Combining Deep Learning and Handcrafted Radiomics for Classification of Suspicious Lesions on Contrast-enhanced Mammograms

Combining Deep Learning and Handcrafted Radiomics for Classification of Suspicious Lesions on Contrast-enhanced Mammograms

Manon P. L. Beuque, Dipl Ing • Marc B. I. Lobbes, MD, PhD • Yvonka van Wijk, PhD • Yousif Widaatalla, MD, MSc • Sergey Primakov, MSc • Michael Majer, MD • Corinne Balleyguier, MD • Henry C. Woodruff, PhD • Philippe Lambin, MD, PhD**

From the Department of Precision Medicine (M.P.L.B., Y.v.W., Y.W., S.P., H.C.W., P.L.) and Department of Radiology and Nuclear Medicine (M.B.I.L.), GROW School for Oncology and Reproduction, Maastricht University, Universiteitssingel 40, 6229 ER Maastricht, the Netherlands; Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, Maastricht, the Netherlands (M.B.I.L., H.C.W., P.L.); Department of Medical Imaging, Zuyderland Medical Center, Sittard-Geleen, the Netherlands (M.B.I.L.); Department of Imaging, Institut Gustave Roussy, Université Paris Saclay, Villejuif, France (M.M., C.B.); and Biomaps, UMR1281 INSERM, CEA, CNRS, Université Paris-Saclay, Villejuif, France (C.B.). Received July 25, 2022; revision requested September 8; revision received March 17, 2023; accepted April 13. **Address correspondence to** P.L. (email: *philippe.lambin@maastrichtuniversity.nl*).

The authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 no. 694812 – Hypoximmuno) and ERC-2020-PoC: 957565-AUTO.DISTINCT. The authors also acknowledge financial support from EUROSTARS (DEEP-MAM: ESTAR19103), the European Union's Horizon research and innovation programme under grant agreement MSCA-ITN-PREDICT no. 766276, CHAIMELEON no. 952172, EuCanImage no. 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY no. UM 2017-8295), IMI-OPTIMA no. 101034347, AIDAVA (HORIZON-HLTH-2021-TOOL-06) no.101057062, and EUCAIM (DIGITAL-2022-CLOUD-AI-02) no.101100633.

* H.C.W. and P.L. are co–senior authors.

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

See also the editorial by Bahl and Do in this issue.

Radiology 2023; 307(5):e221843 • https://doi.org/10.1148/radiol.221843 • Content codes: BR AI

Background: Handcrafted radiomics and deep learning (DL) models individually achieve good performance in lesion classification (benign vs malignant) on contrast-enhanced mammography (CEM) images.

Purpose: To develop a comprehensive machine learning tool able to fully automatically identify, segment, and classify breast lesions on the basis of CEM images in recall patients.

Materials and Methods: CEM images and clinical data were retrospectively collected between 2013 and 2018 for 1601 recall patients at Maastricht UMC+ and 283 patients at Gustave Roussy Institute for external validation. Lesions with a known status (malignant or benign) were delineated by a research assistant overseen by an expert breast radiologist. Preprocessed low-energy and recombined images were used to train a DL model for automatic lesion identification, segmentation, and classification. A handcrafted radiomics model was also trained to classify both human- and DL-segmented lesions. Sensitivity for identification and the area under the receiver operating characteristic curve (AUC) for classification were compared between individual and combined models at the image and patient levels.

Results: After the exclusion of patients without suspicious lesions, the total number of patients included in the training, test, and validation data sets were 850 (mean age, 63 years ± 8 [SD]), 212 (62 years ± 8), and 279 (55 years ± 12), respectively. In the external data set, lesion identification sensitivity was 90% and 99% at the image and patient level, respectively, and the mean Dice coefficient was 0.71 and 0.80 at the image and patient level, respectively. Using manual segmentations, the combined DL and handcrafted radiomics classification model achieved the highest AUC (0.88 [95% CI: 0.86, 0.91]) ($P < .05$ except compared with DL, handcrafted radiomics, and clinical features model, where $P = .90$). Using DL-generated segmentations, the combined DL and handcrafted radiomics model showed the highest AUC (0.95 [95% CI: 0.94, 0.96]) ($P < .05$).

Conclusion: The DL model accurately identified and delineated suspicious lesions on CEM images, and the combined output of the DL and handcrafted radiomics models achieved good diagnostic performance.

Supplemental material is available for this article.

Full-field digital mammography (FFDM) continues to be the primary imaging tool for the detection of breast cancer. However, the diagnostic accuracy of FFDM is decreased in breasts with dense fibroglandular tissue (1), and the specificity of FFDM for detecting cancer is moderate (2). Hence, there remains a clinical need to increase the diagnostic accuracy of FFDM by using either supplemental imaging modalities, such as US or breast MRI, or technically advanced mammography, such as digital breast tomosynthesis or contrast-enhanced mammography (CEM).

Compared with FFDM, CEM has better diagnostic performance in terms of both sensitivity and specificity. Although CEM has a high sensitivity for identifying breast cancer, specificity remains moderate (3). In addition, the currently described diagnostic performance of CEM is based on studies (3) using visual assessment of the images by radiologists without the aid of computerized techniques.

Studies suggest that the diagnostic accuracy of FFDM might be improved with the help of machine learning (ML)–based image analysis. McKinney et al (4) showed that for some FFDM examinations, expert radiologists were unable to provide a correct diagnosis, whereas the ML model did. However, the ML model

Abbreviations

AUC = area under the receiver operating characteristic curve, CEM = contrast-enhanced mammography, DL = deep learning, FFDM = full-field digital mammography, ML = machine learning

Summary

A deep learning algorithm was able to accurately identify and delineate suspicious lesions on contrast-enhanced mammograms, and the combined outputs of this tool and a handcrafted radiomics model achieved good diagnostic performance.

Key Results

- In this retrospective study of 1601 patients, contrast-enhanced mammograms that showed suspicious lesions in patients who were recalled were used to train (n = 850) and test (n = 212) a deep learning (DL) model, which identified 99% of lesions on an external data set (n = 279).
- For DL model segmentations, lesion classification (malignant vs benign) using the DL model achieved the highest sensitivity (90% [319 of 353 lesions]), while the combination of DL and handcrafted radiomics achieved the highest area under the receiver operating characteristic curve (0.95).

would sometimes be unable to recognize “obvious” cases (ie, those easily detected by expert radiologists). Many studies using ML on FFDM images have already been performed, for example, using handcrafted radiomics models (5,6) to classify breast lesions (7) and deep learning (DL) to identify and segment lesions (8,9), but the combination of these two methods in breast cancer imaging has rarely been reported. Briefly, handcrafted radiomics refers to the extraction of predefined quantitative features from regions of interest within the images, which are subsequently used as input into ML algorithms to perform regression or classification tasks, while DL uses the entire input image to automatically construct internal representations of the phenotype to perform detection, localization, and classification predictions.

In this study, we aimed to develop a comprehensive ML tool able to fully automatically identify, segment, and classify breast lesions on the basis of CEM images in recall patients. In our approach, a DL model was first trained to identify and segment suspicious lesions on CEM images and classify them as benign or malignant. Furthermore, handcrafted radiomics classification models based on both manually and automatically delineated regions of interest and clinical parameters were trained, evaluated, and combined with the DL predictions.

Materials and Methods

Study Sample

In this retrospective study, images and clinical data were collected in 1601 consecutive patients who underwent CEM mostly for recall assessment of breast lesions following screening at Maastricht UMC+ between 2013 and 2018. Patients with inconclusive findings at FFDM and/or US, suspicious (palpable) findings during physical examination, or who underwent mammography as an alternative to breast MRI when MRI was contraindicated were also included. The requirement for informed consent was waived by the institutional review board (approval no. METC 2019–0995). Data were collected using the picture archiving and communication system and anonymized. Patients were excluded if their examination was deemed negative (ie, no suspicious lesion was found) by an expert radiologist (M.B.I.L., with 13 years of experience in CEM) (Fig 1).

Images and clinical data were collected as an external validation data set from Gustave Roussy Institute between 2015 and 2019, for which informed consent was waived (approval no. 2022–140). The data from both institutes have not been reported in any prior publications.

Imaging

The acquisition protocol for CEM images was as described previously (10,11). In short, an iodinated contrast agent (iopromide [Ultravist 300, Bayer Healthcare] at Maastricht UMC+ and mostly iobitridol [Xenetix 350, Guerbet] at Gustave Roussy) was intravenously administered 2 minutes before the acquisition of dual-energy mammograms (Senographe Essential with

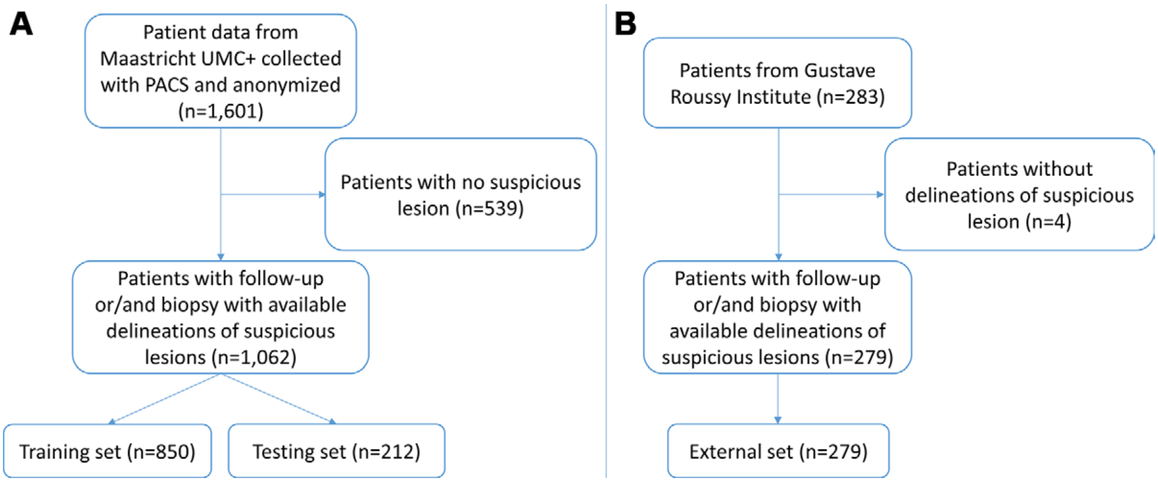
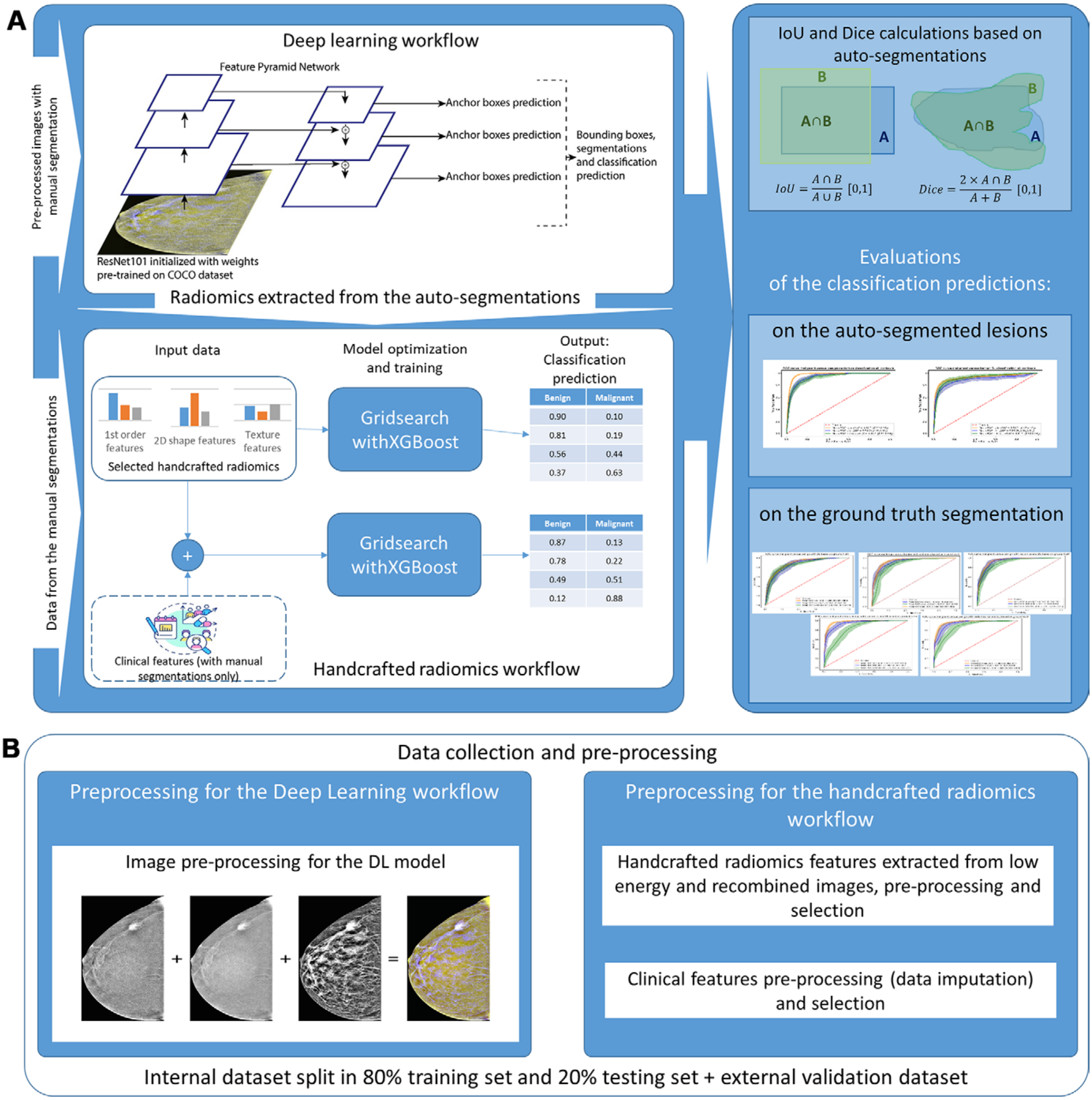

A
Patient data from Maastricht UMC+ collected with PACS and anonymized (n=1,601)
Patients with no suspicious lesion (n=539)
Patients with follow-up or/and biopsy with available delineations of suspicious lesions (n=1,062)
Training set (n=850)
Testing set (n=212)
B
Patients from Gustave Roussy Institute (n=283)
Patients without delineations of suspicious lesion (n=4)
Patients with follow-up or/and biopsy with available delineations of suspicious lesions (n=279)
External set (n=279)

Figure 1: Flowcharts describe patient inclusion for **(A)** the training and test data sets and **(B)** the external validation data set of the machine learning models. PACS = picture archiving and communication system.

Senobright CEM upgrade or Senographe Pristina, GE Healthcare) of both breasts in the mediolateral oblique and craniocaudal views. This resulted in a low-energy image equivalent to FFDM (10) and a recombined image in which areas of contrast material accumulation could be assessed (11), both of which were used for analysis. Lesions on all images were delineated using Medical Image Merge software version 4.1 (MIM Software) by a research assistant (Y.W.) supervised by a certified breast radiologist (M.B.I.L.) aided by information retrieved from the patient records and radiology reports. The final ground truth (diagnosis) of each delineation was assigned based on results obtained after review of the pathology reports and/or 2-year follow-up reports. In our study, the term *breast lesion* comprises architectural distortions, asymmetries, masses, and clusters of suspicious calcifications.


A
Pre-processed images with manual segmentation
Deep learning workflow
Feature Pyramid Network
Anchor boxes prediction
Anchor boxes prediction
Anchor boxes prediction
Bounding boxes, segmentations and classification prediction
ResNet101 initialized with weights pre-trained on COCO dataset
Radiomics extracted from the auto-segmentations
Data from the manual segmentations
Input data
1st order features
2D shape features
Texture features
Selected handcrafted radiomics
Model optimization and training
Gridsearch withXGBoost
Output: Classification prediction
+
Gridsearch withXGBoost
Clinical features (with manual segmentations only)
Handcrafted radiomics workflow
IoU and Dice calculations based on auto-segmentations
B
A∩B
A
B
A∩B
A
Evaluations of the classification predictions:
on the auto-segmented lesions
on the ground truth segmentation
B
Data collection and pre-processing
Preprocessing for the Deep Learning workflow
Image pre-processing for the DL model
+
+
=
Preprocessing for the handcrafted radiomics workflow
Internal dataset split in 80% training set and 20% testing set + external validation dataset

$$IoU = \frac{A \cap B}{A \cup B} \; [0,1] \qquad Dice = \frac{2 \times A \cap B}{A + B} \; [0,1]$$

Benign	Malignant
0.90	0.10
0.81	0.19
0.56	0.44
0.37	0.63

Benign	Malignant
0.87	0.13
0.78	0.22
0.49	0.51
0.12	0.88

Handcrafted radiomics features extracted from low energy and recombined images, pre-processing and selection

Clinical features pre-processing (data imputation) and selection

Figure 2: Workflow of the identification, delineation, and classification of suspicious lesions as malignant or benign with use of handcrafted radiomics models and deep learning (DL). **(A)** The DL workflow uses preprocessed images for input (see **B**) and a mask region-based convolutional neural network to predict bounding boxes, segmentations, and classification of lesions as malignant or benign; the handcrafted radiomics workflow uses as input preselected handcrafted radiomics (see **B**) with clinical features when making predictions with use of manual annotations, and an XGBoost model is trained on this data to predict benign versus malignant, using a grid-search function to fine-tune the parameters of this model; those models are then evaluated on the external data set in combination and individually. COCO = common objects in context, IoU = intersection over union, 2D = two-dimensional. **(B)** Data collection and preprocessing: The low-energy and recombined images are preprocessed and combined to use as input for the DL workflow; for the handcrafted radiomics workflow, handcrafted radiomics features are extracted from low-energy and recombined images, preprocessed, and preselected, and the clinical features are preprocessed using data imputation and preselected.

Table 1: Patient Characteristics

Clinical Characteristic	Training Data Set (*n* = 850)	Test Data Set (*n* = 212)	External Data Set (*n* = 279)	*P* Value for Training vs Test Data Sets	*P* Value for Test vs External Data Sets	*P* Value for Training vs External Data Sets
No. of lesions	850	212	319			
Age (y)*	63 ± 8	62 ± 8	55 ± 12	.87	.08	.10
Menopause status				.99	.88	.92
Premenopause	96 (11)	18 (8.5)	93 (33)			
Perimenopause	65 (7.6)	14 (6.6)	25 (9.0)			
Postmenopause	504 (59)	136 (64)	156 (56)			
Not reported	185 (22)	44 (21)	5 (1.8)			
No. of pregnancies*	1.9 ± 1.3	2.0 ± 1.2	2.2 ± 1.7	.86	<.001	<.001
No. of children*	1.7 ± 1.0	1.8 ± 1.1	1.9 ± 1.5	.95	<.001	<.001
Medication				.99	.72	.84
None	426 (50)	93 (44)	215 (77)			
Oral contraceptive pill	226 (27)	70 (33)	37 (13)			
Hormone replacement therapy	17 (2.0)	5 (2.3)	17 (6.1)			
Not reported	181 (21)	44 (21)	10 (3.6)			
Family history of breast cancer				>.99	.72	.72
Positive	123 (14)	31 (14)	94 (34)			
Negative	551 (65)	139 (66)	170 (61)			
Not reported	176 (21)	42 (20)	15 (5.4)			
Personal history of breast cancer				.92	.87	.91
Positive	10 (1.2)	1 (0.5)	10 (3.6)			
Negative	666 (78)	169 (80)	259 (93)			
Not reported	174 (21)	42 (20)	10 (3.6)			
Cup size						
A–C	418 (49)	104 (49)	164 (59)	.98	.08	.01
D–F	241 (28)	62 (29)	72 (26)	.80	.38	.40
Larger than F	10 (1)	3 (1.4)	3 (1.1)	.73	>.99	>.99
Not reported	181 (21)	43 (20)	40 (14)	.75	.08	.01
Disease characteristics per lesion						
No special type	227 (27)	58 (27)	163 (51)	.85	<.001	<.001
Ductal carcinoma in situ	63 (7.4)	19 (9.0)	9 (2.8)	.48	.01	.01
Other carcinoma	69 (8.1)	10 (4.7)	23 (7.2)	.09	.25	.60
Cyst	310 (36)	80 (38)	75 (23)	.73	<.001	<.001
Fibroadenoma	68 (8.0)	17 (8.0)	29 (9.1)	.99	.68	.56
Negative	6 (0.7)	1 (0.5)	8 (2.5)	>.99	>.99	.03
Not reported	107 (13)	27 (13)	12 (3.8)	.95	<.001	<.001

Note.—Unless otherwise specified, data are numbers of patients, with percentages in parentheses. For continuous variables, the Mann-Whitney *U* test was used for two independent samples. For categorical variables, if every category had fewer than 10 samples, the Fisher exact test was used; otherwise, a two-proportion *z* test was used. *P* < .05 was considered to indicate statistically significant difference.

* Data are means ± SDs.

Automatic Identification and Delineation of Suspicious Lesions with Use of Mask Region-based Convolutional Neural Network

The 1062 patients were split randomly into training and test data sets at a ratio of 4:1. This split resulted in 850 patients for the training data set and 212 patients for the test data set.

Low-energy and recombined CEM images were first preprocessed to filter out noise or irrelevant details (eg, removal of foreign objects and background), and image size was also reduced to limit computational costs (12). Because the preprocessing of CEM images for DL is not standardized, we used a series of preprocessing steps, including contrast adjustment, intensity normalization, and merging low-energy and recombined images into one image. Details of this process are reported in the first section of Appendix S1 and in Figure S1.

The DL model was trained on the preprocessed CEM images to identify (ie, generate a bounding box around the lesion of interest), delineate, and classify lesions as either benign or malignant using Mask R-CNN with a ResNet101 feature pyramid network backbone (13). The details of the DL model and associated metrics can be found in the second section of Appendix S1. The code used during this study is accessible on GitHub: *https://github.com/precision-medicine-um/Radiomics_for_CEM/*.

After preprocessing of the image data, the Mask R-CNN model was trained on 1810 images from 850 patients in the training data set, tested on 454 images from 212 patients, and validated on 590 images from 279 patients in the external data set.

The DL model was trained for 30 epochs, and the best weights were obtained from epoch 13, at which point the model had the lowest total loss on the test data set.

Table 2: Identification and Segmentation Results of the Deep Learning Model

Level and Parameter	Test Data Set	External Data Set	*P* Value
Delineations per lesion			
Accuracy (%)	64 (279/436)	73 (431/590)	.002
Sensitivity (%)	85 (371/436)	90 (532/590)	.01
Mean Dice coefficient	0.65	0.71	<.001
Delineations per patient			
Accuracy (%)	80 (170/212)	88 (245/279)	.02
Sensitivity (%)	94 (200/212)	99 (275/279)	.009
Mean Dice coefficient	0.75	0.80	.007

Note.—Data in parentheses are numbers of lesions or patients, as specified. *P* values were calculated using the *z* test for the accuracy and sensitivity and the Mann-Whitney *U* test for mean Dice coefficient.

Development and Combination of Models for Lesion Classification

Handcrafted radiomics features were selected from both the manual and DL-generated segmentations. After feature selection, a subset was used to train an ML model, which returned a probability of malignancy. To test if the clinical features had any added predictive value, the same feature selection method was applied on the radiomics and clinical features, and the model was retrained. The clinical features included were number of pregnancies, number of children, family history of breast cancer, personal history of breast cancer, age, menopause status, medication use (ie, hormone replacement therapy or oral contraceptives), and cup size. Moreover, for every bounding box generated by the DL model, it also produced a benign or malignant (ie, 0 or 1) classification and a confidence score, which, when added, yielded a classification probability. To understand the importance given to the features selected by the different models, Shapley additive explanations, or SHAP, values were calculated for the training data set. For details about this process, see the third section of Appendix S1.

To combine the DL and handcrafted radiomics models, we averaged their classification probabilities to arrive at a single classification prediction. We repeated this process for the combined radiomics and clinical features model as well as for the combined DL, radiomics, and clinical features model.

Statistical Analyses

The statistical analyses were conducted by M.P.L.B. using Python version 3.7 (Table S1). We reported patients' characteristics per data set and their differences. For the continuous variables, we used the Mann-Whitney *U* test for two independent samples. For categorical variables, if every category had fewer than 10 samples, we used the Fisher exact test; otherwise, we used a two-proportion *z* test. We considered $P < .05$ to indicate statistically significant difference.

Identification sensitivity, accuracy, and mean Dice coefficient of the segmentations were reported per image and per patient on the test and external data sets. The Dice coefficient was computed per lesion and reported in a violin plot (Fig S2). We used the same tests as stated in the previous paragraph.

In a post hoc subanalysis of examinations where the DL model did not identify the presence of a lesion (ie, false-negative findings), we calculated proportion *z* tests ($\alpha = .05$) on the false-negative results and reviewed the images with the same breast radiologist as before to establish potential causes for these false-negative findings.

Lesion classification performance measures—including area under the receiver operating characteristic curve (AUC), sensitivity, specificity, accuracy, and F1 score—were computed for human- and DL-generated delineations in the external data set at both the lesion and patient levels. The calibration curves obtained with the DL and handcrafted radiomics methods on the test data set are provided in Figure S3. The method used to obtain the predictions per patient is described in the last section of Appendix S1. The thresholds used to obtain binary predictions for benign versus malignant lesions were selected based on the statistics obtained in the training data set with the Youden index (14). We listed the results obtained when the binary predictions of the two best-performing models were in agreement, and we reported the percentage of cases for which the models agreed. The 95% CIs were computed for AUCs, specificities, and sensitivities with use of bootstrapping, which resampled the data sets 2000 times, and Tukey tests were performed between the different metrics to assess significant differences for $\alpha = .05$. The complete workflow is presented in Figure 2.

Results

Patient Characteristics

We excluded 543 patients without suspicious lesions (Fig 1). The total number of patients included in the training, test, and validation data sets were 850 (mean age, 63 years ± 8

[SD]), 212 (62 years ± 8), and 279 (55 years ± 12), respectively. The clinical characteristics of the patients are described in Table 1. The training and test data sets were nonsignificantly different for all patient characteristics. The majority of the patient characteristics were similar in the three data sets, but the distribution of disease characteristics per lesion in the external data set was significantly different across most categories.

Identification and Delineation of the Lesions

The results presented in Table 2 show that the model performed better on the external data set than on the test data set. At the patient level, 99% of the lesions were found in the external data set, and 94% in the test data set (*P* = .009). The distribution of the Dice scores can be seen in Figure S2.

Analysis of False-Negative Findings

The DL model failed to identify lesions in four of 279 patients (1.4%) in the external validation data set, in 12 of 212 (5.7%) patients in the test data set, and in 43 of 850 (5.1%) patients in the training data set. The proportion of false-negative findings was different in the external data set compared with the training (*P* = .009) and test (*P* = .009) data sets. The expert radiologist's interpretation of lesions not detected by the DL model are displayed in Figure S4. An example of a lesion seen on a recombined image is available in Figure S5.

Of the false-negative lesions, seven of 12 (58%) in the test data set and 28 of 43 (65%) in the training data set were calcifications. Figures 3 and 4 show example images of unidentified and identified calcifications, respectively, with use of the DL model.

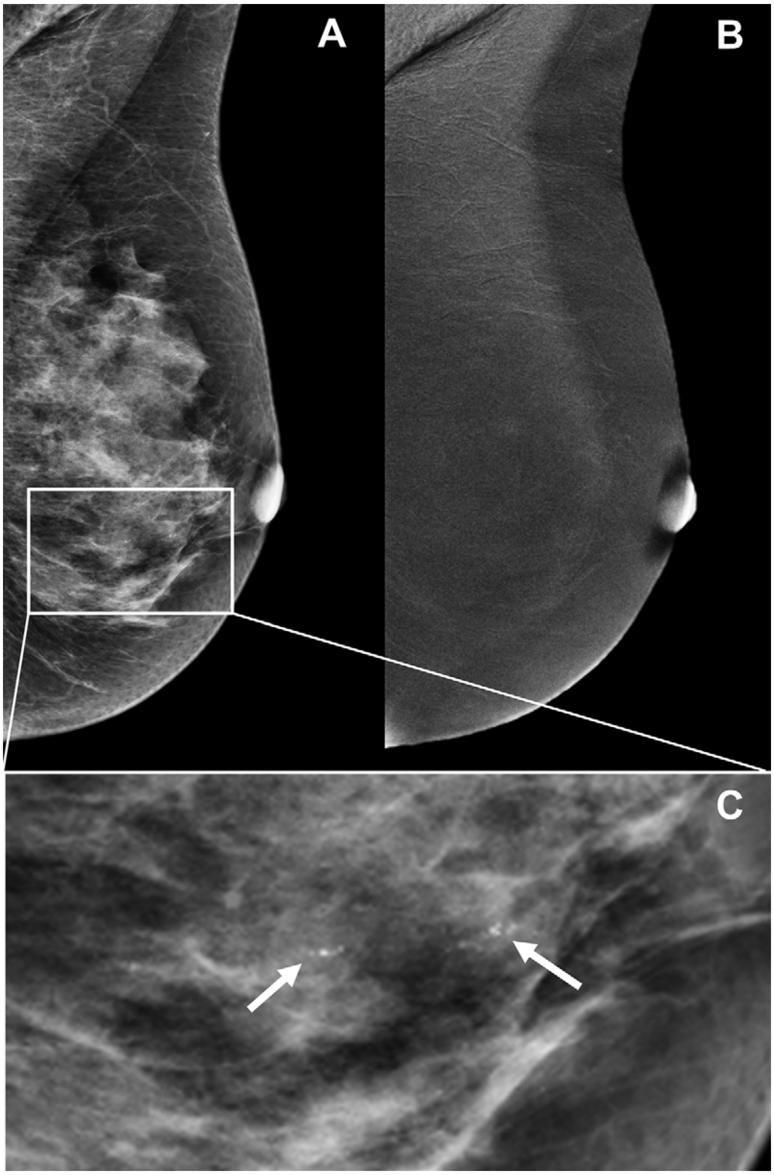


Figure 3: Example contrast-enhanced mammograms of a false-negative finding of suspicious calcifications by the deep learning (DL) model. On the **(A)** mediolateral oblique low-energy image in the left breast of a 58-year-old woman, a cluster of fine linear and branching calcifications was observed by a radiologist, but a negative finding was reported by the DL model. **(B)** On the recombined image, no enhancement was observed. The model did not provide any delineation. However, stereotactic vacuum-assisted core-needle biopsy showed ductal carcinoma in situ. **(C)** Magnification of **A** shows cluster of fine linear and branching calcifications (arrows).

Classification of Lesions as Benign or Malignant

The optimal parameters found with grid-search for the handcrafted radiomics models are reported in Table S2, and the feature importance is available in Figure S6, which includes the summary plot of the SHAP values. The receiver operating characteristic curves are available in Figure 5 for the manual segmentations and Figure 6 for the automatic segmentations. For the predicted classification obtained on the manual segmentations, we observed, based on the intersections of the CIs, that all models tested appeared to overfit on the training data set. However, the CIs of the receiver operating characteristic curves obtained with the predictions on the test data set always overlapped with the CIs based on the predictions obtained in the training data set. For the automatically generated segmentation, the DL model did not overfit on the training data set, but the handcrafted radiomics model did.

The classification results using the external data set are provided in Table 3. For classification per lesion based on the manual segmentations, the combination of the handcrafted radiomics and DL model classifications yielded the highest AUC (0.88 [95% CI: 0.86, 0.91]) and sensitivity (83% [95% CI: 79, 87]), and the handcrafted-radiomics model yielded the highest specificity (80% [95% CI: 75, 85]). For classification per patient based on the manual segmentations, the highest AUC (0.88 [95% CI: 0.84, 0.93]) and sensitivity (89% [95% CI: 85, 93]) were found with the DL model, and the highest specificity (83% [95% CI: 75, 90]) was found using the handcrafted radiomics model. For the automatically generated segmentations at the lesion level, the combined handcrafted radiomics and DL model obtained the highest AUC (0.95 [95% CI: 0.94, 0.96]) and specificity (86% [95% CI:

84, 87]), while the DL model alone yielded the highest sensitivity (90% [95% CI: 87, 93]). For classification per patient based on the automatically generated segmentations, the handcrafted radiomics model yielded the highest specificity (74% [95% CI: 64, 84]), while the highest sensitivity was obtained with the DL model alone (100% [95% CI: 100, 100]). The combination DL and handcrafted radiomics model achieved the highest AUC (0.91 [95% CI: 0.86, 0.95]). Accuracy and F1 score are presented in Table S3. The calibration curves are shown in Figure S3.

The handcrafted radiomics and DL model predictions based on manual segmentations agreed for 76% of the lesions, and the AUC, specificity, and sensitivity within that subset were 0.95, 80%, and 97%, respectively. For the per-patient predictions, the handcrafted radiomics and DL models agreed 92% of the time, and the AUC, specificity, and sensitivity were 0.93, 78%, and 96%, respectively, on that subset. For the automated segmentations, the models agreed on 84% of the lesions and all of the patients. The AUC, specificity, and sensitivity were 0.96, 89%, and 95% per lesion and 0.91, 59%, and 98% per patient, respectively. Results achieved with other combinations are reported in Table 3.

Discussion

We saw in the literature that handcrafted radiomics and deep learning (DL) models individually achieve good performance in lesion classification (benign or malignant) at full-field digital mammography (7–9). In this study, we aimed to build and validate a workflow that would find suspicious lesions on contrast-enhanced mammograms and give a classification of benign or malignant according to handcrafted radiomics and DL models. Additionally, we assessed the added value of clinical features and handcrafted radiomics to classify the manually delineated and automatically delineated lesions. Our DL model found 532 of 590 lesions (90%) on the external validation data set while correctly identifying 275 of 279 patients with lesions (99%). For the classification of lesions, and for most performance evaluation measures, the combined handcrafted radiomics and DL model performed best on the manual delineations (area under the receiver operating characteristic curve [AUC], 0.88) ($P < .05$ compared with the other methods except for DL combined with radiomics and clinical features, where $P = .90$), as well as on the DL model–generated segmentations (AUC, 0.95) ($P < .05$ compared with the other methods). Hence, we concluded that our identification and classification model performed at a level that would make it potentially generalizable.

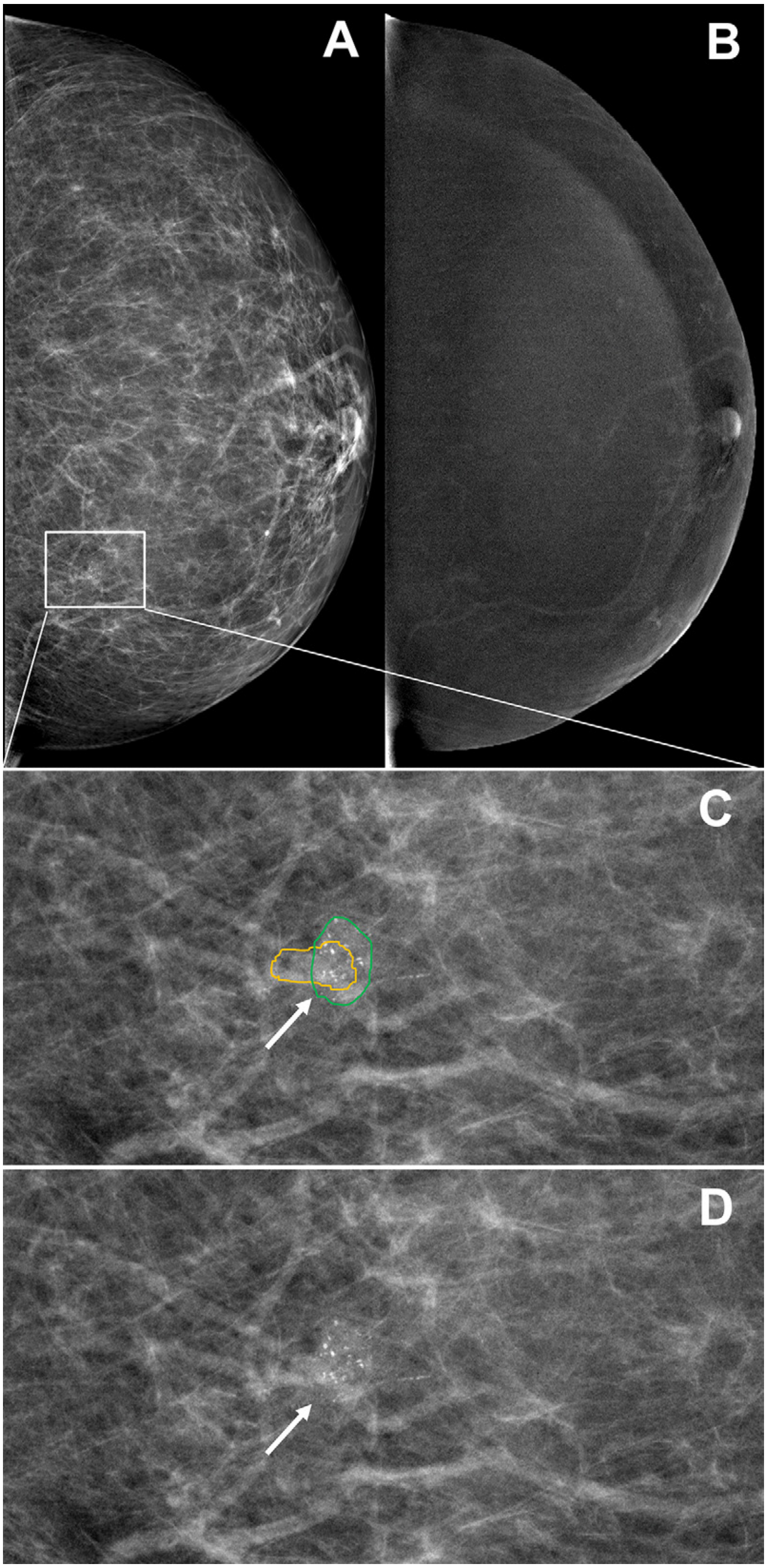


Figure 4: Example contrast-enhanced mammograms of a correct finding of suspicious calcifications by the deep learning (DL) model. **(A, C, D)** Low-energy images in the left breast of a 58-year-old woman show a small cluster of fine calcifications (outlines [green for ground truth, yellow for prediction] in **C**; arrows in **C** and **D**) detected by the DL model, with subtle nonmass enhancement at the site of the calcifications on the **(B)** recombined image. Subsequent stereotactic vacuum-assisted core-needle biopsy showed ductal carcinoma in situ.

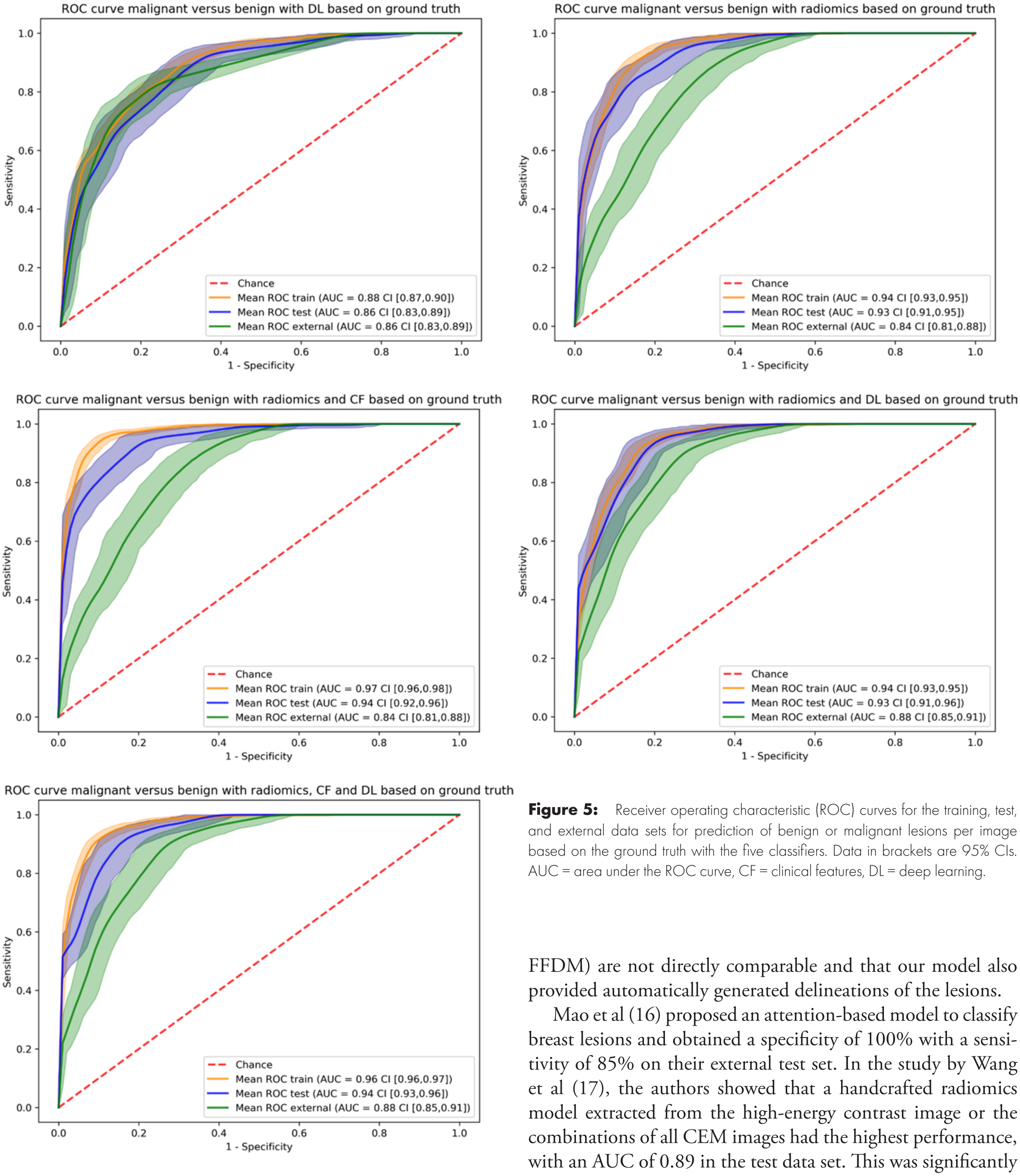


Figure 5: Receiver operating characteristic (ROC) curves for the training, test, and external data sets for prediction of benign or malignant lesions per image based on the ground truth with the five classifiers. Data in brackets are 95% CIs. AUC = area under the ROC curve, CF = clinical features, DL = deep learning.

This is, to our knowledge, the first study to provide a full workflow for identification, segmentation, and classification of suspicious lesions at CEM and to compare the results between handcrafted radiomics and DL models. A similar study was performed on FFDM with use of DL only, reporting a sensitivity of 90% and a false-positive rate of 30% for identification of malignant lesions (15), which is similar to our study. It is important to note that these imaging modalities (CEM and FFDM) are not directly comparable and that our model also provided automatically generated delineations of the lesions.

Mao et al (16) proposed an attention-based model to classify breast lesions and obtained a specificity of 100% with a sensitivity of 85% on their external test set. In the study by Wang et al (17), the authors showed that a handcrafted radiomics model extracted from the high-energy contrast image or the combinations of all CEM images had the highest performance, with an AUC of 0.89 in the test data set. This was significantly better than using the low-energy contrast (generally accepted to be roughly equivalent to FFDM [18]), which achieved an AUC of 0.87. Although these studies showed promising results for automatically classifying benign and malignant lesions at CEM with use of ML approaches, they were limited by a relatively small training set (n = 159) and lack of external validation (17), or the external set on which the results were reported was relatively small (n = 46) and no CIs were given (16), which does not allow the reader to conclude that their models would

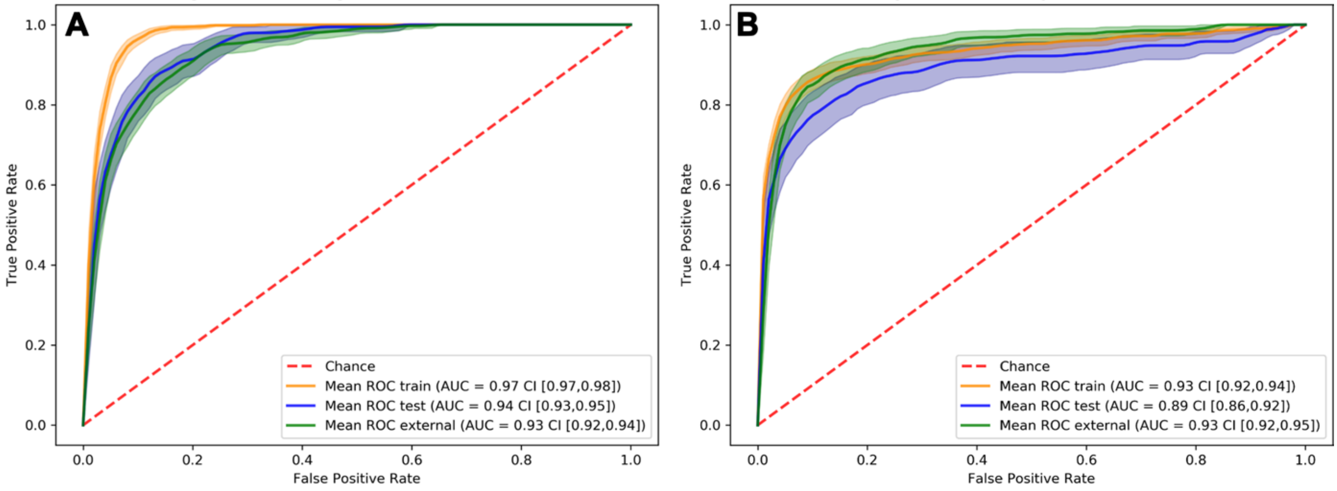


Figure 6: Receiver operating characteristic (ROC) curves for the training, test, and external data sets for prediction of other or malignant lesions per image based on the automatic segmentations with the two classifiers, the **(A)** radiomics-based model and **(B)** deep learning model. Data in brackets are 95% CIs. AUC = area under the ROC curve.

Table 3: Performance Measures of the Different Model Combinations on the Manual Segmentations and DL-generated Segmentations

	Per Lesion			Per Patient		
Model	AUC	Specificity (%)	Sensitivity (%)	AUC	Specificity (%)	Sensitivity (%)
	Manual Segmentations					
Radiomics	0.84 [0.81, 0.88]*	80 (182/227) [75, 85]†	67 (242/363) [62, 71]	0.83 [0.77, 0.88]*	83 (83/100) [75, 90]	63 (113/179) [56, 70]
DL	0.86 [0.83, 0.89]	75 (170/227) [69, 81]	83 (302/363) [79, 87]‡	0.88 [0.84, 0.93]	73 (73/100) [64, 81]	89 (160/179) [85, 93]
Radiomics + clinical features	0.84 [0.81, 0.88]*	76 (173/227) [71, 82]	74 (269/363) [69, 79]	0.83 [0.77, 0.88]*	79 (79/100) [71, 86]	66 (119/179) [60, 73]
DL + radiomics	0.88 [0.86, 0.91]§	78 (177/227) [72, 84]	83 (302/363) [79, 87]‡	0.88 [0.83, 0.92]§	77 (77/100) [69, 85]§	80 (144/179) [74, 86]
DL + radiomics + clinical features	0.88 [0.86, 0.91]§	79 (179/227) [73, 84]	80 (289/363) [76, 84]	0.88 [0.83, 0.92]§	78 (78/100) [70, 86]§	78 (140/179) [72, 84]
Agreed labels	0.95 [0.92, 0.97]	80 (128/160) [74, 86]†	97 (279/288) [95, 99]	0.93 [0.89, 0.96]	78 (71/91) [69, 86]	96 (161/167) [93, 99]
	Automatically Generated Segmentations					
Radiomics	0.93 [0.92, 0.94]	82 (2027/2463) [81, 84]	89 (315/353) [86, 92]	0.89 [0.85, 0.94]	74 (55/74) [64, 84]	88 (150/171) [83, 92]
DL	0.93 [0.92, 0.95]	83 (2043/2463) [81, 84]	90 (319/353) [87, 93]	0.87 [0.82, 0.92]	45 (33/74) [34, 57]	100 (171/171) [100, 100]
DL + radiomics	0.95 [0.94, 0.96]	86 (2106/2463) [84, 87]	90 (317/353) [87, 93]	0.91 [0.86, 0.95]‖	59 (44/74) [48, 70]‖	98 (168/171) [96, 100]‖
Agreed labels	0.96 [0.95, 0.97]	89 (1828/2049) [88, 91]	95 (297/313) [92, 97]	0.91 [0.86, 0.95]‖	59 (44/74) [48, 70]‖	98 (168/171) [96, 100]‖

Note.—Data in parentheses are numbers of lesions or patients, and data in brackets are 95% CIs. The clinical feature used in the model was age. "Agreed labels" metrics are calculated for the cases in which the deep learning (DL) and handcrafted radiomics models agreed on the predicted label (benign or malignant). Unless specified otherwise, for each of the columns separately, every table entry is significantly different from the rest of the data presented in that column (*P* < .05). AUC = area under the receiver operating characteristics curve.

* The handcrafted radiomics model was not statistically significantly different from the radiomics + clinical features model for AUC.

† The handcrafted radiomics model was not statistically significantly different from the agreed labels.

‡ The DL model was not statistically significantly different from the DL + radiomics model.

§ The DL + radiomics model was not statistically significantly different from the DL + radiomics + clinical features model.

‖ The DL + radiomics model was not statistically significantly different from the agreed labels.

perform similarly on a data set acquired externally. Our study is notable for its large training data set, the validation of the model on an external data set, and the combination of handcrafted radiomics and DL.

Regarding the classification performance achieved by our best-performing model on manual segmentations, the AUC is comparable with those obtained by radiologists across multiple studies. In the meta-analysis by Suter et al (19), the pooled AUC for eight studies classifying suspicious lesions was 0.89, similar to the result obtained by our model based on the combination of DL and handcrafted radiomics (AUC, 0.88). In fact, for the cases in which our models are most certain (ie, for which the DL and handcrafted radiomics models agree), we report an AUC of 0.95 per lesion and 0.93 per patient for the manual segmentations and an AUC of 0.96 per lesion and 0.91 per patient for the automated segmentations.

Our study had several limitations. First, the models in this study were not optimized to identify calcifications, which do not always enhance at CEM. Most false-negative findings were a result of this limitation. In the literature, contradictory results regarding the benefit of CEM compared with FFDM for the classification of calcifications by radiologists have been reported. It is also possible that the resolution of the images was too low for our model to identify certain calcifications. A solution to this problem could be to combine the model for the identification and classification of lesions at CEM with a different model that would specifically target calcifications with FFDM (or the low-energy images) only. Second, we only evaluated the delineations of lesions for which there was either a biopsy or prolonged follow-up, which can be a potential source of bias, but it is theoretically possible that other benign lesions of nonclinical importance were present on the image, potentially making false-positive identifications actually true-positive findings. Moreover, we did not further analyze the external validation cases, as privacy legislation did not permit us to retrospectively assess the in-depth data of these patients. Third, the identification algorithm was not tested on images that did not depict lesions, as the CEM scans were acquired after suspicious lesions had already been identified during screening. Fourth, the segmentations and the evaluation of the models were made by one certified radiologist with 13 years of experience reviewing CEM images. Independent review by multiple breast radiologists would be preferable to limit bias. Fifth, the systems used in this study are not yet equipped with automated breast density measurement software, nor are other available tools validated for use in CEM. Hence, we were not able to present our results per breast density category. Sixth, we hypothesized that the images were independent and could be used independently to train a model. However, the data sets contained two different views for each patient and were likely correlated. To overcome this, a follow-up study could use our best model, train two models on the two different views, and combine the result with a voting algorithm. The final limitation was the use of the same data to choose the method to combine the DL and handcrafted radiomics models and then to evaluate the performance of that combination. This could have led to some degree of overestimation. As a future direction, to confirm our findings and support the utility of our model, a follow-up study in which FFDM is replaced by CEM might be interesting to conduct to compare the performance of those systems and to test our algorithm's capacity to detect lesions.

In conclusion, our deep learning algorithm was able to automatically delineate and identify the majority of suspicious lesions seen at contrast-enhanced mammography and showed good performance for finding malignant lesions.

Author contributions: Guarantors of integrity of entire study, **M.P.L.B., H.C.W., P.L.**; 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, **M.P.L.B., M.B.I.L., Y.v.W., M.M., H.C.W.**; clinical studies, **M.B.I.L., M.M., C.B., P.L.**; experimental studies, **M.P.L.B., M.B.I.L., Y.v.W., S.P., H.C.W.**; statistical analysis, **M.P.L.B., Y.v.W., H.C.W.**; and manuscript editing, **M.P.L.B., M.B.I.L., Y.W., M.M., C.B., H.C.W., P.L.**

Data sharing statement: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: **M.P.L.B.** No relevant relationships. **M.B.I.L.** Research grant from GE Healthcare; speaker's fees from GE Healthcare, Tromp Medical, Bayer, and Guerbet; medical advisory board member for Bayer, Hologic, and GE Healthcare. **Y.v.W.** No relevant relationships. **Y.W.** No relevant relationships. **S.P.** Grant from Predict Innovative Training Network. **M.M.** No relevant relationships. **C.B.** No relevant relationships. **H.C.W.** Minority shares in Radiomics. **P.L.** Consulting fees from Radiomics and Benelux Health Ventures; co-inventor on patents licensed to Radiomics and has licensed software to Radiomics; unpaid advisory board member for Radiomics, Communicare Biotech, LivingMed Biotech, Bactam, and Convert Pharmaceuticals; minority shares in Radiomics and Communicare Solutions.

References

1. Mori M, Akashi-Tanaka S, Suzuki S, et al. Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-field digital mammography in a population of women with dense breasts. Breast Cancer 2017;24(1):104–110.
2. Zeeshan M, Salam B, Khalid QSB, Alam S, Sayani R. Diagnostic accuracy of digital mammography in the detection of breast cancer. Cureus 2018;10(4):e2448.
3. Cozzi A, Magni V, Zanardo M, Schiaffino S, Sardanelli F. Contrast-enhanced mammography: a systematic review and meta-analysis of diagnostic performance. Radiology 2022;302(3):568–581.
4. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94. [Published correction appears in Nature 2020;586(7829):E19.]
5. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48(4):441–446.
6. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295(2):328–338.
7. Conti A, Duggento A, Indovina I, Guerrisi M, Toschi N. Radiomics in breast cancer classification and prediction. Semin Cancer Biol 2021;72:238–250.
8. Baccouche A, Garcia-Zapirain B, Castillo Olea C, Elmaghraby AS. Connected-UNets: a deep learning architecture for breast mass segmentation. NPJ Breast Cancer 2021;7(1):151.
9. Ueda D, Yamamoto A, Onoda N, et al. Development and validation of a deep learning model for detection of breast cancers in mammography from multi-institutional datasets. PLoS One 2022;17(3):e0265751.
10. Lobbes MBI, Lalji U, Houwers J, et al. Contrast-enhanced spectral mammography in patients referred from the breast cancer screening programme. Eur Radiol 2014;24(7):1668–1676.
11. Jochelson MS, Lobbes MBI. Contrast-enhanced mammography: state of the art. Radiology 2021;299(1):36–48.
12. Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of

medical images in deep learning. arXiv 2003.04696 [preprint]. https://arxiv.org/abs/2003.04696.

13. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision, 2017; 2980–2988.
14. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–35.
15. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Sci Rep 2018;8(1):4165.
16. Mao N, Zhang H, Dai Y, et al. Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study. Br J Cancer 2023;128(5):793–804.
17. Wang S, Mao N, Duan S, et al. Radiomic analysis of contrast-enhanced mammography with different image types: classification of breast lesions. Front Oncol 2021;11:600546.
18. Lalji UC, Jeukens CRLPN, Houben I, et al. Evaluation of low-energy contrast-enhanced spectral mammography images by comparing them to full-field digital mammography using EUREF image quality criteria. Eur Radiol 2015;25(10):2813–2820.
19. Suter MB, Pesapane F, Agazzi GM, et al. Diagnostic accuracy of contrast-enhanced spectral mammography for breast lesions: a systematic review and meta-analysis. Breast 2020;53:8–17.
20. Perek S, Kiryati N, Zimmerman-Moreno G, Sklair-Levy M, Konen E, Mayer A. Classification of contrast-enhanced spectral mammography (CESM) images. Int J CARS 2019;14(2):249–257.
21. Reza AM. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for real-time image enhancement. J VLSI Signal Process Syst Signal Image Video Technol 2004;38(1):35–44.
22. Gaiser H, Liscio E, et al. fizyr/keras-maskrcnn 0.2.2. Zenodo.org. https://doi.org/10.5281/ZENODO.3250666. Published June 20, 2019. Accessed April 28, 2022.
23. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, 2014; 740–755.
24. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 1412.6980 [preprint]. https://arxiv.org/abs/1412.6980.
25. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–e107.
26. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112–118.