Radiology: Artificial Intelligence

AI IN BRIEF

Development and Validation of an AI-driven Mammographic Breast Density Classification Tool Based on Radiologist Consensus

Veronica Magni, MD • Matteo Interlenghi, MSc* • Andrea Cozzi, MD • Marco Alì, MSc, PhD • Christian Salvatore, MSc, PhD • Alcide A. Azzena, MD • Davide Capra, MD • Serena Carriero, MD • Gianmarco Della Pepa, MD • Deborah Fazzini, MD • Giuseppe Granata, MD • Caterina B. Monti, MD, PhD • Giulia Muscogiuri, MD • Giuseppe Pellegrino, MD • Simone Schiaffino, MD • Isabella Castiglioni, MSc, MBA • Sergio Papa, MD • Francesco Sardanelli, MD*

From the Department of Biomedical Sciences for Health (V.M., A.C., D.C., C.B.M., F.S.) and Postgraduate School in Radiodiagnostics (A.A.A., S.C., G.D.P., G.G., G.M., G.P.), Università degli Studi di Milano, Milan, Italy; DeepTrace Technologies, Milan, Italy (M.I., C.S.); Unit of Diagnostic Imaging and Stereotactic Radiosurgery, C.D.I. Centro Diagnostico Italiano, Milan, Italy (M.A., D.F., S.P.); Bracco Imaging, Milan, Italy (M.A.); Department of Science, Technology and Society, University School for Advanced Studies IUSS Pavia, Palazzo del Broletto, Piazza della Vittoria 15, 27100 Pavia, Italy (C.S.); Unit of Radiology, IRCCS Policlinico San Donato, San Donato Milanese, Italy (S.S., F.S.); Institute of Biomedical Imaging and Physiology, Consiglio Nazionale delle Ricerche, Segrate, Italy (I.C.); and Department of Physics, Università degli Studi di Milano-Bicocca, Milan, Italy (I.C.). Received July 16, 2021; revision requested September 8; revision received February 23, 2022; accepted March 3. **Address correspondence to** C.S. (e-mail: *salvatore@deeptracetech.com*).

* V.M. and M.I. contributed equally to this work.

Authors declared no funding for this work.

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

Radiology: Artificial Intelligence 2022; 4(2):e210199 • https://doi.org/10.1148/ryai.210199 • Content codes: AI BR

Mammographic breast density (BD) is commonly visually assessed using the Breast Imaging Reporting and Data System (BI-RADS) four-category scale. To overcome inter- and intraobserver variability of visual assessment, the authors retrospectively developed and externally validated a software for BD classification based on convolutional neural networks from mammograms obtained between 2017 and 2020. The tool was trained using the majority BD category determined by seven board-certified radiologists who independently visually assessed 760 mediolateral oblique (MLO) images in 380 women (mean age, 57 years ± 6 [SD]) from center 1; this process mimicked training from a consensus of several human readers. External validation of the model was performed by the three radiologists whose BD assessment was closest to the majority (consensus) of the initial seven on a dataset of 384 MLO images in 197 women (mean age, 56 years ± 13) obtained from center 2. The model achieved an accuracy of 89.3% in distinguishing BI-RADS a or b (nondense breasts) versus c or d (dense breasts) categories, with an agreement of 90.4% (178 of 197 mammograms) and a reliability of 0.807 (Cohen κ) compared with the mode of the three readers. This study demonstrates accuracy and reliability of a fully automated software for BD classification.

Supplemental material is available for this article.

Breast density (BD) is defined as the proportion of fibroglandular tissue relative to the total volume of the breast, as commonly assessed on mammograms (1). High BD is an independent risk factor for breast cancer and reduces mammography sensitivity by masking underlying lesions (2,3). The importance of reliable BD reporting was further heightened when legislation in the United States mandated that women be notified of their BD, as decisions about supplemental screening with US and MRI are based on mammographic density (4).

In clinical practice, BD is visually assessed on two-view mammograms, most commonly with the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) four-category scale (5): a (almost entirely fatty), b (scattered fibroglandular), c (heterogeneously dense), and d (extremely dense). However, this and other classification systems are affected by intra- and interobserver variability (6–8). To overcome the suboptimal reliability of visual assessment, fully automated software was proposed, including artificial intelligence (AI) systems, providing contrasting results (9–12). Considerable differences in BD classification persist when comparing repeated automatic assessment or comparing visual assessment by different human readers (13–15).

Therefore, the aim of this study was twofold: *(a)* to develop and externally validate an AI-driven fully automated software for BD classification mimicking a consensus-based human visual assessment; and *(b)* to assess its agreement and reliability in a clinical setting.

Materials and Methods

This work received no specific support from any funding agency. M.I. is an employee and shareholder of DeepTrace Technologies; I.C. owns DeepTrace Technologies shares; C.S. is CEO and shareholder of DeepTrace Technologies; and F.S. is a member of the scientific advisory board of DeepTrace Technologies. All the other authors are neither employees nor consultants of DeepTrace Technologies, and they had full control of all the data and information presented in this article.

This bicentric study was approved by the ethics committees of IRCCS Ospedale San Raffaele (protocol code SenoRetro, approved on November 9, 2017, amended on May 12, 2021) and of Fondazione IRCCS Ca' Granda

Abbreviations

AI = artificial intelligence, BD = breast density, BI-RADS = Breast Imaging Reporting and Data System, MLO = mediolateral oblique

Summary

A developed and externally validated radiologist consensus–based artificial intelligence–driven tool had high accuracy and agreement with radiologists in classifying nondense versus dense breasts.

Key Points

- The developed and externally validated radiologist consensus–based artificial intelligence (AI) tool for mammographic breast density classification achieved 89.3% accuracy in the nondense (Breast Imaging Reporting and Data System [BI-RADS] category a or b) versus dense (BI-RADS category c or d) classification task.
- The AI tool showed an agreement of 90.4% (178 of 197 mammograms) with radiologists' assessments and a reliability of 0.807 (Cohen κ) in the external validation mammography dataset.

Keywords

Mammography, Breast, Convolutional Neural Network (CNN), Deep Learning Algorithms, Machine Learning Algorithms

Ospedale Maggiore Policlinico (protocol code 1666, approved on October 14, 2020). Informed consent was waived due to the study's retrospective design.

The AI model (TRACE4BDensity; DeepTrace Technologies) (Appendix E1 [supplement]), consisting of three convolutional neural networks (ResNet50 [16]), was used to automatically classify BD of mediolateral oblique (MLO) views according to BI-RADS categories for three binary visual classification tasks: a and b (nondense breasts) versus c and d (dense breasts); a versus b; and c versus d; the last differentiation being a potential need considering the results of the DENSE (Dense Tissue and Early Breast Neoplasm Screening) trial (17). The final BD category of a given mammogram is assigned by the tool through assessment of both MLO views, using the most dense BD classification in case of difference between the two breasts.

AI Training and Tuning

The tool was developed, trained, cross-validated, and internally tested (Appendix E1 [supplement]) on a dataset built by retrieving all mammograms obtained in an organized population-based screening program at IRCCS Policlinico San Donato, Milan, Italy (center 1) between March 6 and May 2, 2017, acquired with Giotto Image 3DL (IMS, Italy) or Senographe Pristina 3D (GE Healthcare, France) systems. Mammograms presenting artifacts or technical limitations, breast implants, or lesions proven to be malignant were excluded. For each patient included, images of both the right and left breast were kept in the same partition during training, thus preventing the possibility that a patient could have images of one breast in the training partition and images of the other breast in the validation partition for the cross-fold validation.

To establish a reference labeling based on a consensus in human visual assessment, BD of MLO views was independently assessed by seven board-certified radiologists (among them S.S., F.S.). The majority category of individual classifications (mode) for each MLO view was obtained and used as a consensus-based reference for training (Appendix E2 [supplement]).

AI External Validation

The model was externally validated on a dataset of mammograms consecutively obtained for spontaneous screening between September 6 and October 4, 2020, at Centro Diagnostico Italiano, Milan, Italy (center 2), acquired using a Lorad Selenia system (Hologic).

To assess the agreement and reliability of the software compared with human classification, the three radiologists closest to the majority assessment of individual classifications (mode) performed independent BD classification of all mammograms. Having verified that the difference between the mode of the seven readers and the mode of the three readers closest to the mode was negligible (<2% in each class, Table E1 [supplement]), the external testing was carried out with the three readers to reduce the overall reading time. Again, the mode was calculated for each examination and used as a consensus-based reference.

Statistical Analysis

Results from visual and automatic BD classification were reported as percentages across BI-RADS categories. Overall and BI-RADS category-wise agreement between the model and the mode of the three radiologists closest to consensus were reported as percentages, while reliability was evaluated with Cohen κ statistics, reported with 95% CIs and interpreted according to the Landis and Koch scale (18). All statistical analyses were performed using software (SPSS version 26.0; IBM SPSS). A *P* value less than .05 was considered to indicate a significant difference.

Results

AI Training and Tuning

After application of the exclusion criteria on 974 preliminarily retrieved MLO views from center 1, 174 MLO views (six due to presence of lesions subsequently proven to be malignant, 52 due to the presence of breast implants, and 116 due to artifacts or technical limitations) were excluded. A final dataset of 800 MLO views from 400 women (mean age, 57 years ± 6 [SD]) was used for BD classification by human readers (Appendix E2 [supplement]) and for software training and internal testing (Appendix E1 [supplement]). For the latter purpose, the dataset was split into a fourfold cross-validation set (760 MLO views from 380 women; mean age, 57 years ± 6) and into an internal testing set (40 MLO views from 20 women; mean age, 57 years ± 7).

AI External Validation

For the external validation, 402 MLO views of 201 mammograms consecutively obtained at center 2 were retrieved. After the exclusion of 18 MLO views (12 due to the presence of breast implants and six due to image artifacts in breasts previously surgically treated), 384 MLO views of 197 mammo-

Table 1: Frequency Table of AI and Human Readings in the Four-Category Breast Density Classification Task on Mammograms of External Testing Dataset

External Validation	HR Category a	HR Category b	HR Category c	HR Category d	Total
AI category a	16	2	0	0	18
AI category b	7	61	14	0	82
AI category c	0	5	60	13	78
AI category d	0	0	2	17	19
Total	23	68	76	30	197

Note.—Data shown are number of mammograms. Breast Imaging Reporting and Data System: category a (almost entirely fatty), category b (scattered fibroglandular), category c (heterogeneously dense), category d (extremely dense). AI = artificial intelligence, HR = human readings.

Table 2: Agreement between AI and Human Readings in the Breast Density Classification Tasks (Four-Category and Nondense vs Dense Breasts) on Mammograms of External Testing Dataset

External Validation	HR Category a (n = 23)	HR Category b (n = 68)	HR Category c (n = 76)	HR Category d (n = 30)	Overall
AI-HR agreement in four BD categories	69.6% (16/23) [47.1, 86.8]	89.7% (61/68) [79.9, 96.8]	78.9% (60/76) [68.1, 87.5]	56.7% (17/30) [37.4, 74.5]	78.2% (154/197) [71.7, 83.7]
AI-HR agreement in nondense vs dense breasts	94.5% (86/91) [87.6, 98.2]		86.8% (92/106) [78,8, 92.6]		90.4% (178/197) [85.3, 94.1]

Note.—Agreement percentages are reported with 95% CIs in brackets. Breast Imaging Reporting and Data System: category a (almost entirely fatty), category b (scattered fibroglandular), category c (heterogeneously dense), category d (extremely dense); nondense breasts (category a or b), dense breasts (category c or d). AI = artificial intelligence, HR = human readings, BD = breast density.

Table 3: Reliability between AI and Human Readings in the Breast Density Classification Tasks (Four-Category and Nondense vs Dense Breasts) on Mammograms of External Testing Dataset

External Validation	Category a	Category b	Category c	Category d	Overall (Four-Category, Unweighted κ)	Overall (Four-Category, Linear-weighted κ)	Overall (Nondense vs Dense)
Cohen κ	0.755 [0.615, 0.895]	0.699 [0.559, 0.838]	0.638 [0.498, 0.777]	0.650 [0.511, 0.790]	0.677 [0.586, 0.768]	0.759 [0.694, 0.825]	0.807 [0.667, 0.947]

Note.—Cohen κ are reported with 95% CIs in brackets. Breast Imaging Reporting and Data System: category a (almost entirely fatty), category b (scattered fibroglandular), category c (heterogeneously dense), category d (extremely dense); nondense breasts (category a or b), dense breasts (category c or d).

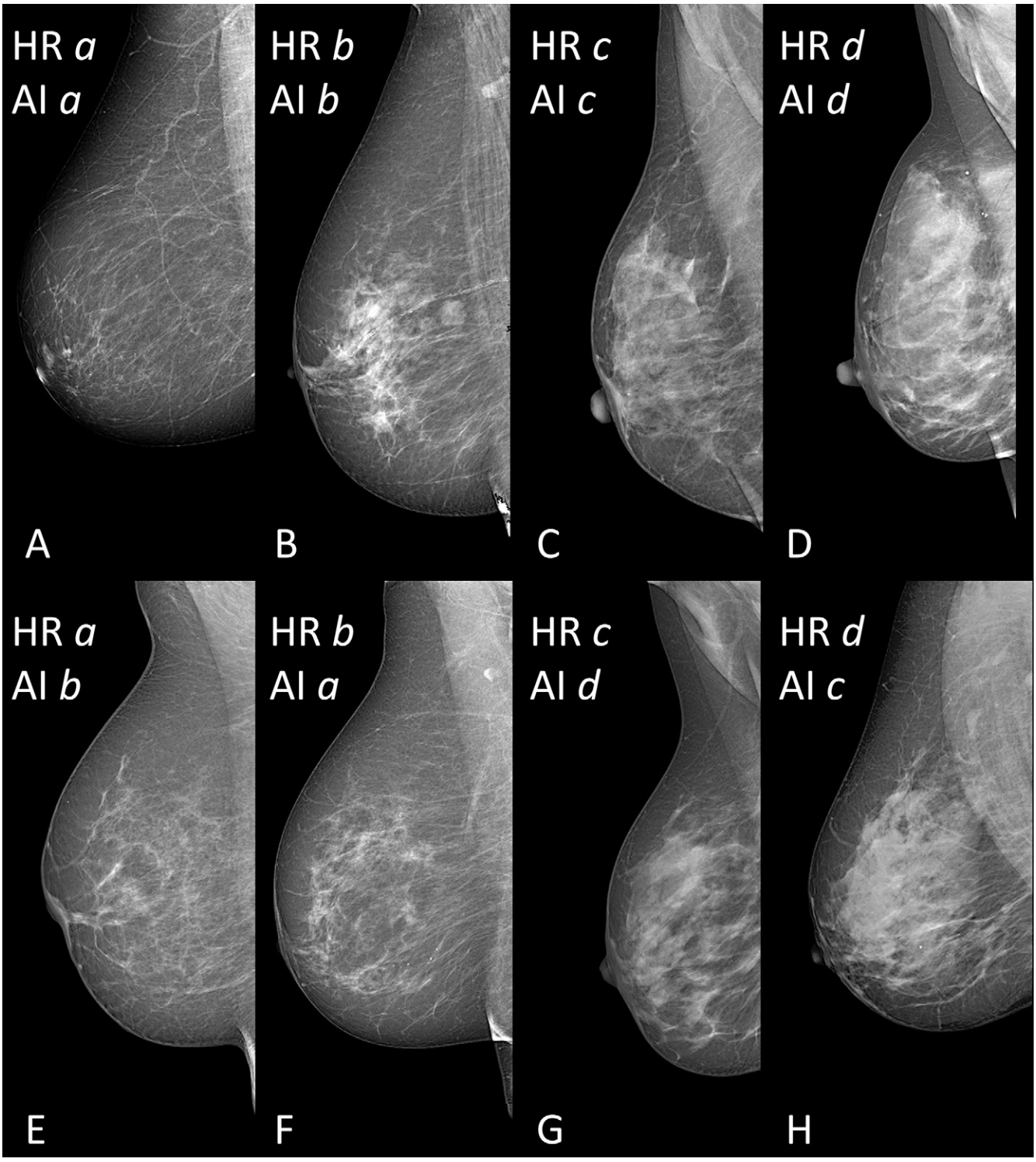
grams from 197 women (mean age, 56 years ± 13) were automatically classified by the tool and assessed by the three human readers closest to consensus.

The model showed an 89.3% accuracy (343 of 384; 95% CI: 85.8, 92.2) in the two-category BD classification (nondense vs dense breasts) and a 75.0% accuracy (288 of 384; 95% CI: 70.4, 79.3) in the four-category BD classification.

The distribution of BD categories assigned by the model and human readers are reported in Table 1. Human-software agreement was found in 154 of 197 mammograms (78.2%; 95% CI: 71.7, 83.7), ranging from 56.7% (category d) to 89.7% (category b). Human-software reliability analysis yielded an overall linear-weighted Cohen κ of 0.759 (95% CI: 0.694, 0.825). Furthers details, including within category and reader-specific data, are given in Tables 1–3.

Disagreements were found always between adjacent categories: two of 43 (4.7%) mammograms were classified as category a by the software but category b by the readers; 14 of 43 (32.6%) as category b by the software but category c by the readers; seven of 43 (16.3%) as category b by the software but category a by the readers; 13 of 43 (30.2%) as category c by the software but category d by the readers; five of 43 (11.6%) as category c by the software but category b by the readers; and two of 43 mammograms (4.7%) were classified as category d by the software but category c by the readers.

When considering the two-category BD classification task (nondense vs dense breasts), human-software agreement was 90.4% (95% CI: 85.3, 94.1; 178 of 197 mammograms). Reliability analysis showed Cohen κ of 0.807 (95% CI: 0.667, 0.947).



Selection of mammographic mediolateral oblique views of breasts with different breast density from women between 51 and 68 years of age. **(A–D)** Examples of human readers (HR)–artificial intelligence (AI) agreement for category a (68 years), b (66 years), c (51 years), and d (54 years); **B** shows an example of a breast with a benign mass. **(E–H)** Examples of HR-AI disagreement; **E** was classified as category a by HR, and as category b by AI (67 years); **F** was classified as category b by HR, and as category a by AI (68 years); **G** was classified as category c by HR, and as category d by AI (55 years); **H** was classified as category d by HR, and as category c by AI (52 years). Breast Imaging Reporting and Data System: category a (almost entirely fatty), category b (scattered fibroglandular), category c (heterogeneously dense), category d (extremely dense).

The Figure shows mammograms where agreement (Fig A–D) and disagreements (Fig E–H) occurred. No specific pattern was identified where the tool failed.

Discussion

We developed and externally validated an AI software for mammographic BD assessment, which had an 89.3% accuracy for

the nondense versus dense breasts classification task and substantial agreement of 90.4% and reliability (Cohen κ 0.807) with radiologist readings. Distinguishing dense from nondense breasts is indeed the clinically most relevant task and may drive the potential referral to supplemental screening (19). This result offers a robust way to overcome the variability of human visual assessment.

This study had limitations. First, six images with lesions subsequently proven to be malignant were excluded to avoid training the classifier on mammograms of cases for which the breast contralateral to that containing the diagnosed cancer is usually considered for assigning the woman's BD category. Indeed, when the radiologist finds a lesion, BD is classified considering the lesion background or the contralateral breast. We consider this omission to be negligible as it would amount to 0.8% of the MLO views in the training set. Second, reliability analysis was performed only by three human readers, although they were the readers closest to the majority assessment. Finally, the model does not provide quantitative BD measurements; however, the BI-RADS category system provides the breast cancer risk component attributable to BD for the Tyrer-Cuzick predictive model (20), thus the model can be used to drive referral to supplemental screening. Moreover, disagreement has been proven between automated quantitative BD assessment and human BD assessment, as there is currently no reference standard to validate BD measurement.

In conclusion, an AI software based on radiologist consensus was developed and externally validated, being able to automatically classify dense versus nondense breasts on mammograms according to BI-RADS categories.

Acknowledgments: The authors wish to express their gratitude to the colleagues who contributed to the organization and human reading phase of this study: Adrienn Benedek, MD; Luca Alessandro Carbonaro, MD; Maria Iodice, MD; Laura Menicagli, MD; Cristian Giuseppe Monaco, MD; Diana Spinelli, MD; and Rubina Manuela Trimboli, MD, PhD.

Author contributions: Guarantors of integrity of entire study, **V.M., M.I., C.S., M.A., I.C., S.P., F.S.**; 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, **V.M., M.I., A.C., M.A., C.S., A.A.A., S.C., G.D.P., D.F., I.C., S.P., F.S.**; clinical studies, **V.M., A.C., S.C., G.D.P., D.F., G.G., G.P., S.S., M.A., S.P., F.S.**; statistical analysis, **V.M., M.I., A.C., M.A., C.S., S.C., I.C.;** and manuscript editing, **V.M., M.I., A.C., M.A., C.S., S.C., I.C., S.P., F.S.**

Disclosures of conflicts of interest: V.M. No relevant relationships. **M.I.** CTO and employee of DeepTrace Technologies. DeepTrace Technologies is a spin-off of Scuola Universitaria Superiore IUSS, Pavia, Italy; shareholder in DeepTrace Technologies. **A.C.** No relevant relationships. **M.A.** Scientific advisor for Bracco Imaging. **C.S.** CEO of DeepTrace Technologies. DeepTrace Technologies is a spin-off of Scuola Universitaria Superiore IUSS, Pavia, Italy; shareholder in DeepTrace Technologies. **A.A.A.** No relevant relationships. **D.C.** No relevant relationships. **S.C.** No relevant relationships. **G.D.P.** No relevant relationships. **D.F.** No relevant relationships. **G.G.** No relevant relationships. **C.B.M.** No relevant relationships. **G.M.** No relevant relationships. **G.P.** No relevant relationships. **S.S.** Honoraria for lectures from GE Healthcare; support for attending meetings/travel from GE Healthcare. **I.C.** Shareholder in DeepTrace Technologies. **S.P.** No relevant relationships. **F.S.** Member of Scientific Advisory Board for DeepTrace Technologies.

References

1. Nazari SS, Mukherjee P. An overview of mammographic density and its association with breast cancer. Breast Cancer 2018;25(3):259–267.
2. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. RadioGraphics 2015;35(2):302–315.
3. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev 2006;15(6):1159–1169.
4. Keating NL, Pace LE. New Federal Requirements to Inform Patients About Breast Density: Will They Help Patients? JAMA 2019;321(23):2275–2276.
5. Sickles EA, D'Orsi CJ, Bassett LW. ACR BI-RADS Mammography. ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. Reston, Va: American College of Radiology, 2013.
6. Damases CN, Hogg P, McEntee MF. Intercountry analysis of breast density classification using visual grading. Br J Radiol 2017;90(1076):20170064.
7. Alomaim W, O'Leary D, Ryan J, Rainford L, Evanoff M, Foley S. Variability of Breast Density Classification Between US and UK Radiologists. J Med Imaging Radiat Sci 2019;50(1):53–61.
8. Sprague BL, Conant EF, Onega T, et al. Variation in Mammographic Breast Density Assessments Among Radiologists in Clinical Practice: A Multicenter Observational Study. Ann Intern Med 2016;165(7):457–464.
9. Castiglioni I, Rundo L, Codari M, et al. AI applications to medical images: From machine learning to deep learning. Phys Med 2021;83:9–24.
10. Sartor H, Lång K, Rosso A, Borgquist S, Zackrisson S, Timberg P. Measuring mammographic density: comparing a fully automated volumetric assessment versus European radiologists' qualitative classification. Eur Radiol 2016;26(12):4354–4360.
11. Yeo I, Akwo J, Ekpo E. Automated mammographic density measurement using Quantra™: comparison with the Royal Australian and New Zealand College of Radiology synoptic scale. J Med Imaging (Bellingham) 2020;7(3):035501.
12. Moshina N, Roman M, Sebuødegård S, Waade GG, Ursin G, Hofvind S. Comparison of subjective and fully automated methods for measuring mammographic density. Acta Radiol 2018;59(2):154–160.
13. Youk JH, Gweon HM, Son EJ, Kim JA. Automated Volumetric Breast Density Measurements in the Era of the BI-RADS Fifth Edition: A Comparison With Visual Assessment. AJR Am J Roentgenol 2016;206(5):1056–1062.
14. Brandt KR, Scott CG, Ma L, et al. Comparison of Clinical and Automated Breast Density Measurements: Implications for Risk Prediction and Supplemental Screening. Radiology 2016;279(3):710–719.
15. Lehman CD, Yala A, Schuster T, et al. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. Radiology 2019;290(1):52–58.
16. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 27–30, 2016. Piscataway, NJ: IEEE, 2016; 770–778.
17. Bakker MF, de Lange SV, Pijnappel RM, et al. Supplemental MRI Screening for Women with Extremely Dense Breast Tissue. N Engl J Med 2019;381(22):2091–2102.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.
19. Expert Panel on Breast Imaging; Weinstein SP, Slanetz PJ, et al. ACR Appropriateness Criteria® Supplemental Breast Cancer Screening Based on Breast Density. J Am Coll Radiol 2021;18(11S):S456–S473.
20. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 2004;23(7):1111–1130.