Journal of Breast Imaging, 2020, Vol. 2, Issue 4, 304–314
doi:10.1093/jbi/wbaa033
Scientific Review
Received: December 20, 2019; Editorial Acceptance: April 13, 2020
Published Online: June 19, 2020

Scientific Review

Artificial Intelligence: A Primer for Breast Imaging Radiologists

Manisha Bahl, MD, MPH[1,*]

[1]Massachusetts General Hospital, Department of Radiology, Boston, MA

*Address correspondence to M.B. (e-mail: mbahl1@mgh.harvard.edu)

Abstract

Artificial intelligence (AI) is a branch of computer science dedicated to developing computer algorithms that emulate intelligent human behavior. Subfields of AI include machine learning and deep learning. Advances in AI technologies have led to techniques that could increase breast cancer detection, improve clinical efficiency in breast imaging practices, and guide decision-making regarding screening and prevention strategies. This article reviews key terminology and concepts, discusses common AI models and methods to validate and evaluate these models, describes emerging AI applications in breast imaging, and outlines challenges and future directions. Familiarity with AI terminology, concepts, methods, and applications is essential for breast imaging radiologists to critically evaluate these emerging technologies, recognize their strengths and limitations, and ultimately ensure optimal patient care.

Key words: artificial intelligence; machine learning; deep learning; breast imaging; mammography.

Introduction

Artificial intelligence (AI) in radiology is quickly progressing from pilot and feasibility studies to clinical implementation. These recent advances in AI have been driven by advanced computer algorithms, increased availability of large datasets, and improved computing power (1,2). Traditional computer-aided detection (CAD), introduced in the 1990s, is based on features perceived by humans (eg, density and shape), whereas AI algorithms discover the features that are necessary to classify the data and have the potential to discover useful features that are currently unknown or beyond the limits of human detection (3–5).

Artificial intelligence is poised to enhance the quality and value of radiology's contribution to patient care and improve radiologists' workflows (2,6–11). Advances in AI technologies have led to marked improvements in their clinical utility. In breast imaging, for example, AI has the potential to enhance radiologists' accuracy by improving sensitivity for the detection of breast cancers and reducing false-positive assessments (12–16). Beyond improving the accuracy of interpretation, AI has the potential to help radiologists accurately assess an individual woman's risk of breast cancer, guide decision-making regarding high-risk lesions, decrease interpretation times, quickly identify cancer-free mammograms and therefore reduce workload, predict the risk of concurrent invasive cancer in patients with ductal carcinoma in situ (DCIS), provide early prediction of neoadjuvant chemotherapy response, and predict lymph node metastasis in patients with breast cancer (17–27).

This article reviews key terminology and concepts, discusses common AI models and methods to validate and evaluate these models, describes emerging AI applications in breast imaging, and outlines challenges and future directions. Specific AI applications discussed in this review include algorithms to increase breast cancer detection, improve clinical efficiency, and accurately assess breast cancer risk. Familiarity with the concepts presented in this review is essential for breast imaging radiologists to be able to critically evaluate emerging AI technologies and recognize their strengths and limitations.

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

© Society of Breast Imaging 2020. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Key Messages

- Advances in artificial intelligence (AI) technologies have led to techniques that could increase breast cancer detection, improve clinical efficiency in breast imaging practices, and guide decision-making regarding screening and prevention strategies.
- Artificial intelligence models require both internal and external validation, and techniques used for model evaluation and interpretability include confusion matrices, receiver operating characteristic curves, and heat maps.
- Familiarity with AI terminology, concepts, methods, and applications is essential for breast imaging radiologists to critically evaluate these emerging technologies, recognize their strengths and limitations, and ultimately ensure optimal patient care.

Terms and Techniques

In order to understand and critically evaluate AI literature, breast imaging radiologists must be familiar with AI terminology and concepts, common AI models, and methods to validate and evaluate these models.

Artificial Intelligence and Its Subfields

Artificial intelligence is a branch of computer science dedicated to developing computer algorithms that emulate intelligent human behavior, such as learning, recognizing patterns, reasoning, solving problems, making decisions, and self-correcting (2). Artificial intelligence is a broad umbrella term that encompasses machine learning (ML) and deep learning (DL) (1,28) (Figure 1, Table 1). ML is a subfield of AI in which the computer learns from provided data without being explicitly programmed. The ML algorithm is developed to maximize the fit between the input (eg, text or images) and output (eg, classification), and it can then be applied to new data (2). DL is a subfield of ML that relies on neural networks with multiple layers to progressively extract higher-level features from raw data (2,29–32) (Figure 2). Network architectures with numerous and large layers are "deep" learning neural networks, as opposed to "shallow" learning neural networks with only a few layers (33). The various layers can be used to detect complex features, such as shapes, from simpler features, such as image intensities, to decode image data (31).

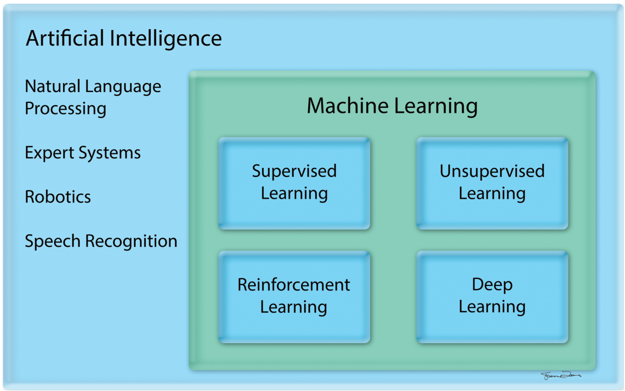

Figure 1. Hierarchy of artificial intelligence fields (28).

Learning Processes

The learning process for ML algorithms can be supervised, unsupervised, or based on reinforcement (1,2,9). In supervised learning, the algorithm is provided with labeled data (eg, mammographic images labeled as positive or negative for breast cancer) (1,34). Two examples of supervised learning are classification (in which the output is categorical or a class) and regression (in which the output is numeric or continuous) (2). Supervised learning requires a large amount of data for learning (and thus computational power), accurate labeling, and an agreed-upon definition of the ground truth (ie, the "correct labels" or "true labels" for the data) (2). In unsupervised learning, the algorithm is provided with unlabeled data, and the ML algorithm clusters or organizes the data to uncover underlying patterns (1,34). An example of unsupervised learning is clustering (in which the data are partitioned or clustered into classes). A hybrid approach is semisupervised learning, in which a large amount of unlabeled data and a small number of labeled examples are provided to the computer (1). In reinforcement learning, the algorithm learns from positive and negative feedback without being taught (eg, a robot learning to walk or an autonomous car learning to avoid other cars) (9,35).

Most published AI models in the breast imaging literature, which will be discussed in subsequent sections, utilize supervised learning, in which the computer is provided with labeled data. For example, when developing an algorithm for breast cancer detection, the mammographic images provided to the computer are labeled as positive or negative for breast cancer. While a traditional ML algorithm would rely on human-engineered (or manually designed) features based on clinicians' knowledge and experience (eg, density or shape), DL algorithms learn the features that are necessary to classify the mammographic images as positive or negative, improve with exposure to more data, and have the potential to discover features and relationships that are currently unknown or imperceptible to humans (3,36) (Figure 3). If a large enough training dataset is provided, AI systems based on DL could potentially classify data better than if human-engineered features were used (1).

Common Models

Examples of ML models include support vector machine, random forest, and neural networks. Support vector machine is used in the setting of large numbers of features to discriminate data into two or more classes (3,37). The algorithm finds a straight or curved line, or "hyperplane," to separate the classes with as wide a margin as possible. Random forest uses an ensemble of decision trees based on random subsets

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

Table 1. Definitions of AI Terminology

Term	Definition
Artificial intelligence	Branch of computer science dedicated to developing computer algorithms that emulate intelligent human behavior, such as learning, recognizing patterns, reasoning, solving problems, making decisions, and self-correcting.
Classification	A supervised learning method to predict class membership of an observation.
Deep learning	A subfield of machine learning that relies on neural networks with multiple layers to progressively extract higher-level features from raw data.
External validation	Validation of a model using data from a source that is different from the training data.
Ground truth	Correct labels (or true labels) for data, as determined by experts or other reference standards.
Hidden layer	A synthetic layer in a neural network between the input layer (ie, the features) and the output layer (ie, the prediction).
Internal validation	Validation of a model using data from the same source as the training data.
Machine learning	A subfield of artificial intelligence in which computers learn without being explicitly programmed.
Neural network	A multi-layer network that resembles the connectivity of neurons in the brain.
Overfitting	Occurs when a model is trained to predict the training dataset so well that it may fail to make a good prediction on new data.
Regression	A supervised learning method to predict output with continuous value.
Reinforcement learning	A type of machine learning in which the algorithm learns from positive and negative feedback without being taught.
Supervised learning	A type of machine learning in which the algorithm is provided with labeled training data.
Test set	A subset of the dataset that is used to evaluate the model.
Training set	A subset of the dataset that is used to develop the model.
Unsupervised learning	A type of machine learning in which the algorithm is provided with training data without corresponding labels.
Validation set	A subset of the dataset that is used to fine-tune the model's parameters.

References (2,30,41).

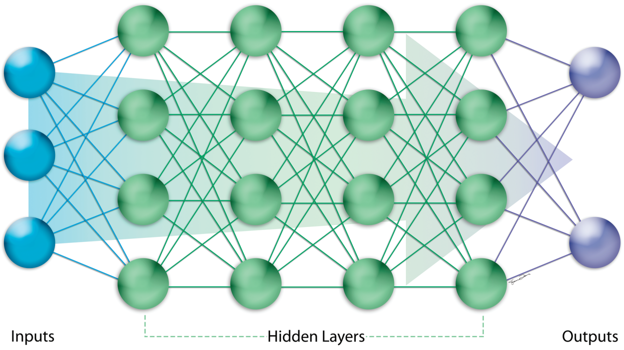

Figure 2. Structure of a neural network. A neural network is composed of groups of nodes with consecutive layers—an input layer, one or more hidden layers, and an output later.

of features from the training data (37). When a new input (eg, image) is presented, a prediction from each tree is made (eg, "benign" or "malignant"), and then the best solution is generated through "voting" by each of the trees.

Neural networks, which form the basis of DL models, resemble the connectivity of neurons in the brain (2). In a neural network, there are layers of connected nodes, each of which receives inputs from other nodes. After a node receives information, and if a certain threshold is met, it then transmits a signal to other nodes. This signal may vary in its strength (or weight)—that is, weak inputs lead to a small signal, while appropriate inputs lead to a strong signal. Initially during development, node inputs and weights are randomly set, and the output is compared to the true label provided by a human (if supervised learning is used). The neural connections and weights are then altered to produce the output again. This process is repeated many times, each time improving the connections and weights to maximize the likelihood that the generated output matches the true label provided by a human.

Convolutional neural networks (CNNs) are the most common type of neural networks used for image analysis, as these networks can perform well with two-dimensional and volumetric images (31,38). Convolutional neural networks have an input layer (which receives input data), one or more hidden layers (which extract patterns within the data), and an output layer (which produces the results of data processing) (31). Better performance is generally achieved with deeper architecture (ie, more hidden layers) that can be used to extract more features; however, the addition of more layers can also lead to "overfitting" the network, which occurs when the algorithm is trained to predict the training dataset so well that it does not perform well with new, previously unseen data (39,40). Better performance of CNNs is also achieved with exposure to more data.

Model Validation

Internal validation refers to the validation of a model using data from the same source as the training data, and external validation refers to the validation of a model using data from a source that is different from the training data (41). Common

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

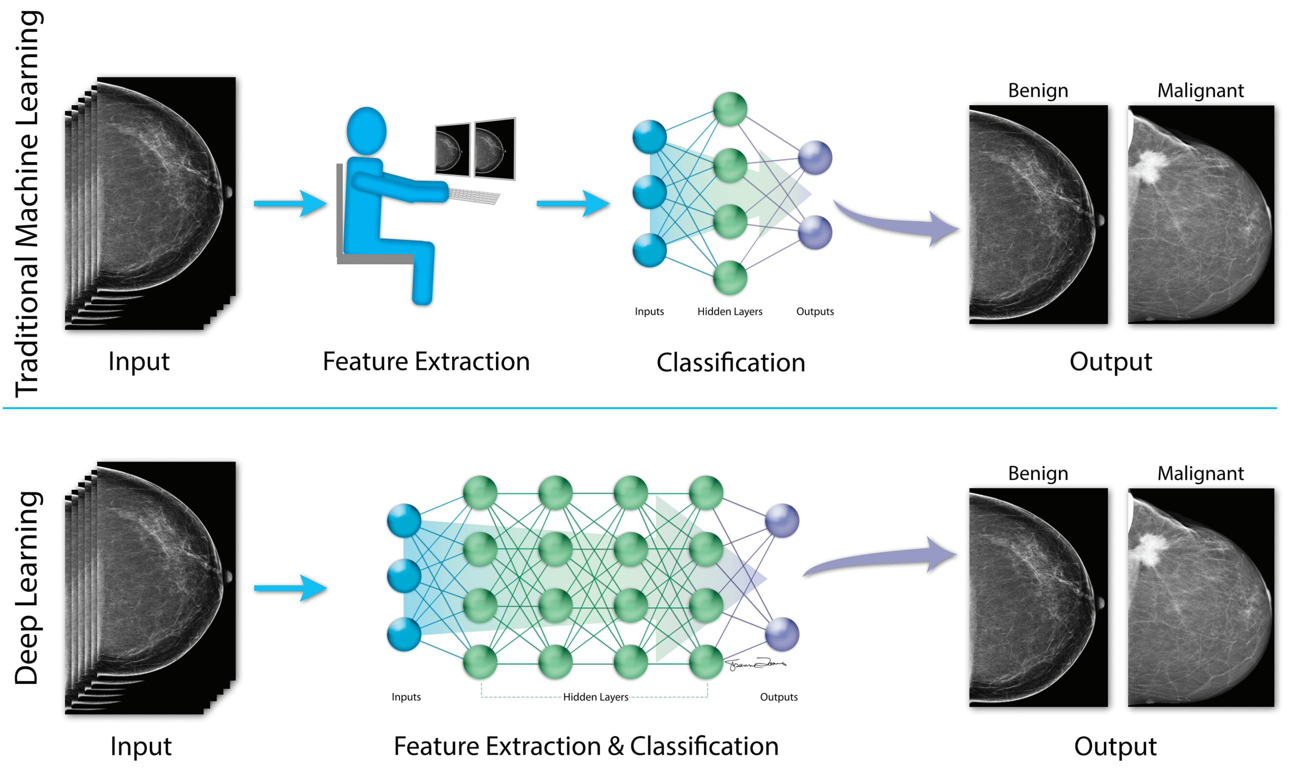

Figure 3. Comparison of traditional ML and DL (3,36). Most published models in the breast imaging literature utilize supervised learning, in which the computer is provided with labeled data (eg, inputs are mammographic images that are labeled as "benign" or "malignant"). The training processes for traditional ML models and DL models differ in that the traditional ML model is based on human-engineered features, whereas the DL model learns the features that are necessary to classify the mammographic images as "benign" or "malignant" without human input. Once trained, the traditional ML model and the DL model could then classify a previously unseen mammographic image as "benign" or "malignant." Abbreviations: DL, deep learning; ML, machine learning.

internal validation methods include random split and k-fold cross-validation, as described below (40,41). Model performance should also be validated with a completely external dataset (eg, data collected by independent investigators at a different site) to confirm its generalizability for clinical practice (42).

In a random split, the dataset is randomly divided into a training set, validation set, and test set. In radiology, the training process occurs with a set of images for which the "ground truth" is known—that is, the "correct labels" or "true labels" for the images, as determined by experts or other reference standards (eg, diagnostic tests or pathology results) (32). The model parameters (or variables) are updated iteratively until the fit between the input (eg, images) and output (eg, classification) is maximized (40,43). The validation set is independent from the training set and is used to fine-tune the model's parameters, while a test set is used to evaluate the model performance (40). Ideally, the training, validation, and test sets should be independent, without overlap (32,43). Of note, in the field of AI, the term "validation" can refer to internal versus external validation of a model or can refer to the fine-tuning stage of model development (42).

In k-fold cross-validation, multiple pairs of training and test sets are created from one dataset (40). The dataset is split into k different subsets, with k chosen based on the size of the dataset. The cross-validation process is repeated, with each of the subsets used once as the test set and all other subsets combined to form the training set. These results are then averaged to provide a single estimation. This technique can be useful for smaller datasets but requires more computational power.

Model Evaluation

Confusion matrices or receiver operating characteristic (ROC) curves can be used to evaluate model performance (44,45). A confusion matrix provides information about the classification performance of a model on test data for which the true labels are known (Figure 4) (45). The information is presented in a table format, in which each column represents instances of the predicted label and each row represents instances of the true label (or vice versa). Using the table, the reader can easily visualize if the model is "confusing" two classes (ie, if the model is mislabeling one class as another one).

An ROC curve is plotted on a graph, in which the x-axis is the false-positive rate (or 1-specificity) and the y-axis is the true-positive rate (or sensitivity) (Figure 5) (44). Each point on the ROC curve represents a different decision threshold, with tradeoffs between the false-positive and true-positive rates; that is, as the sensitivity increases so does the false-positive rate. The accuracy of the test can be summarized by the area under the ROC curve (AUC), which can range from 0 to 1.0. An ROC curve with an AUC of 1.0 represents

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

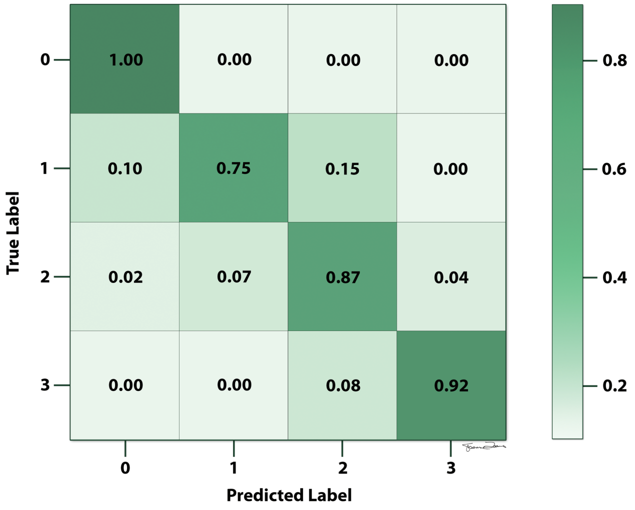

Figure 4. Example of a confusion matrix. The reader can visualize if the model is "confusing" two classes (ie, if the model is mislabeling one class as another one). In this example, the true label of 0 is predicted with 100% accuracy, the true label of 1 is predicted with 75% accuracy, the true label of 2 is predicted with 87% accuracy, and the true label of 3 is predicted with 92% accuracy.

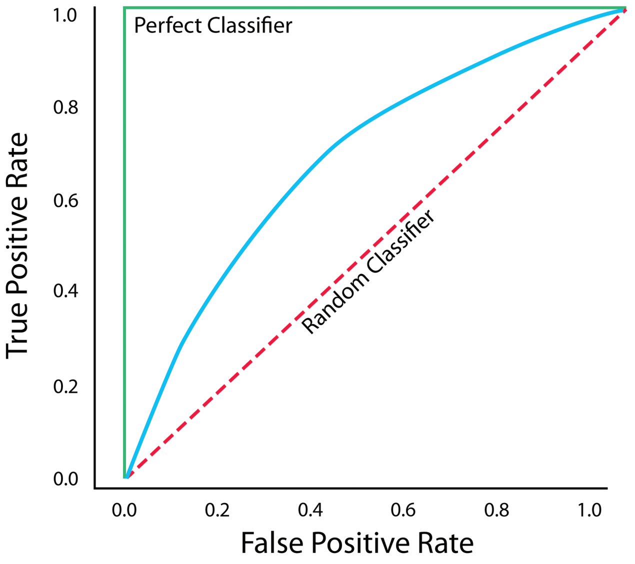

Figure 5. Example of a receiver operating characteristic curve. The green line represents a perfect classifier (with an area under the curve [AUC] of 1.0), and the dotted red line represents a random classifier (with an AUC of 0.5). Models with AUCs above 0.5 (blue line) have at least some ability to discriminate between classes, with better models having AUCs closer to 1.0.

a perfect classifier, in which the sensitivity is 1.0 when the false-positive rate is 0. An ROC curve with an AUC of 0.5 represents a random classifier. Models with AUCs above 0.5 have at least some ability to discriminate between classes, with better models having AUCs closer to 1.0.

Model Interpretability

Certain models, such as those based on neural networks, are considered "black boxes" in that the imaging features or patterns used by the model to make predictions may not be readily evident to radiologists. One method to address this lack of interpretability is heat maps (or saliency maps), which are used to indicate the most salient regions within images and thus draw attention to the specific regions that contribute most to the corresponding output by the model (Figure 6) (46,47). For example, in breast imaging, a heat map of a mammogram can be generated by color-coding the image on a pixel-wise basis based on the likelihood of breast cancer, thus demonstrating regions that are most commonly encountered in patients with breast cancer (eg, red) and without breast cancer (eg, green).

Breast Imaging Applications

Applications of AI in three domains will be reviewed: breast cancer detection, clinical efficiency, and risk stratification. These published models are largely DL-based algorithms using a supervised learning process unless otherwise noted.

Breast Cancer Detection

Multiple studies evaluating traditional CAD for breast cancer have had mixed results (48–54). Traditional CAD systems are programmed to identify or mark imaging features that are known to be associated with a specific disease (eg, breast cancer), whereas AI systems learn how to extract imaging features that are visible or invisible to the human eye (31). For example, a traditional CAD algorithm could identify a feature (eg, calcifications) and determine if the feature (eg, calcifications) is present or absent on a mammographic image, but such a system may not reliably differentiate between benign calcifications and those seen in the setting of DCIS. In contrast, an AI algorithm would be trained to focus on the outcome of DCIS and, after exposure to a large amount of data, could learn to identify specific features or patterns of features that are associated with DCIS. In addition, a strength of an AI system is that it could continually improve with exposure to additional data, whereas a traditional CAD system would require modifications by humans in order to improve its performance (54).

In a recent multireader study, investigators compared the cancer detection performance of radiologists interpreting screening mammograms with and without the support of a DL-based AI system (14). The AI system provided radiologists with certain decision support tools, including traditional lesion markers, local cancer likelihood scores activated by clicking on specific areas, and a cancer likelihood score based on the entire examination. Each of the 14 radiologists interpreted 240 digital mammograms (enriched with 100 cancers), once with and once without the AI system. The AUC was higher with the AI system (0.89 versus 0.87, $P = 0.002$), but this improvement was observed with less-experienced radiologists and not with expert radiologists. The observed differences in performance were statistically significant but quite small, bringing into question whether integration of the

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

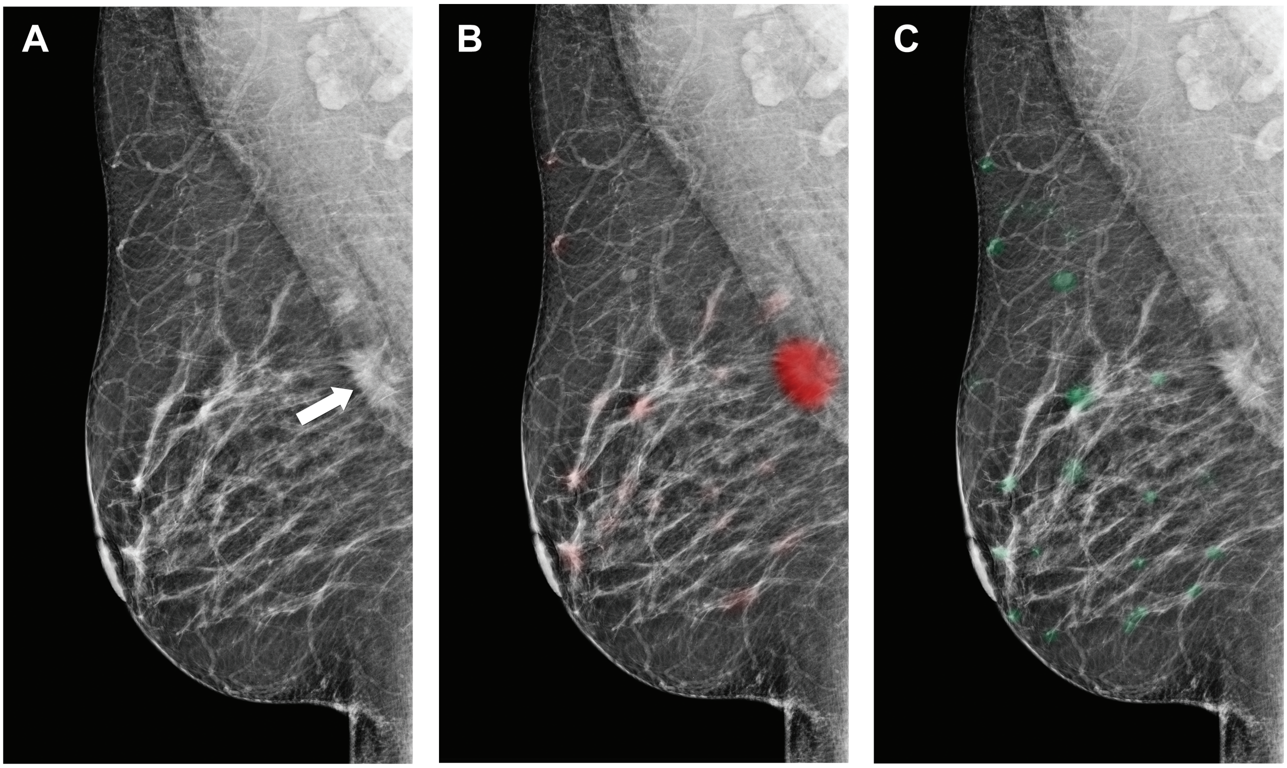

Figure 6. Examples of heat maps or saliency maps. Mediolateral oblique views of the right breast demonstrate an invasive ductal cancer in the superior aspect of the breast at posterior depth (**A**, arrow), with overlying malignant heat map in red (**B**) and overlying benign heat map in green (**C**).

AI system into clinical practice would meaningfully impact performance metrics and clinical outcomes (55).

In a subsequent study with the same AI system, the investigators compared the stand-alone performance of the AI system to that of 101 radiologists (15). The analysis included digital mammograms from four different vendors and interpretations by radiologists from the U.S. and Europe. The authors reported that the stand-alone performance of the AI system was noninferior to that of the radiologists' (AUC of 0.84 versus 0.81) and that the AI system had a higher AUC than 61.4% of the radiologists. The results suggest that this AI system could serve as a stand-alone first or second reader in screening programs and that it could help radiologists with varying levels of training and experience achieve performance benchmarks. Although this AI system performs well, it could potentially be strengthened if it incorporated comparisons to patients' prior examinations and to the contralateral breasts. As the two studies done by this group of investigators on the specific AI system were retrospective in design and based on reader studies with enriched sets of mammograms, a prospective evaluation in the clinical setting would be necessary before widespread deployment.

Researchers from New York University trained and evaluated a convolutional neural network with more than 225 000 exams, which achieved an AUC of 0.90 in predicting the presence of breast cancer (47). A reader study was then conducted with 14 readers (12 attending radiologists, a resident, and a medical student), with each reader interpreting 720 screening mammograms. The model achieved an AUC of 0.88, while individual readers achieved AUCs of 0.71 to 0.86. A recent algorithm trained with more than 170 000 exams collected from South Korea, the U.S., and the UK improved radiologists' performance from an AUC of 0.81 to 0.88 ($P < 0.0001$) (56). The algorithms presented in these two studies are promising diagnostic support tools but require validation in real-world clinical environments (57).

Researchers from Google Health (Mountain View, CA) and DeepMind Technologies (London, UK) recently published a report on the performance of an AI system using datasets from the UK and the U.S. (58). The UK dataset consisted of screening mammograms from 25 856 women obtained at 2 centers in England, and the U.S. dataset consisted of cancer-enriched screening mammograms from 3097 women obtained at 1 center. Absolute reductions of 5.7% and 1.2% in false positives (U.S. and UK, respectively) and 9.4% and 2.7% in false negatives (U.S. and UK, respectively) were demonstrated with the AI system. When the AI system was retrained with UK data only and performance was measured on unseen U.S. data, the system continued to outperform radiologists but by a smaller margin, suggesting that AI systems could benefit from fine-tuning with local data. In an independent study of six radiologists, all of whom were eligible to interpret mammograms in the U.S. but did not uniformly receive breast imaging fellowship training,

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

the AUC for the AI system was greater than the AUC for the average radiologist by an absolute margin of 11.5%. In addition, in a simulation in which the AI system was used in the double-reading system (which is done in the UK), the AI system maintained noninferior performance compared to the second reader and also reduced the workload of the second reader by 88%. Limitations of the study include the nonrepresentative U.S. dataset, which came from a single center and was enriched with cancer cases, and the use of images primarily from a single manufacturer (59). Clinical trials will be needed to evaluate the utility of this AI system in real-world practice.

The studies discussed above focus on mammographic imaging only. In a recent study, investigators integrated both mammographic imaging and clinical information in a combined "ML-DL model" to predict one-year breast cancer risk, which achieved an AUC of 0.91 (12). Images alone achieved an AUC of 0.88, and clinical data alone achieved an AUC of 0.78 (12). In addition, the model identified breast cancer in 48% of women (34 of 71) in whom the radiologist had interpreted the examination as negative but in whom cancer was detected within one year (12). The authors suggest that this model holds promise as a second reader for mammographic exams; however, it does not yet offer localization of specific findings, only a global probability of cancer for the entire breast, and would require validation across different vendors and facilities.

Clinical Efficiency

The shortage of radiologists subspecialized in breast imaging, combined with high volumes of screening examinations, has fueled interest in methods to increase efficiency while maintaining (or ideally improving) performance metrics (60). In one recent study, investigators developed a DL model to identify mammograms as cancer-free with high confidence, in order to improve workflow efficiency and improve performance (26). The DL model was trained and validated on a retrospective cohort of more than 235 000 mammograms. For the validation set, the model threshold was chosen to minimize the likelihood of a true-positive assessment, in order to maximize the mammograms triaged as true-negative examinations while maintaining high sensitivity for cancer detection. The DL model was subsequently tested on more than 26 000 mammograms and achieved an AUC of 0.82, with similar predictive accuracies across all age groups and races. A simulated workflow, in which radiologists only read mammograms not triaged as cancer-free by the DL model, demonstrated improved specificity (94.2% versus 93.5%, $P < 0.01$) and unchanged sensitivity (90.1% versus 90.6%) when compared to the usual workflow, in which radiologists interpreted all examinations.

Other studies have also demonstrated the potential of DL models to confidently identify mammograms as cancer-free and thus decrease radiologist workload (22,24). While these methods could decrease the number of cases that require interpretation by a radiologist, it remains to be seen whether this workload reduction would lead to less overall time spent on image interpretation (and thus allow more time for other tasks), or whether the radiologist would then devote more time to mammograms of higher complexity (61). Rapid and reliable identification of an examination as negative might also be useful in underserved communities, in which there is limited access to medical expertise (62).

Increased utilization of digital breast tomosynthesis since its approval by the Food and Drug Administration in 2011, coupled with longer times required to interpret these examinations when compared to digital two-dimensional (2D) mammography, has led to interest in methods that maximize reading efficiency while maintaining or improving performance metrics (63–66). However, the application of AI algorithms to tomosynthesis is limited by differences in the appearance of breast tissue with different vendors (which are larger than differences observed with digital 2D mammography) and relatively small training datasets (46). This second limitation can be mitigated by transfer learning, in which a pretrained model is fine-tuned with a new dataset; that is, the parameters of a model trained with a large dataset of digital 2D mammographic exams are copied, and the new model with the copied parameters is fine-tuned with the smaller tomosynthesis dataset (1,46). A recent reader study with tomosynthesis, in which 24 radiologists each interpreted 260 tomosynthesis examinations with and without an AI system, found that reading time decreased by an average of more than 30 seconds (from 64 seconds without AI to 30 seconds with AI) while improving AUC (0.80 to 0.85, $P < 0.01$), sensitivity, specificity, and the abnormal interpretation rate (19). The behavior of radiologists in this reader study may differ from actual clinical practice, and factors that could impact clinical practice include radiologists' level of confidence in their independent interpretations, their level of confidence in the performance of the AI system, the interpretability of the AI system (ie, the rationale being used to guide its predictions), and the user friendliness of the AI system (67).

Risk Stratification

Existing breast cancer risk prediction models are calibrated to provide risk estimates at the population level, but accurate risk assessment at the individual level is needed to inform decisions about screening regimens and prevention strategies (68). Breast density has been integrated into risk prediction models recently, but it is not likely to capture all of the rich information within a mammographic image (69,70). In one recent study, investigators developed a DL model to predict breast cancer risk by using cancer-free mammographic images and patient outcomes (ie, those who did and did not develop subsequent breast cancer) (20). The model was tested on cancer-free mammographic images from 2283 women, of whom 278 subsequently developed breast cancer. The model output was a score reflecting the likelihood of developing breast cancer. The correlation between the score and automated breast density

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

measures was low, indicating that the score was not simply a reflection of density. The DL model achieved a higher AUC than a model based on age and a breast density measure (0.65 versus 0.60, $P < 0.001$). The model also had significantly fewer false negatives than the best density model.

Other studies have also shown that image-based DL models may offer more accurate risk prediction than density-based models and existing clinical tools such as the Tyrer–Cuzick model (21,25). In one study, the best model to predict breast cancer risk within five years incorporated image-based information in addition to traditional risk factors (eg, family history) (AUC of 0.70) (25). Higher AUCs may not be possible for models that predict future disease (as opposed to models that predict current disease) (20). Future work could incorporate other sources of information to strengthen the models and could also shed light on the imaging patterns used by the DL models to predict risk, although such "black box" models may not ever be entirely understandable (21,71). Further research is needed to validate these risk prediction models across institutions and mammography vendors (70).

Artificial intelligence can also be used for risk stratification and clinical decision-making support in other domains of breast imaging. For example, the current management of high-risk breast lesions varies widely, and AI may be useful to guide clinical decision-making with regard to surveillance versus surgical excision (72). In one study, investigators developed an ML model using supervised learning to predict the risk of upgrade of high-risk lesions diagnosed by core needle biopsy to cancer at surgery (17). The random forest model, which was based on traditional features (eg, patient age and high-risk lesion histologic results) and text features from the biopsy pathology reports, was trained with 671 high-risk lesions and tested with 335 high-risk lesions. If the ML model were used to determine which high-risk lesions should be surgically excised versus surveilled, then 97.4% of cancers would have been diagnosed at surgery and 30.6% of benign surgeries could have been avoided. Of note, the one patient with cancer who was misclassified by the ML model had a history of Cowden syndrome, which is a feature that the model was not trained to recognize as an indicator of risk. Given this patient's increased risk of breast cancer, she likely would have undergone surgical excision regardless of the assessment made by the ML model. The incorporation of more varied clinical data (eg, the presence of genetic syndromes) and other sources of data (eg, mammographic images and pathology slides) could potentially strengthen the model and improve its predictive ability (73).

Challenges and Future Directions

One of the major challenges in AI algorithm development is the task of collecting and curating large datasets with appropriate labeling, which may require trained professionals (74–76). Furthermore, if the purpose of the model is to identify a rare disease, then it must be trained with a dataset of sufficient size to ensure exposure to various types and subtleties of disease presentation (76). Certain techniques (eg, transfer learning) can be used in cases of limited data, but these methods do not obviate the need for adequate representations of the disease of interest in the training dataset (1,32,76–78). A second challenge is the interpretability of AI systems that are developed, with uncertainty about acceptance by radiologists, other clinicians, and patients if there is no or limited human involvement and no accessible rationale about the decision-making process (46,75,76,79–83).

Prior to widespread deployment, AI systems must be validated in real-world clinical settings and across vendors and institutions (46,76,82,84). In addition, radiologists will require training to understand the appropriate use and limitations of each tool (76). These tools could be used in various clinical scenarios. For example, a triage approach would involve rapid identification of negative cases such that radiologists could spend more time on more complex examinations or other tasks, or the system would run in the background to identify cases that need to be evaluated more urgently; a replacement approach would use AI systems for stand-alone imaging interpretation; and, an add-on approach would provide decision support to radiologists or perform time-consuming tasks (1).

Conclusion

Artificial intelligence has incredible potential to improve the diagnostic accuracy and operational efficiency of breast imaging radiologists. Similar to other emerging technologies, however, AI systems require thorough evaluation in the clinical setting and on multiple, diverse imaging datasets before widespread adoption. Robust AI systems can complement our training, experience, and intelligence to improve accuracy and efficiency, and familiarity with the terminology, concepts, methods, and limitations of AI techniques is essential to ensure optimal patient care.

Acknowledgments

I thank Susanne L. Loomis, MS, FBCA, Radiology Educational Media Services, Massachusetts General Hospital, Boston, MA, for creating the figures in this manuscript.

Funding

This work was supported by the National Cancer Institute under the National Institutes of Health (K08CA241365). The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest Statement

None declared.

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

References

1. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *Radiographics* 2017;37(7):2113–2131.
2. Tang A, Tam R, Cadrin-Chênevert A, et al.; Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian Association of Radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J* 2018;69(2):120–135.
3. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. *AJR Am J Roentgenol* 2017;208(4):754–760.
4. Fuchsjäger M. Is the future of breast imaging with AI? *Eur Radiol* 2019;29(9):4822–4824.
5. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology* 2019;290(3):590–606.
6. Dreyer KJ, Geis JR. When machines think: radiology's next frontier. *Radiology* 2017;285(3):713–718.
7. Recht M, Bryan RN. Artificial intelligence: threat or boon to radiologists? *J Am Coll Radiol* 2017;14(11):1476–1480.
8. Brink JA. Artificial intelligence for operations: the untold story. *J Am Coll Radiol* 2018;15(3 Pt A):375–377.
9. Choy G, Khalilzadeh O, Michalski M, et al. Current applications and future impact of machine learning in radiology. *Radiology* 2018;288(2):318–328.
10. Tajmir SH, Alkasab TK. Toward augmented radiologists: changes in radiology education in the era of machine learning and artificial intelligence. *Acad Radiol* 2018;25(6):747–750.
11. Arieno A, Chan A, Destounis SV. A review of the role of augmented intelligence in breast imaging: from automated breast density assessment to risk stratification. *AJR Am J Roentgenol* 2019;212(2):259–270.
12. Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* 2019;292(2):331–342.
13. Mayo RC, Kent D, Sen LC, Kapoor M, Leung JWT, Watanabe AT. Reduction of false-positive markings on mammograms: a retrospective comparison study using an artificial intelligence-based CAD. *J Digit Imaging* 2019;32(4):618–624.
14. 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.
15. Rodríguez-Ruiz A, Lång K, Gubern-Merida A, et al. Standalone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019;111(9):916–922.
16. Watanabe AT, Lim V, Vu HX, et al. Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography. *J Digit Imaging* 2019;32(4):625–637.
17. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. *Radiology* 2018;286(3):810–818.
18. Shi B, Grimm LJ, Mazurowski MA, et al. Prediction of occult invasive disease in ductal carcinoma in situ using deep learning features. *J Am Coll Radiol* 2018;15(3 Pt B):527–534.
19. Conant EF, Toledano AY, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol Artif Intell* 2019;1(4):e180096.
20. Dembrower K, Liu Y, Azizpour H, et al. Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology* 2019;294(2):265–272.
21. Ha R, Chang P, Karcich J, et al. Convolutional neural network based breast cancer risk stratification using a mammographic dataset. *Acad Radiol* 2019;26(4):544–549.
22. Kyono T, Gilbert FJ, van der Schaar M. Improving workflow efficiency for mammography using machine learning. *J Am Coll Radiol* 2020;17(1 Pt A):56–63.
23. Lo Gullo R, Eskreis-Winkler S, Morris EA, Pinker K. Machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy. *Breast* 2020;49:115–122.
24. Rodríguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol* 2019;29(9):4825–4832.
25. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 2019;292(1):60–66.
26. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. *Radiology* 2019;293(1):38–46.
27. Zhou LQ, Wu XL, Huang SY, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology* 2020;294(1):19–28.
28. IBM Analytics. Data science and machine learning. Available at: https://www.ibm.com/analytics/machine-learning. Accessed 15 March 2020.
29. Giger ML. Machine learning in medical imaging. *J Am Coll Radiol* 2018;15(3 Pt B):512–520.
30. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018;18(8):500–510.
31. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp* 2018;2(1):35.
32. Do S, Song KD, Chung JW. Basics of deep learning: a radiologist's guide to understanding published radiology articles on deep learning. *Korean J Radiol* 2020;21(1):33–41.
33. Burt JR, Torosdagli N, Khosravan N, et al. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. *Br J Radiol* 2018;91(1089):20170545.
34. Syed AB, Zoga AC. Artificial intelligence in radiology: current technology and future directions. *Semin Musculoskelet Radiol* 2018;22(5):540–545.
35. Stone P. Reinforcement learning. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer, 2017.
36. Inteliment Technologies. Let's understand the difference between machine learning vs. deep learning. Available at: https://www.inteliment.com/blog/our-thinking/lets-understand-the-difference-between-machine-learning-vs-deep-learning/. Accessed 15 December 2019.
37. Robertson S, Azizpour H, Smith K, Hartman J. Digital image analysis in breast pathology – from image processing techniques to artificial intelligence. *Transl Res* 2018;194:19–35.

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

38. Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. *Clin Radiol* 2019;74(5):357–366.
39. Abdelhafiz D, Yang C, Ammar R, Nabavi S. Deep convolutional neural networks for mammography: advances, challenges and applications. *BMC Bioinformatics* 2019;20(Suppl 11):281.
40. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322(18):1806–1816.
41. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016;18(12):e323.
42. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286(3):800–809.
43. Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers – from the Radiology editorial board. *Radiology* 2019;294(3):487–489.
44. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29–36.
45. Ting KM. Confusion matrix. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer, 2017.
46. Geras KJ, Mann RM, Moy L. Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. *Radiology* 2019;293(2):246–259.
47. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging* 2020;39(4):1184–1194.
48. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007;356(14):1399–1409.
49. Cole EB, Zhang Z, Marques HS, Edward Hendrick R, Yaffe MJ, Pisano ED. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. *AJR Am J Roentgenol* 2014;203(4):909–916.
50. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL; Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175(11):1828–1837.
51. Fazal MI, Patel ME, Tye J, Gupta Y. The past, present and future role of artificial intelligence in imaging. *Eur J Radiol* 2018;105:246–250.
52. Katzen J, Dodelzon K. A review of computer aided detection in mammography. *Clin Imaging* 2018;52:305–309.
53. Gao Y, Geras KJ, Lewin AA, Moy L. New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. *AJR Am J Roentgenol* 2019;212(2):300–307.
54. Sechopoulos I, Mann RM. Stand-alone artificial intelligence – the future of breast cancer screening? *Breast* 2020;49:254–260.
55. Bahl M. Detecting breast cancers with mammography: will AI succeed where traditional CAD failed? *Radiology* 2019;290(2):315–316.
56. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digital Health* 2020;2(3):E138–E148.
57. Dustler M. Evaluating AI in breast cancer screening: a complex task. *Lancet Digital Health* 2020;2(3):E106–E107.
58. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577(7788):89–94.
59. Pisano ED. AI shows promise for breast cancer screening. *Nature* 2020;577(7788):35–36.
60. Wing P, Langelier MH. Workforce shortages in breast imaging: impact on mammography utilization. *AJR Am J Roentgenol* 2009;192(2):370–378.
61. Kontos D, Conant EF. Can AI help make screening mammography "lean"? *Radiology* 2019;293(1):47–48.
62. Mayo RC, Leung J. Artificial intelligence and deep learning – radiology's next frontier? *Clin Imaging* 2018;49:87–88.
63. Bernardi D, Ciatto S, Pellegrini M, et al. Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time. *Br J Radiol* 2012;85(1020):e1174–e1178.
64. Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology* 2013;267(1):47–56.
65. Dang PA, Freer PE, Humphrey KL, Halpern EF, Rafferty EA. Addition of tomosynthesis to conventional digital mammography: effect on image interpretation time of screening examinations. *Radiology* 2014;270(1):49–56.
66. Hooley RJ, Durand MA, Philpotts LE. Advances in digital breast tomosynthesis. *AJR Am J Roentgenol* 2017;208(2):256–266.
67. Hsu W, Hoyt AC. Using time as a measure of impact for AI systems: implications in breast screening. *Radiol Artif Intell* 2019;1(4):e190107.
68. Amir E, Freedman OC, Seruga B, Evans DG. Assessing women at high risk of breast cancer: a review of risk assessment models. *J Natl Cancer Inst* 2010;102(10):680–691.
69. Brentnall AR, Cuzick J, Buist DSM, Bowles EJA. Long-term accuracy of breast cancer risk assessment combining classic risk factors and breast density. *JAMA Oncol* 2018;4(9):e180174.
70. Bahl M. Harnessing the power of deep learning to assess breast cancer risk. *Radiology* 2020;294(2):273-274.
71. Sitek A, Wolfe JM. Assessing cancer risk from mammograms: deep learning is superior to conventional risk models. *Radiology* 2019;292(1):67–68.
72. Falomo E, Adejumo C, Carson KA, Harvey S, Mullen L, Myers K. Variability in the management recommendations given for high-risk breast lesions detected on image-guided core needle biopsy at U.S. academic institutions. *Curr Probl Diagn Radiol* 2019;48(5):462–466.
73. Shaffer K. Can machine learning be used to generate a model to improve management of high-risk breast lesions? *Radiology* 2018;286(3):819–821.
74. Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol* 2018;15(3 Pt B):504–508.
75. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin* 2019;69(2):127–157.
76. Chan HP, Samala RK, Hadjiiski LM. CAD and AI for breast cancer-recent development and challenges. *Br J Radiol* 2020;93(1108):20190580.

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022

77. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88.
78. Akkus Z, Cai J, Boonrod A, et al. A survey of deep-learning applications in ultrasound: artificial intelligence-powered ultrasound for improving clinical workflow. *J Am Coll Radiol* 2019;16(9 Pt B):1318–1328.
79. Liew C. The future of radiology augmented with artificial intelligence: a strategy for success. *Eur J Radiol* 2018;102:152–156.
80. SFR-IA Group, CERF, French Radiology Community. Artificial intelligence and medical imaging 2018: French Radiology Community white paper. *Diagn Interv Imaging* 2018;99(11):727–742.
81. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Radiology* 2019;293(2):436–440.
82. Lee CI, Elmore JG. Artificial intelligence for breast cancer imaging: the new frontier? *J Natl Cancer Inst* 2019;111(9): 875–876.
83. Mendelson EB. Artificial intelligence in breast imaging: potentials and limitations. *AJR Am J Roentgenol* 2019;212(2):293–299.
84. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices* 2019;16(5):351–362.

Downloaded from https://academic.oup.com/jbi/article/2/4/304/5859939 by guest on 21 January 2022