	1095/1213 (90.3) [87.4, 93.2]
Clinical only	0.85 (0.82, 0.88)	787/1213 (64.9) [53.2, 76.6]	929/1213 (76.6) [70.1, 83.1]
First examination and excluding findings suspicious for cancer at US only subcohort			
All features	0.96 (0.95, 0.97)	979/1047 (93.5) [90.4, 96.6]	1020/1047 (97.4) [95.7, 99.1]
DM images only	0.95 (0.94, 0.96)	951/1047 (90.8) [87.3, 94.3]	997/1047 (95.2) [93.6, 96.8]
Clinical only	0.86 (0.84, 0.88)	689/1047 (65.8) [57.5, 74.1]	805/1047 (76.9) [70.5, 83.3]
Breast Cancer Surveillance Consortium radiologists		89	
DREAM model	0.87		81‡
Objective 2, identification of normal DM examinations			
General cohort			
All features	0.85 (0.84, 0.86)	648/1016 (63.8) [60.8, 66.8]	754/1016 (74.2) [71.1, 77.3]
DM images only	0.80 (0.79, 0.81)	555/1016 (54.6) [51.5, 57.7]	650/1016 (64) [61.6, 66.4]
Clinical only	0.80 (0.79, 0.81)	533/1016 (52.5) [48.5, 56.5]	656/1016 (64.6) [62.6, 66.6]
Subcohort excluding findings suspicious for cancer found at US only			
All features	0.85 (0.84, 0.86)	374/597 (62.6) [58.2, 67.0]	439/597 (73.5) [71.1, 75.9]
DM images only	0.79 (0.78, 0.80)	304/597 (51) [47.9, 54.1]	367/597 (61.4) [58.7, 64.1]
Clinical only	0.78 (0.77, 0.79)	252/597 (42.2) [36.4, 48.0]	346/597 (58) [55.0. 61.0]

Table 3 *(continues)*

For the objective of predicting malignancy in the general cohort, the ML-DL model achieved an AUC of 0.91 (95% CI: 0.89, 0.93). Images alone achieved an AUC of 0.88 (95% CI: 0.86, 0.90), and clinical data alone achieved an AUC of 0.78 (95% CI: 0.75, 0.81). By adding clinical features to the image-based features in the ML-DL model, the AUC improved by 3.2%, and the specificity by 6.7% at sensitivity of 87%. Because the model's output is a probability and not a final

Table 3 *(continued)*: Results of the Prediction Objectives Compared with Deep-Learning Models on the Breast Level

Prediction Objective	AUC*	Specificity with Sensitivity of 87%†	Specificity with Sensitivity of 80%†
First examination subcohort			
All features	0.88 (0.87, 0.89)	341/472 (72.3) [70.6, 74.0]	372/472 (78.8) [77.5,
DM images only	0.84 (0.83, 0.85)	297/472 (62.9) [60.4, 65.4]	339/472 (71.8) [69.3, 74.3]
Clinical only	0.85 (0.84, 0.86)	313/472 (66.4) [64.5, 68.3]	346/472 (73.4) [71.1, 75.7]
First examination and subcohort excluding findings suspicious for cancer found at US only			
All features	0.88 (0.87, 0.89)	210/289 (72.6) [69.4, 75.8]	230/289 (79.6) [77.1, 82.1]
DM images only	0.83 (0.82, 0.84)	163/289 (56.5) [52.0, 61.0]	196/289 (67.8) [65.0, 70.6]
Clinical only	0.83 (0.82, 0.84)	188/289 (65.2) [64.1, 66.3]	207/289 (71.7) [69.7, 73.7]
Normal identification with deep learning§	0.61	20‡	30‡

Note.—Unless otherwise indicated, data are numerator/denominator. Individual level results are reported in Table E8 (online). AUC = area under the receiver operating characteristic curve, The excluding US-only suspicious findings subcohort excluded examinations in which the digital mammography final BI-RADS assessment was 1 or 2 and the US final BI-RADS assessment was 3 or higher. DREAM = Dialogue for Reverse Engineering Assessments and Methods, DM = digital mammography.

* Data in parentheses are 95% confidence intervals.

† Data in parentheses are percentages; data in brackets are 95% confidence intervals.

§ Results were previously obtained by Geras et al (32).

decision, we set a threshold for a final decision by choosing an operation point of 87% sensitivity on the validation set, consistent with the average radiologist operation point from the literature (6). On the basis of images and clinical data, the ML-DL model correctly interpreted 34 of 71 (48%) mammographic examinations in women with false-negative findings interpreted by radiologists in the test set. Without clinical data, the ML-DL model correctly read 32 of 71 (45%) false-negative findings on mammograms. By using a second operation point of 99% specificity, the sensitivity of the ML-DL model was 52% (187 of 360; 95% CI: 47%, 57%) and it correctly interpreted 11 of 71 (15%) false-negative findings on mammograms in the test set. Without the clinical data, the ML-DL model correctly read eight of 71 (11%) false-negative findings on mammograms in 87% sensitivity operation point. To detect the areas in the image that the model found most suspicious for malignancy, we employed a technique developed by Fong and Vedaldi (19). Figure 4 and Figure E1 (online) show examples for index examinations that had false-negative interpretations by radiologists but were detected by the ML-DL model for the operation point at 87% sensitivity. The identified areas suspicious for malignancy were verified by expert breast radiologists in a retrospective analysis on the basis of follow-up images and reports. In the subcohort that excluded findings suspicious for cancer that only appeared on US images, the ML-DL model performed better (AUCs, 0.94 [95% CI: 0.93, 0.95], 0.91 [95% CI: 0.89, 0.93], and 0.81 [95% CI: 0.78, 0.84] by using combined images and clinical data, images only, and clinical only data, respectively). Performance was similar for the subcohort focusing on individuals who were undergoing their first DM examination (AUCs, 0.94 [95% CI: 0.93, 0.95], 0.93 [95% CI: 0.91, 0.95], and 0.85 [95% CI: 0.82, 0.88], respectively), and in a cohort that combined the two constraints, excluding findings suspicious for cancer that only appeared on US images and by focusing on first-examination individuals, results were AUC of 0.96 (95% CI: 0.95, 0.97), 0.95 (95% CI: 0.94, 0.96), and 0.86 (95% CI: 0.83, 0.89), respectively.

For the objective of a so-called normal examination differentiation, the ML-DL model achieved an AUC of 0.85 (95% CI: 0.84, 0.86) by using both images and clinical information, 0.80 (95% CI: 0.79, 0.81) by using images, and 0.80 (95% CI: 0.79, 0.81) by using clinical information alone. By focusing on women undergoing their first DM examination, our results improved (0.88 [95% CI: 0.87, 0.89], 0.84 [95% CI: 0.83, 0.85], and 0.85 [95% CI: 0.84, 0.86], respectively). In the subcohort that excluded findings suspicious for malignancy found at US and the additional subcohort that included first-examination individuals, the ML-DL model's performance was reduced (0.85 [95% CI: 0.84, 0.86], 0.79 [95% CI: 0.78, 0.80], and 0.78 [95% CI: 0.77, 0.79], respectively). Adding clinical features in addition to image-based features improved the AUC obtained by the ML-DL model by 6.8%, and specificity (at sensitivity of 87%) by up to 16.8%. By using an operation point with high sensitivity (99%), the specificity was 22% (497 of 2259; 95% CI: 15%, 29%).

Figure 5 shows, in descending order, the top 15 clinical features that had the most influence on positive biopsy prediction and normal differentiation on the general test cohort (Fig 5, *A*, *B*, respectively), and for first-examination individuals (Fig 5, *C*, *D*, respectively). Please see Appendix E1 (online) for feature

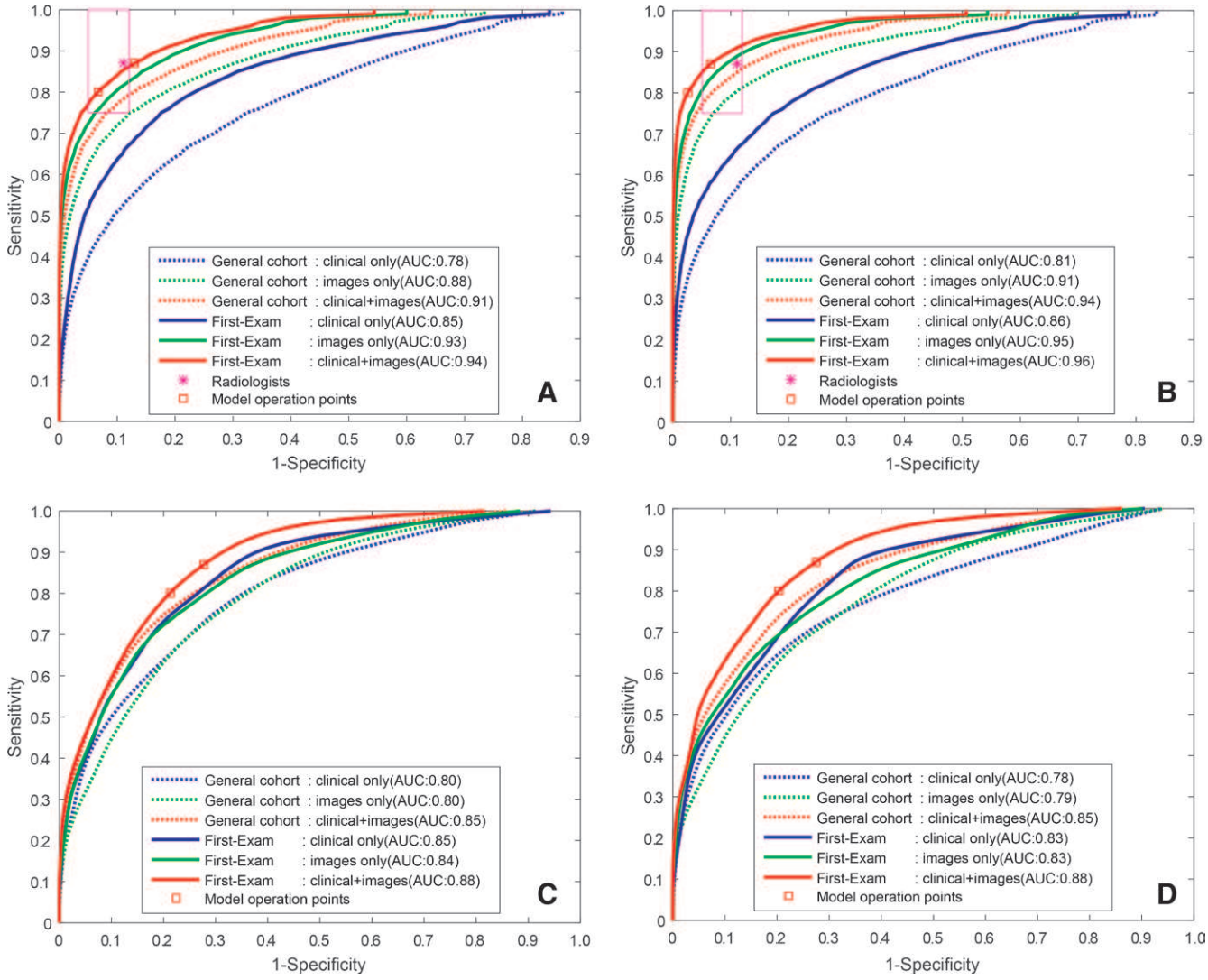

Figure 3: Comparison of the classification performance on the two prediction objectives and four cohorts. Performance is reported on three sets of features (clinical data based, images based, and based on both imaging and clinical features). The magenta rectangle (*A*, *B*) for the malignancy prediction objective corresponds to the American benchmark for the acceptable range of screening digital mammography. The red squares (*A–D*) represent the machine learning–deep learning model's specificity at sensitivity of 80% and 87%. *A*, Results of the malignancy prediction objective in the general cohort (complete test set). *B*, Results of the malignancy prediction objective in the subcohort that excluded women with findings suspicious for cancer that only appeared on US images (ie, excluding examinations in which digital mammography depicted Breast Imaging Reporting and Data System [BI-RADS] category 1–2 and US depicted BI-RADS ≥3 lesions). *C*, Results of the normal examinations identification objective in the general cohort (complete test set). *D*, Results for the normal examinations identification objective in the subcohort that excluded findings suspicious for cancer that only appeared on US images. AUC = area under the receiver operating characteristic curve.

contribution analysis corresponding to this figure. Table E9 (online) reports our sensitivity analyses on subsets of clinical features, including all features except those of symptoms that may lead to a diagnostic examination (lump, nipple discharge, and nipple retraction), and all features without past breast density and BI-RADS (from previous radiologist reports if any exist).

Discussion

The machine learning (ML)–deep learning (DL) model exhibits the potential to reduce the likelihood of breast cancer misdiagnosis. Importantly, the algorithm identified 34 of 71 (48%) false-negative findings missed by radiologists in an operation point set at 87% sensitivity.

Integrating imaging and clinical information in our ML-DL model achieved an AUC of 0.91 (95% CI: 0.89, 0.93) with 77.3% specificity at 87% sensitivity for prediction of biopsy positive for cancer, and an AUC of 0.85 (95% CI: 0.84, 0.86) on the identification of normal examinations. In the subcohort where radiologists' final BI-RADS assessment could be estimated by using DM alone, without the assistance of US, the performance was better for malignancy prediction. Moreover, the results are well within the acceptable range of radiologists at screening DM as described by the Breast Cancer Surveillance Consortium benchmark (6) (sensitivity 75%; and specificity, 88%–95%) for all cohorts. Compared with existing clinically based risk models (7), our prediction with clinical data alone outperformed the Gail

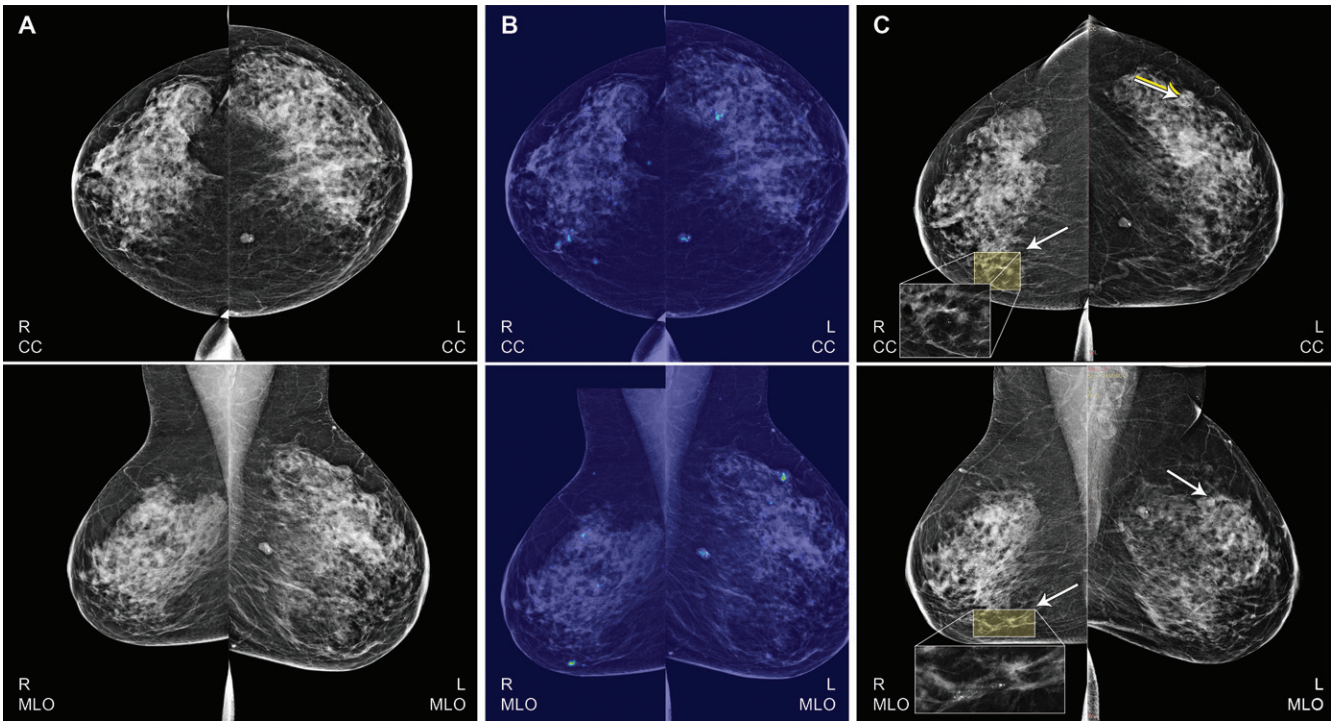

Figure 4: Images in a 64-year-old woman show prediction of malignancy in an examination interpreted as false-negative findings. *A*, At routine mammography, cysts and a solid mass in the left breast were reported as unchanged from an examination preformed 2 years earlier. In her right breast, small microcalcifications at the lower inner quadrant were not reported. *B*, The machine learning–deep learning model classification, as viewed by the technique by Fong and Vedaldi (19); the heat-map color ranged from blue (not suspicious for cancer) to red (highly suspicious for cancer). *C*, A 0.8-cm lesion at 12 o'clock (arrows on images in left [L] breast) was depicted at a short follow-up examination 6 months later because of nipple discharge. The lesion was interpreted as BI-RADS category 3 and the pathologic result was invasive ductal carcinoma. Six months after the diagnosis in her left breast, the microcalcifications in her right (R) breast were found to be ductal carcinoma in situ with a small invasive ductal carcinoma. CC = craniocaudal, MLO = mediolateral oblique.

model (21). Because DL algorithms often lack interpretability, combining DL with clinical data can shed light on the results obtained. First, by offering a careful cohort selection, we can avoid or adjust for biases. Second, by using clinically centered features, physicians may be able to transcend correlation-based predictions into causal networks of clinical factors leading to a diagnosis.

Our results regarding identification of highly probable healthy individuals by using only clinical information show the potential for personalized screening methods by training ML algorithms on readily available rich clinical data.

Previous studies have applied ML and DL of breast cancer to relatively small sets, typically less than 2500 individuals, on the basis of subsets of the Digital Database for Screening Mammography data set, INBreast data set, or Benchmarking Data sets for Breast Cancer (22–24). Some studies report results on full images (25–27), whereas others focus on region-of-interest patch classification (28–30). Recently, the digital mammography Dialogue for Reverse Engineering Assessments and Methods Dialogue for Reverse Engineering Assessments and Methods (known as DREAM) challenge (31) provided, to our knowledge, the largest existing DM data set confirmed with tissue diagnosis, consisting of 86 000 individuals. Their objective was to develop an automatic algorithm for breast cancer screening classification wherein only global information of biopsies positive for cancer was provided, first without clinical information and then with a limited set of features. The winning team obtained an AUC of 0.87, and specificity of 81% at a sensitivity of 80% (31). Another large study used 103 000 images from 23 000 examinations (32) and focused on breast cancer screening with BI-RADS categories 0, 1, or 2, corresponding to examinations that are incomplete, normal, or with benign findings.

Risk prediction models may further improve, by considering genetic information, hormone measurements, and breast density (33). Indeed, several studies have already shown significant improvement by adding breast density (34,35). Wu et al (36) used Gail features with and without mined mammographic features by employing a logistic regression-based model that resulted in an AUC of 0.71 versus 0.60, respectively. This, however, relied on the radiologist's analysis and interpretation of the index DM.

Our study had limitations. In our general cohort, we reported a lower number of relatives in breast cancer family history of women with biopsy positive for cancer. This could be explained by the exclusion of women with a personal history of breast cancer from the cohort. This exclusion only affected the number of women with family history that were found to have a biopsy positive for cancer, without affecting the number of women without a biopsy positive for cancer. We expect this counter-intuitive finding to be present in other studies excluding breast cancer survivors. One approach for

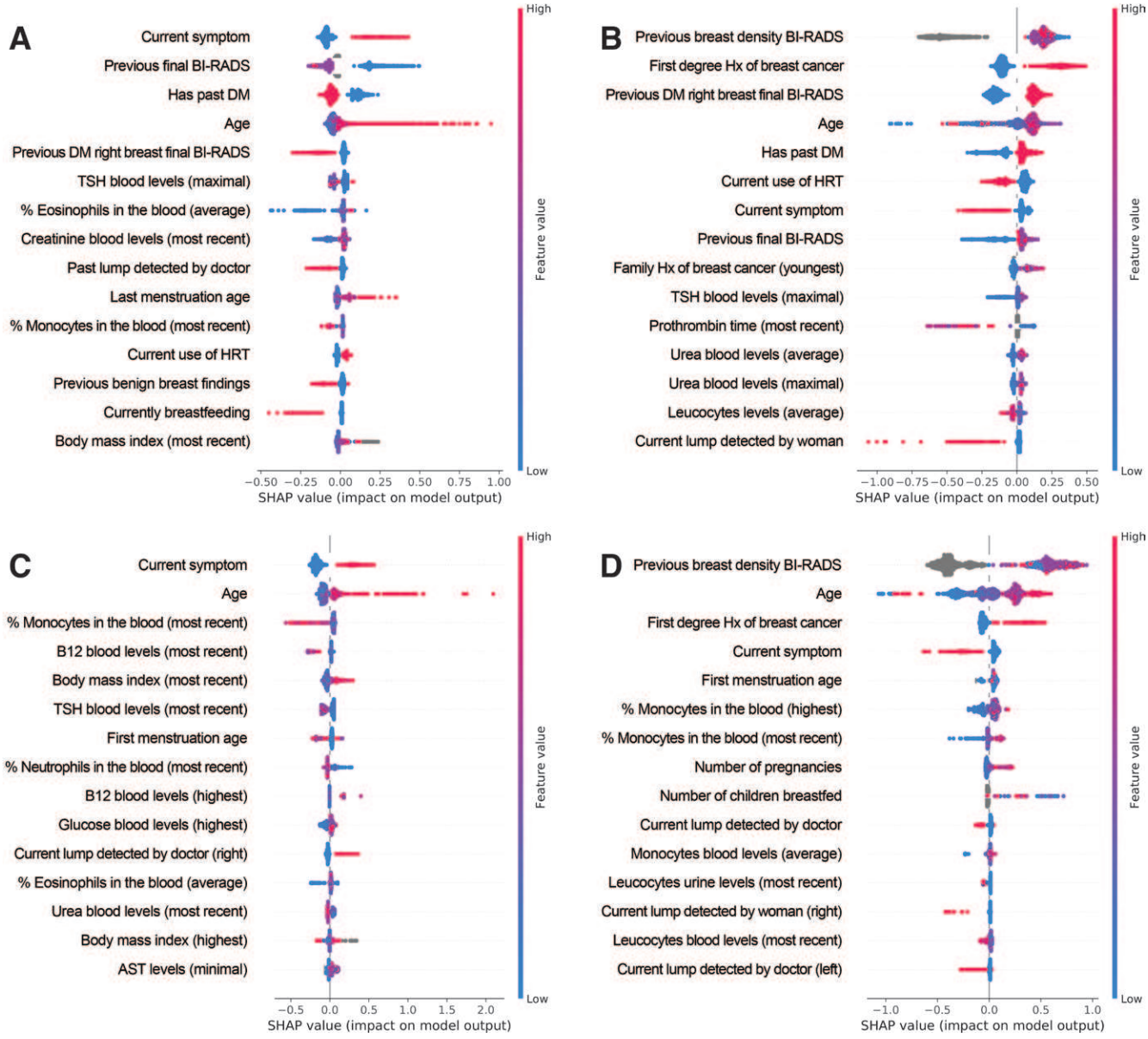

Figure 5: Clinical features contribution. The features are ordered on the y-axis in a descending order according to their mean absolute impact on prediction of biopsy positive for cancer. Each dot represents the Shapley additive explanations (SHAP) value for a specific feature and a specific woman. SHAP algorithm takes into account all possible combinations of features with and without that specific feature to evaluate its contribution to the prediction. The farther a dot is from 0 on the x-axis, the more effect (positive or negative) this feature had on the machine learning–deep learning model output for this particular woman. A dot's color indicates the feature's original value using a color bar between low (blue) and high (magenta) values; missing data are gray. The color scale was calculated for each feature separately on the basis of the women's feature values. Values that were higher than the 95th percentile and lower than the 5th percentile were trimmed. *A*, Top 15 highest contributing clinical features for prediction of biopsy positive for cancer as evaluated in the entire sample set. *B*, Top 15 highest contributing clinical features for normal examination identification in the entire sample set. *C*, Top 15 highest contributing clinical features for prediction of biopsy positive for cancer in the test set of the subcohort of first examinations only. *D*, Top 15 highest contributing clinical features for normal examination identification in the test set of the subcohort of first examinations only. AST = aspartate aminotransferase, BI-RADS = Breast Imaging Reporting and Data System, DM = digital mammography, HRT = hormone replacement therapy, Hx = history, TSH = thyroid-stimulating hormone.

correcting this selection bias is to limit the cohort to women undergoing their first mammographic examination, rendering this exclusion criterion moot. Consequently, we reported our results for first-examination subcohorts in addition to the general cohort, in which the selection bias was corrected, and we obtained improved results. The model was trained by using images from Assuta Medical Centers facilities, which use one mammography vendor (Hologic, Bedford, Mass); the clinical data originated exclusively from Maccabi Health Services facilities. Therefore, these results must be validated across different vendors, facilities, and populations around the world. Variability in the clinical data available in different facilities is expected, but the fact that we identified highest-contributing features for each prediction objective should

help reproduce these results in other facilities. Because of the process by which data were transferred from Maccabi Health Services, many women were excluded on the basis of a single nonmalignant DM examination without sufficient follow-up to determine that their results are without suspicious lesions that require further diagnostic workup. On the other hand, many women with benign findings were introduced into the cohort. We addressed this issue by sampling cohort members on the basis of their real-world distribution. The distinction between screening and diagnostic studies at Assuta Medical Centers was not well defined; we addressed this by analyzing only the standard views available at screening examinations. Finally, our ability to fully understand the ML-DL malignancy detection capabilities was complicated by two main factors. The ML-DL model does not yet offer a localization of the finding, only a global probability for the entire breast. We mitigated this by inferring the areas that most contributed to the global prediction by using the technique by Fong and Vedaldi (19). We currently do not have data to differentiate between different types of findings such as calcifications or mass that would help us to better analyze our results.

In conclusion, we developed a combined machine learning (ML) and deep learning (DL) model trained on a data set of linked mammograms and health records that improved previous risk models and obtained performance in the acceptable range of radiologists for breast cancer screening. The model did not perform better than radiologists, it performed differently. In a scenario where double reading at screening mammography is not available, as is the case at Assuta Medical Centers, we believe that the use of this model as a second reader could be beneficial. In general, the ML-DL model does not use the same tools as those accessible to radiologists. For example, it does not yet incorporate comparisons to previous mammography as is performed by human radiologists. Another example is the use of US images, which is a common part of the screening process at Assuta Medical Centers. We believe that, in the future, incorporating these data and additional clinical data such as genetic information can further improve the ML-DL model's performance. ML technology emphasizes the need for linking data sets from multiple modalities to improve the accuracy of breast cancer detection and save experts' valuable time on high-probability healthy individuals. In particular, this model's ability to lower false-negative results by half is of immediate clinical relevance.

Acknowledgments: We thank Tal El-Hay, PhD, and Chen Yanover, PhD, for helpful discussions throughout the work on this manuscript. We thank Natalie Shapira, MSc, and Sivan Ravid, MSc, for their contributions during project kickoff. We thank Ehud Klein, MD, from Maccabi Health Services and Yuliana Weinstein, MD, from Assuta Medical Centers for sharing their experience. We thank Chani Sacharen, BEng, for her help in editing the manuscript. We thank the IT departments at IBM, Maccabi, and Assuta for their dedicated support. Last, we thank the entire Healthcare Informatics group in IBM Research–Haifa for useful brainstorming and support.

Author contributions: Guarantors of integrity of entire study, A.A.B., M.C., E.H., M.G.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, A.A.B., M.C., A.S., R.M., S.N., G.K., Y.G., V.S., M.R.Z., M.G.; clinical studies, E.H., V.S., M.G.; statistical analysis, A.A.B., M.C., A.S., A.H., E.K., G.K., V.S., M.R.Z.; and manuscript editing, A.A.B., M.C., E.K., G.K., Y.G., V.S., M.R.Z., M.G.

Disclosures of Conflicts of Interest: **A.A.B.** disclosed no relevant relationships. **M.C.** disclosed no relevant relationships. **Y.S.** disclosed no relevant relationships. **A.S.** disclosed no relevant relationships. **A.H.** disclosed no relevant relationships. **R.M.** disclosed no relevant relationships. **E.B.** disclosed no relevant relationships. **E.H.** disclosed no relevant relationships. **S.N.** disclosed no relevant relationships. **E.K.** disclosed no relevant relationships. **G.K.** disclosed no relevant relationships. **Y.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed employment from IBM. Other relationships: disclosed no relevant relationships. **V.S.** disclosed no relevant relationships. **M.R.Z.** disclosed no relevant relationships. **M.G.** disclosed no relevant relationships.

References

1. American Cancer Society. Global Cancer Facts & Figures. 3rd ed. Atlanta, Ga: American Cancer Society, 2015.
2. Giger ML, Karssemeijer N, Schnabel JA. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annu Rev Biomed Eng 2013;15(1):327–357.
3. Kopans DB. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. Cancer 2002;94(2):580–581; author reply 581–583.
4. Sickles EA, D'Orsi CJ, Bassett LW. American College of Radiology Breast Imaging Reporting and Data System Atlas (ACR BI-RADS Atlas). Reston, Va: American College of Radiology, 2013.
5. Katalinic A, Bartel C, Raspe H, Schreer I. Beyond mammography screening: quality assurance in breast cancer diagnosis (The QuaMaDi Project). Br J Cancer 2007;96(1):157–161.
6. Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology 2017;283(1):49–58.
7. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat 2012;132(2):365–377.
8. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402–2410.
9. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–118 [Published correction appears in Nature 2017;546(7660):686.] https://doi.org/10.1038/nature21056.
10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318(22):2199–2210.
11. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225 [cs, stat]. http://arxiv.org/abs/1711.05225. Published November 2017. Accessed September 10, 2018.
12. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
13. Lehman CD, Yala A, Schuster T, et al. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. Radiology 2019;290(1):52–58.
14. Gilbert FJ, Astley SM, Gillan MG, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med 2008;359(16):1675–1684.
15. van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 2011;261(3):719–732.
16. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology 2019;290(2):305–314.
17. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. 2016; 785–794.
18. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Vol 4. 2017; 12.
19. Fong R, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. 2017 IEEE International Conference on Computer Vision (ICCV). October 2017; 3449–3457.
20. Zadrozny B. Learning and Evaluating Classifiers under Sample Selection Bias. ICML, 2004.
21. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81(24):1879–1886.
22. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. LOCATION: Medical Physics Publishing, 2000; 212–218.
23. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. Acad Radiol 2012;19(2):236–248.

24. Moura DC, López MAG, Cunha P, et al. Benchmarking Datasets for Breast Cancer Computer-Aided Diagnosis (CADx). In: Ruiz-Shulcloper J, Sanniti di Baja G, eds. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Vol 8258. Berlin, Germany: Springer, 2013; 326–333.
25. Carneiro G, Nascimento J, Bradley AP. Unregistered multiview mammogram analysis with pre-trained deep learning models. In: Navab N, Hornegger J, Wells W, Frangi A, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Cham, Switzerland: Springer, 2015; 652–660.
26. Dhungel N, Carneiro G, Bradley AP. Fully automated classification of mammograms using deep residual neural networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, 2017; 310–314.
27. Zhu W, Lou Q, Vang YS, Xie X. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins D, Duchesne S, eds. Medical Image Computing and Computer Assisted Intervention − MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science, vol 10435. Cham, Switzerland: Springer, 2017; 603–611.
28. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–312.
29. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Programs Biomed 2016;127:248–257.
30. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. Invest Radiol 2017;52(7):434–440.
31. D.R.E.A.M. The digital mammography DREAM challenge. https://www.synapse.org/Digital_Mammography_DREAM_challenge. Published 2017. Accessed DATE.
32. Geras KJ, Wolfson S, Shen Y, Kim S, Moy L, Cho K. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047. http://arxiv.org/abs/1703.07047. Published 2017. Accessed DATE.
33. Howell A, Anderson AS, Clarke RB, et al. Risk determination and prevention of breast cancer. Breast Cancer Res 2014;16(5):446.
34. Tice JA, Cummings SR, Ziv E, Kerlikowske K. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. Breast Cancer Res Treat 2005;94(2):115–122.
35. Chen J, Pee D, Ayyagari R, et al. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. J Natl Cancer Inst 2006;98(17):1215–1226.
36. Wu Y, Abbey CK, Chen X, et al. Developing a utility decision framework to evaluate predictive models in breast cancer risk estimation. J Med Imaging (Bellingham) 2015;2(4):041005.