Taylor & Francis
Taylor & Francis Group

Expert Review of Medical Devices

ISSN: 1743-4440 (Print) 1745-2422 (Online) Journal homepage: https://www.tandfonline.com/loi/ierd20

Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice

Nehmat Houssami, Georgia Kirkpatrick-Jones, Naomi Noguchi & Christoph I. Lee

To cite this article: Nehmat Houssami, Georgia Kirkpatrick-Jones, Naomi Noguchi & Christoph I. Lee (2019) Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice, Expert Review of Medical Devices, 16:5, 351-362, DOI: 10.1080/17434440.2019.1610387

To link to this article: https://doi.org/10.1080/17434440.2019.1610387

Accepted author version posted online: 18 Apr 2019.
Published online: 03 May 2019.

EXPERT REVIEW OF MEDICAL DEVICES
2019, VOL. 16, NO. 5, 351–362
https://doi.org/10.1080/17434440.2019.1610387

REVIEW

OPEN ACCESS

Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice

Nehmat Houssami[a], Georgia Kirkpatrick-Jones[a], Naomi Noguchi[a] and Christoph I. Lee[b,c,d]

[a]The University of Sydney, Faculty of Medicine and Health, Sydney School of Public Health (A27), Sydney, Australia; [b]Department of Radiology, University of Washington School of Medicine, Seattle, WA, USA; [c]Department of Health Services, University of Washington School of Public Health, Seattle, WA, USA; [d]Hutchinson Institute for Cancer Outcomes Research, Seattle, WA, USA

ABSTRACT

Introduction: Various factors are driving interest in the application of artificial intelligence (AI) for breast cancer (BC) detection, but it is unclear whether the evidence warrants large-scale use in population-based screening.

Areas covered: We performed a scoping review, a structured evidence synthesis describing a broad research field, to summarize knowledge on AI evaluated for BC detection and to assess AI's readiness for adoption in BC screening. Studies were predominantly small retrospective studies based on highly selected image datasets that contained a high proportion of cancers (median BC proportion in datasets 26.5%), and used heterogeneous techniques to develop AI models; the range of estimated AUC (area under ROC curve) for AI models was 69.2–97.8% (median AUC 88.2%). We identified various methodologic limitations including use of non-representative imaging data for model training, limited validation in external datasets, potential bias in training data, and few comparative data for AI versus radiologists' interpretation of mammography screening.

Expert opinion: Although contemporary AI models have reported generally good accuracy for BC detection, methodological concerns, and evidence gaps exist that limit translation into clinical BC screening settings. These should be addressed in parallel to advancing AI techniques to render AI transferable to large-scale population-based screening.

ARTICLE HISTORY
Received 4 February 2019
Accepted 18 April 2019

KEYWORDS
Artificial intelligence; breast cancer; data bias; mammography; population screening

1. Introduction and aims

Intelligent computer systems have existed and have made a mark in society for several decades. Interest in artificial intelligence (AI) research and development spans the technology, communication, industry, health, and government including security and defense sectors [1]. At present, the convergence of novel AI techniques, massive computer processing capabilities, and widespread growth of digital capture and storage of data in general and specifically in science and health, is transforming the application of AI in diverse areas. In cancer, as in other areas of healthcare, AI systems are being developed, explored and evaluated for disease detection, prognostication and as support strategies for clinical decision-making.

In the context of breast cancer, ongoing research using AI for early detection includes a global effort attempting to develop advanced machine learning algorithms for interpreting screening mammograms to potentially improve breast cancer screening by reducing false-positives [2,3]. The potential application of AI in breast cancer diagnostics extends to imaging modalities and also pathology interpretation, for example, AI has been shown to augment identification of metastatic breast cancer in whole-slide images of sentinel lymph node biopsy [4]. We focus on early detection of breast cancer in this work to gauge the potential role of contemporary AI systems in screening practice.

We performed a scoping review, a form of structured evidence synthesis (similar to a systematic review) describing a broad research field, with the aims of (a) identifying and summarizing current knowledge on the application of AI in the early detection of (screening for) breast cancer; (b) mapping key evidence concepts in the application of AI in breast screening, specifically whether AI has been evaluated as a stand-alone screening strategy or as a complement (aid) to screen-reading (that is, an aid to human interpretation of mammograms) to determine transferability to the screening context; and (c) defining gaps in the available evidence to highlight areas meriting more research. The scoping review *did not aim to assess technical or statistical aspects* in the development of AI models and strategies, rather it focused on the evidence in applied research using AI techniques to determine readiness for real-world breast screening practice or screening trials, and to inform future research in the AI space as it relates to breast cancer screening.

2. Methods

We performed a scoping review to assess and summarize, in a structured manner, the evidence on the use of AI in breast

CONTACT Nehmat Houssami nehmat.houssami@sydney.edu.au The University of Sydney, Faculty of Medicine and Health, Sydney School of Public Health (A27), Edward Ford Building, NSW 2006, Australia

Article highlights

- This scoping review, a form of structured evidence synthesis describing a broad research field, summarizes knowledge from 23 studies that evaluated artificial intelligence (AI) for automated BC detection
- Majority of studies were small, retrospective studies that trained and tested AI models using cancer-enriched image datasets (median proportion cancer-positive 26.5%)
- AI techniques were heterogeneous, but a predominance of models was developed using convolutional neural networks (CNN); most studies validated developed models (frequently using cross-validation) but few tested the model in an independent dataset
- A consistently reported measure of accuracy for the AI models was the area under the receiver-operating characteristic curve (AUC): estimated AUC was 69.2–97.8% across studies
- Methodological concerns include substantial uncertainty regarding the quality of the imaging data used to train AI models in terms of limited applicability (external validity) of developed models
- There was potential for bias due to use of unbalanced imaging data (that does not represent the spectrum in real-world screening) to train and test models; hence, algorithms may not perform well when applied or tested in actual screening practice
- There were limited comparative data on AI versus human interpretation of breast screening examinations
- Current evidence is limited to AI algorithm development in digital (2D) mammography; none of the studies used digital breast tomosynthesis (3D-mammography) to train or test models
- We identify current gaps in the evidence including the need for large prospective studies that develop and test AI using real-world screening data and more efforts in the clinical translation of AI systems into routine breast cancer screening practice.

cancer detection. We anticipated a range of study designs exploring various AI methods in different applied contexts in breast cancer detection, we therefore undertook a scoping review to address this broad research area – given the heterogeneity of research in this field, conventional data synthesis using standard systematic reviews or meta-analysis would not be appropriate [5]. Scoping reviews allow evidence mapping and synthesis from a variety of studies and sources to address broad research questions and to identify evidence gaps [5,6]. To develop the methods of the scoping review, we considered a framework and recommendations on scoping review methodology [5–7] as well as a reporting checklist (an extension of PRISMA) specific to scoping reviews (PRISMA ScR) [8].

2.1. Literature search and eligible studies

A literature search was conducted (2010–2018) as shown in Appendices 1–2; the search timeframe was chosen to factor advances in AI methods and capabilities. The review focused on summarizing the evidence on the application of AI in breast cancer detection (screening) without study design restriction. Studies were eligible for inclusion in our review on the basis of the following criteria: (a) the purpose of the study was to assess an AI approach or strategy in breast cancer screening or detection; (b) reported quantitative data on performance (accuracy) or screening or clinical outcome measures for the AI approach relative to a reference standard and/or an established comparator (for example, an ascertained database or radiologists' interpretation); (c) undertook the evaluation in women or screening examinations from women *without* being restricted solely to women with breast cancer or to those who have had tissue biopsy. Studies were not eligible: if they evaluated AI in phantom (simulated) lesions or in simulation models, or if they described AI techniques or compared data-mining algorithms (or dealt with the development thereof) without application as described in the inclusion criteria; if they did not provide information on the number of subjects or screens or images included, or if they were based on fewer than 100 subjects (or fewer than 200 images if multiple images were used from an undeclared number of subjects) as this would not yield reliable information for testing of AI strategies in the context of breast screening. Commentary or editorial articles, review articles, and congress abstracts were not eligible for inclusion.

Literature searching and abstract screening to identify potentially eligible studies were performed by one investigator (NH): selection of eligible studies based on the above-defined criteria is shown in the flow-diagram (Appendix 1).

2.2. Data extraction and collation

Study-specific information and data were extracted into an evidence table to summarize the following: purpose or aims of the study, design and methods (including amount and type of data), source population or subjects (and whether study included consecutive screens or subjects), reference standard and/or comparator (if any), class of AI technique, validation (if done), and the main findings reported on accuracy, or screening or clinical outcomes (where reported). Formal quality appraisal is not routinely done in scoping reviews; however, methodological variables were considered in the extracted information to provide an understanding of the quality of the evidence. Extraction of information from eligible studies was based on independent double-extraction (two investigators from NN, GKJ, and NH) using a pre-defined extraction form; discussion and consensus were used to cross-check the extracted information and to resolve disagreement.

The information collated in the evidence tables was used to define the main themes of research and to elucidate the extent that published evidence to date transfers to breast cancer screening. Descriptive statistics (median, range) were used to summarize quantitative information where these were reported by a majority of the eligible studies; where studies reported several estimates for accuracy measures, the median of the reported range was used for that study.

3. Results

There were 23 eligible studies [9–31] based on the literature search strategy (additional details and excluded studies [32–47] shown in Appendix 1). A summary of the eligible studies, including study characteristics, is shown in Table 1; and study findings are reported in Table 2. There were no prospective screening trials or randomized studies. Studies were predominantly retrospective using publicly available or institutional image datasets (Table 1), and the same datasets (or selected subsets thereof) were often used in several studies; however, Parmeggiani et al. [28] reported a prospective cohort study from an institutional screening program, and Ayer et al. [31] reported a retrospective study using a large

Table 1. Summary of the characteristics of studies reporting on artificial intelligence (AI) in breast cancer detection.

Study first author(s)	Purpose or aim (s)	Study design	Data Source for AI development or evaluation	Type of data (input variables where stated)	Number of subjects or images (N)	Subjects' mean or median age	Breast cancer proportion in study	Included consecutive subjects or screens?	Reference Standard
Rodriguez-Ruiz [9]	To compare stand-alone performance of a commercially available AI system to that of radiologists in detecting BC on DM	Retrospective (image databases)	DM exams collected during previous reader studies	DM	2652 DMs	Various ranges across datasets	24.6%	No	Datasets verified with pathology or ≥1 year follow-up
Al-Masni [10]	Describe and evaluate a CAD system based on ROI (region of interest) deep learning techniques, using a CNN referred to as You Only Look Once (YOLO), to detect and classify breast masses on DM.	Retrospective (image database)	DDSM	DM (lesion feature, mammographic mapping)	600 DM from 150 women	NR	50.0%	No	NS, sourced classified database
Ribli [11]	Propose a CAD system based on deep CNNs to detect and classify lesions on mammograms	Retrospective (image database)	DDSM, Inbreast image datasets plus institutional images	Digital and digitized film mammograms	2949 mammograms	NR	NR	No	NS, several image databases
Chougrad [12]	To develop a CAD system by refining existing models using CNN, exploring a transfer learning method, to help classify mammography mass lesions.	Retrospective (image databases)	Merged *subsets* from several image databases: DDSM, BC Digital Repository, and Inbreast databases.	Digital, including digitized film, mammograms (pixel information)	6116 images from 1529 subjects (selected benign or cancer cases, excluded normal)	NR	51.0%	No	NS, sourced classified database (many biopsy-proven)
Bandeira-Diniz [13]	To develop a model to classify breasts on mammography into dense and non-dense breasts, and to develop another model to classify regions of the breast into mass and non-mass regions based on symmetry in dense and non-dense breasts separately using CNN.	Retrospective (imaging database)	DDSM	DM (pixel information, some demographic information)	2,482 images from 1241 women (80% used for model training/ 20% for testing)	NR	NR	No (only images with at least one mass were chosen)	NS, sourced classified database.
Becker [14]	To evaluate the diagnostic accuracy of a multipurpose image analysis software based on deep learning with artificial neural networks for the detection of BC in an independent, dual-centre mammography data set.	Retrospective (image database)	Mammograms from an institution during one year; and BC Digital Repository data set	DM (lesion feature, mammographic mapping)	1146 subjects (images NR)	59.6	7.0%	No	Histology or clinical follow-up
Lotter [15]	To develop a CAD system using CNN to classify whether BC is present on the mammogram	Retrospective (image database)	DDSM	DM	10,480 images from 2620 women	NR	NR	No	Sourced classified database, includes pathology outcomes
Becker [16]	To train a generic deep learning software to classify breast cancer on ultrasound images, and to compare its performance to human readers with variable breast image-reading experience.	Retrospective (examinations done in one year)	Still images from ultrasound examinations in one institution (not real-time ultrasound)	Ultrasound images	632 patients (70% used for training/30% for validation)	53 (SD: 15)	13.0%	No (18% of initial cohort included); those with normal US or benign lesions excluded	Histology or 24-month clinical follow-up
De Oliviera Silva [17]	To describe and test use of independent component analysis (compared with principal component analysis) in CAD for detecting lesions in dense breasts	Retrospective (image database)	Selected images from DDSM and Mini-MIAS	DM (pixel information)	216 images from 322 patients	NR	NR	No	NS, sourced classified database
Kooi [18]	To develop a model using deep CNN to detect malignant lesions on mammograms and to compare this to a mammography CAD system (based on manually designed features); to assess to what extent clinical information (location, context features, and age) improves model accuracy; and to compare its performance against human readers.	Retrospective (imaging database)	A large-scale screening program in the Netherlands.	DM (pixel information and some clinical information)	40,506 images from 6729 subjects (training), 4303 images from 745 subjects (validation), 18,453 images from 2188 subjects (test set)	NR (women >50 years)	4.7%	NR	Biopsy for cancers, or follow-up
Samala [19]	To translate knowledge learned from non-medical images to medical diagnostic tasks using a multi-task transfer learning DCNN through supervised training of DCNNs (comparing multi-task approach to single-task transfer learning classification methods), and apply this approach in CAD of breast cancer	Retrospective (image database)	DDSM, and mammograms from University of Michigan Health System	Digitized screen-film mammograms (SFM) and DMs	2,242 images (containing 1,057 malignant, 1,397 benign findings) (number of subjects NR)	51.7	43.0% (lesion-based estimate)	No	NS, sourced classified database

(*Continued*)

Table 1. (Continued).

Study first author(s)	Purpose or aim (s)	Study design	Data Source for AI development or evaluation	Type of data (input variables where stated)	Number of subjects or images (N)	Subjects' mean or median age	Breast cancer proportion in study	Included consecutive subjects or screens?	Reference Standard
Carneiro [20]	To develop a model using deep CNN to assess whether unregistered mammographic images (cranio-caudal and medio-lateral oblique views from each breast) can be classified as containing malignant lesions, benign lesions or only normal tissue.	Retrospective (imaging databases)	Inbreast, DDSM and ImageNet (ImageNet used for model training) databases	DM (BI-RADS class plus lesion delineation (shape and features))	410 + 680 images (Inbreast + DDSM) from 287 subjects (ImageNet has non-medical imaging data)	NR	NR	NR	NS, sourced classified databases
Teare [21]	To present a machine-learning based algorithm which utilizes novel techniques (deep CNN) plus false colour-enhancement technique) to detect malignant lesions in digital mammographic images, and to achieve accuracy similar to that of expert radiologists	Retrospective (imaging database)	DDSM and Zebra Mammography Dataset (ZMDS)	DM	761 images from 586 women	NR	32.0%	NR#	Histology or 2-year imaging follow-up
Dhungel [22]	To present an integrated methodology for detecting, segmenting and classifying breast masses from mammograms with minimal user intervention.	Retrospective (image database)	Inbreast dataset	DM	410 images from 115 subjects	NR	28.3%	NR#	NS, sourced classified database
Sun [23]	To propose a semi-supervised learning (SSL) scheme using deep CNN for BC diagnosis, that only requires a small portion of labelled data in training set rather than a large amount of labelled data for training and fine-tuning.	Retrospective (image database)	In-house full-field digital mammography image database	DM (lesion features)	1874 paired mammographic images (subjects NR)	51	45.0% (lesion-based estimate)	NR#	NR
Saraswathi & Srinivasan [24]	To evaluate a fully complex-valued relaxation neural network (FCRN) based system to identify normal, benign and malignant lesions in digital mammographic images, to improve classification accuracy.	Retrospective (imaging database)	MIAS	DM	322 images (number of subjects NR)	NR	NR; 55% (validation sample)	NR#	NS, sourced classified database
Velikova [25]	To obtain a balanced view on the role and place of expert knowledge and learning methods in building Bayesian networks for medical image interpretation, using interpretation of mammograms as the example	Retrospective (image database)	Imaging data from the Dutch breast cancer screening program	DM	795 subjects (images NR)	NR	43.3%	NR#	Cancers verified by pathology reports
Dheeba & Tamil-Selvi [26]	Propose a supervised machine learning algorithm (DEOWNN) for automatic detection of cancerous masses in mammograms.	Retrospective (image database)	MIAS	DM (texture features)	322 images (paired breast images) from 161 subjects	NR	16.0%	NR#	NS, sourced classified database
Dheeba & Tamil-Selvi [27]	To develop a CAD system to detect microcalcification clusters in digital mammograms using Swarm Optimization Neural Network (SONN).	Retrospective (imaging database)	MIAS; clinical mammogram images for validation	DM (texture features)	322 images (paired breast images) from 161 subjects; (validation: 216 images from 54 subjects)	NR	16.0%	NR#	NS, sourced classified database
Parmeggiani [28]	To test an Artificial Neural Network (ANN) system developed to detect BC in mammographic and echographic images on a cohort of women being screened for BC, to potentially develop autonomous AI systems.	Prospective cohort	A screening program in a university-affiliated hospital in Naples, Italy.	Mammograms and ultrasound scans	550 subjects (Images NR)	NR (> 40 years)	13.30%	NR#	Surgery or follow-up
Lesniak [29]	To reduce false positives in mammographic breast cancer screening by using computer aided detection with a support vector machine (SVM) based system	Retrospective (image database)	A BC screening database	Scanned film mammograms (lesion descriptors)	10,064 images from 1539 subjects	NR	32%	NR#	sourced classified database (cancers biopsy-proven)
Huang [30]	To compare the performance of three different hybrid algorithms, particle swarm optimizer (PSO)-based artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS) and a case-based reasoning (CBR) classifier (latter with a logistic regression (LR) or a decision tree (DT) model), in BC diagnosis.	Retrospective (image database)	Machine Learning Repository Mammographic Mass Data Set	DM (BI-RADS score, lesion descriptors), age	815 subjects (images NR)	NR (range 18–96)	46% (lesion-based estimate)	No	NS, sourced classified database
Ayer [31]	To assess whether an artificial neural network (ANN) trained on a large dataset of consecutive screening and diagnostic mammograms can discriminate between benign and malignant disease, and accurately predict the probability of BC	Retrospective, using prospectively collected and interpreted mammogram database	All mammograms performed at Medical College of Wisconsin Breast Care Centre (1999–2004)	Mammographic features (BI-RADS descriptors), routinely collected demographic risk factors	48,744 mammograms from 18,269 subjects	56.5	0.80%	Yes (consecutive mammographic exams)	Biopsy or ascertained by matching with State Cancer Reporting System

BC (breast cancer); CAD (computer-aided diagnosis); CNN (Convolutional Neural Network); DM (digital mammograms); DCNN (Deep Convolutional Neural Network); MIAS (Mammographic Image Analysis Society database); DDSM (Digital Database for Screening Mammography); NR (not reported); NS (not specified).

#Not reported however not likely to have included consecutive subjects or screens based on other information regarding the data source or the proportion with cancer.

Table 2. Summary of the findings of studies reporting on artificial intelligence (AI) in breast cancer detection.

First author	Category of AI	Model validation (specify type if reported)	Was there external validation using an independent dataset?	Area under the curve (AUC) %	Accuracy (other than AUC) %	Sensitivity%	Specificity%
Rodriguez-Ruiz [9]	Deep learning CNNs, feature classifiers, and image analysis algorithms	AI system trained and validated in earlier work	Present study represents a validation in an independent dataset	84.0%[vs average for 101 radiologists = 81.4%, statistically non-inferior based on difference in AUC of 2.6% (95%CI: −0.3 to 5.5)]*	NR	75%-85% across various datasets at radiologists' specificity	NR
Al-Masni [10]	CNN, FC-NN	5-fold cross validation	No, however a modified (augmented image data) version used for testing	87.7%	85.5% (mass location 99.7%)	93.2%	78.0%
Ribli [11]	Faster R-CNN (based on a CNN with additional components)	NR	Participated in the DREAM‡ challenge which was based on an independent dataset (AUC 85%)	95%	NR	90%	NR
Chougrad [12]	Deep CNN	Stratified 5-fold cross validation	yes, however based on selected benign/cancer cases from the MIAS.	Range of estimates using various datasets: 96.0–99.0%	Range of estimates using various datasets: 95.5–98.2%	NR	NR
Bandeira-Diniz [13]	Deep CNN	Not done	No, however subset of 20% of database images (not used for training) saved for testing developed model	NR	Mass/non-mass region classification 91.0% in non-dense breasts, 94.8% in dense breasts.	Mass/non-mass region classification: 91.5% in non-dense, and 90.4% in dense breasts.	Mass/non-mass region classification 90.7% in non-dense, 96.4% in dense breasts
Becker [14]	ANN	Not done	Yes, tested using an independent image dataset	81–85% (training data) [vs radiologists 83–94%]*; 79%-81% (test data) [vs radiologists 77–87%]*	NR	59.8% (training data); 71.6%,73.7% (study 1, 2)	84.4% (training data); 69.6%, 72.0% (study 1, 2)
Lotter [15]	Multi-scale CNN	Cross-validation by patient (8% of data)	No, however subset of 5% of database used to validate the developed model	92%	NR	NR	NR
Becker [16]	Deep CNN	Validated	No, however subset of 30% of images (not used for training) saved for model validation	96% (training set); 84% (validation set) [vs experienced/intermediate human readers 89% (P < 0.05), or inexperienced readers 79% (P < 0.05)]*	NR	84.2% [experienced/intermediate readers 84.2%]	80.4% [experienced/intermediate readers 89/83%]
De Oliviera Silva [17]	Independent component analysis	NR	No	NR	92.7% in non-dense, 79.2% in dense breasts	95.7% in non-dense breasts, 66.7% in dense breasts	90.0% in non-dense breasts, 91.7% in dense breasts.
Kooi [18]	Deep CNN	8-fold cross validation	No, however subset of images (not used for training) used for testing developed model	CNN 92.9% (CAD system 91%); CNN with other variables 94.1%; test set: CNN 85.2% [vs radiologists 91.1% (p = 0.001 for mean AUC]*	–	NR	NR
Samala [19]	Deep CNN	4-fold cross-validation	Yes, independent test set (multi-task DCNN)	78% (single-task transfer learning), 82% (multi-task transfer learning), 76% (lesion-based), 79% (view-based)	NR	NR	NR
Carneiro [20]	Deep CNN	5-fold cross validation (Inbreast dataset)	No	90% (semi-automated approach), 86% (fully automated approach)	NR	69.0–94.0%	66.0–92.0%
Teare [21]	Dual deep CNN	Validated	No	92.2%	84%	91%	80.4%
Dhungel [22]	CNN, RF, BO	5-fold cross validation	No, however subset of images (not used for training) used for testing model	69–76% (minimal user intervention), 80–91% (manual set-up)	84–95% (range for various methods)	98%	70%
Sun [23]	Deep CNN	No	No	88.2% (using mixed labelled and unlabelled data)	82.4% (using mixed labelled and unlabelled data)	79.2–81.0% (for different data weighing functions)	69.4–72.3% (for different data weighing functions)
Saraswathi &	Srinivasan [24]	FCRN	10-fold cross validation	No	67.9% (without cross-validation), 94.7% (with cross-validation)	–	90% (cross-validation)
91% (cross-	validation)						
Velikova [25]	BN	2-fold cross validation	No	62.8–75.5% (range for models learnt with discretised or continuous data)	NR	NR	NR
Dheeba & Tamil-Selvi [26]	DEOWNN	Not done	No	97.80%	NR	96.9%	92.9%
Dheeba & Tamil-Selvi [27]	SONN	Not done	Yes, 216 clinical mammogram images (54 patients) from screening centres	97.6% (for detecting microcalcification clusters); validation 91.3%	NR	NR	NR
Parmeggiani [28]	ANN	Not done	Yes – the ANN was pre-trained using external data and was validated in the study in a different population	NR	NR	80% [73.6–74.2% radiologists alone]*; 88.5% for AI and readers combined	70.9% [54–57.4% radiologists alone]*; 69.5% for AI and readers combined

(*Continued*)

Table 2. (Continued).

First author	Category of AI	Model validation (specify type if reported)	Was there external validation using an independent dataset?	Area under the curve (AUC) %	Accuracy (other than AUC) %	Sensitivity%	Specificity%
Lesniak [29]	SVM, RF and CNN	10-fold cross validation.	No, however independent dataset used for normalisation of data	NR	NR	Mean true positive fraction 68.6 (higher than other CAD systems)	NR
Huang [30]	PSO, ANN, ANFIS, CBR with DT	10-fold cross validation	No, subset of images (not used for training) used for testing models	91.1% (PSO-based ANN), 92.8% (ANFIS), 83.6% (CBR-DT), 79.9% (CBR-LR)	NR	NR	NR
Ayer [31]	ANN	10-fold cross validation	No, subset of images from same dataset not used for training used for testing	96.5 [vs radiologists 93.9 (P < 0.001) aggregate level analysis] *	NR	90.7 [vs radiologists 82.2 (P < 0.001) aggregate level analysis]*	(sensitivity shown at a fixed specificity of 90%)

*Indicates results for radiologists in squared brackets.

‡ DREAM refers to the Digital Mammography challenge (information via https://www.synapse.org/Digital_Mammography_DREAM_Challenge)

CNN (Convolutional Neural Network); SNN (Spiking Neural Network); BN (Bayesian Networks); FC-NN (Fourier Convolutional Neural Networks); ANN (Artificial Neural Network); DCNN (Deep Convolutional Neural Network); SONN (Swarm Optimization Neural Network); SVM (Support Vector Machine); DEOWNN (Differential Evolution Optimized Wavelet Neural Network); FCRN (Fully complex-valued relaxation neural network); PSO (Particle Swarm Optimizer); ANFIS (Adaptive Neuro-Fuzzy Inference System); MIAS (Mammographic Image Analysis Society database); CBR (Case-Based Reasoning); CAD (computer-aided diagnosis); BO (Bayesian optimisation); DT (Decision Trees); RF (Random Forests).

prospectively collected well-defined mammographic database. As shown in Table 1, studies were generally based on relatively modest numbers of images (and hence smaller number of subjects), except for each of the studies from Kooi et al. and Ayer et al. [18,31] which investigated AI systems using relatively large datasets (>40,000 images, or mammographic examinations from >9,000 women). Most studies provided limited information on the methods used to assemble the source imaging datasets and the extent that these were verified in terms of a reference standard, with many studies simply citing the source image dataset [10,12,13,19,20,22–24,26,27,30]. However, several studies described an appropriate reference standard that included histopathology with either clinical follow-up or cancer registry matching to ascertain outcomes [9,14,16,18,21,28,31].

Studies proposed to develop and/or evaluate AI models or techniques for breast cancer detection [9,11,18,21,22,27,28,26], or for diagnosis (classification) or interpretation of mammographic examinations [13,14,15,16,20,23–25,30], or dealt with advancing computer-aided detection (CAD) systems through new AI models [10,12,17,19,29]; and one study investigated AI for discrimination between benign and cancerous lesions jointly with cancer risk prediction [31]. Rodriguez-Ruiz et al. [9] reported a multi-reader study comparing an AI system with radiologists' interpretation of various datasets of screening and clinical mammographic examinations. All studies were based on mammographic images except for the studies from Becker (which used ultrasound scans) [16] and from Parmegianni (which combined ultrasound and mammography screening) [28].

The reported breast cancer proportion across studies ranged between 0.80% and 55.0% for studies reporting this variable, with a median cancer proportion of 26.5% [9,10,12,14,16,18,19,21–31]. With the exception of one study from Ayer et al. [31], studies did not include consecutive screens or subjects (with many reporting selection of cases with abnormalities), or did not report any information on whether consecutive screens were included or the extent of exclusions. This is commensurate with the generally high proportion of cancers described for the datasets used to develop AI models across studies (Table 1) with the exception of the work from Ayer et al. [31].

A brief summary of the AI methods (type of AI and validation) and study-specific results are shown in Table 2. The AI techniques were heterogeneous but there was a predominance of models that were primarily developed using convolutional neural networks (CNN), and AI models generally achieved good accuracy (Table 2). Most of the studies incorporated a validation process (frequently cross-validation) when training AI models or reported results of model testing, generally using subsets of images that were not used for training or by augmenting (modifying) the image datasets to allow testing of the developed model. However, few studies undertook an external validation of the developed AI model using an independent dataset (study-specific details shown in Table 2).

We did not identify any studies that reported clinical outcomes or conventional breast cancer screening metrics (such as cancer detection rates or recall rates). The most consistently reported measure of accuracy for the AI models (Table 2) was the area under the receiver-operating characteristic (ROC) curve, a global measure of accuracy that incorporates the trade-off between sensitivity and specificity: the AUC across studies ranged between 69.2% and 97.8% [9–12,14–16,18–27,30,31], with a median AUC of 88.2%. Several studies reported a range of estimates depending on the techniques used within the study or on whether the training or validation data (or both) were reported (Table 2). Other study-specific results (accuracy, sensitivity, specificity) that were not consistently reported by most studies are also shown in Table 2. Very few studies reported comparisons for AI and human readers: the five studies that did so [9,14,16,18,31] showed mixed findings for AUC, sensitivity, and specificity – study-specific results are shown in Table 2.

4. Discussion

We report a scoping review, a form of structured evidence collation used to address a broad research question, to assess the evidence on AI systems evaluated for breast cancer detection (published since 2010) to gauge AI's potential role in breast screening. We specifically looked for evidence on how the AI models performed and whether there was data comparing their accuracy to human readers in a breast screening context. Available studies indicate a potential role for AI in this clinical scenario, however, there are evidence gaps relevant to future evaluation and application of AI in breast cancer screening.

We found that the published evidence on AI for breast cancer detection was concentrated around model (algorithmic) development, generally independently of real-world clinical or screening evaluation, and overall the evidence does not indicate the readiness of AI systems for real-world breast screening trials or for stand-alone screen-reading. We arrived at that conclusion despite encouraging results for the performance of the AI models, highlighted in the range of reported model AUCs (69.2% to 97.8% across studies, median AUC 88.2%), because there are key evidence gaps that need to be addressed before AI can be rendered more transferable to large-scale screening evaluations. Our conclusion takes into consideration both the rationale for undertaking the scoping review and the methodological concerns we have identified through our work (described in the remainder of the Discussion) that are relevant for future studies in this field.

Several factors prompted us to undertake this work: first, there are large-scale projects developing AI for breast cancer screening [3]; second, the data and statistical sciences driving AI development have advanced substantially in recent years, as has digital imaging data capture and archiving; third, mammography, the only imaging modality to date shown to reduce breast cancer mortality, has evolved into digital breast tomosynthesis (DBT or 3D mammography) technology which contains richer imaging data than conventional mammography; and fourth, the increasing burden of resourcing screen-reading in population-based screening programs that practice double-reading of mammography. In combination, these factors steer a rationale for AI as a candidate technology for future breast screening practice. Hence, we sought to assess the published literature to gauge the readiness of recently investigated AI systems for breast screening application and to inform research directions in this field. We identified several concerns relating to the quality, depth, and representativeness of imaging datasets used to train models, as well as limited comparative data (AI versus human readers), that affect both the applicability and robustness of developed models and raise the possibility of bias. These issues merit attention from researchers developing and evaluating AI models and systems with the intention of deploying these in breast cancer screening practice. We also note that only one of the studies included in this scoping review evaluated a commercially available AI system in a reader study format [9], and there are no prospective evaluations reported in clinical practice settings. This suggests that real-world implementation studies of AI in breast screening may be lagging behind developments in the AI industry or may not be available yet in the peer-reviewed literature.

First, the majority of studies used relatively small datasets, frequently using the same or selected subsets of the same source datasets to train models; and many of the eligible studies provided limited information on the methods used to verify the source datasets in terms of a reference standard. Most of these imaging datasets were enriched with malignant lesions, with studies often selecting images containing suspicious abnormalities. This is reflected in the high percentage of breast cancer in the datasets used to train AI algorithms in the majority of studies (median 26.5%). Whereas this approach supports the feasibility of conducting the 'experiment' and developing an AI model, resulting model performance has unclear (uncertain) applicability to a real-world screening where only around 0.5%–0.8% of screens will contain cancer. Only one study from Ayer et al. [31] had a cancer prevalence approximating that encountered in screening practice, and that study differed from the other studies because it focused on the combination of classification (of mammography findings) and risk prediction. The use of small cancer-enriched datasets presents a methodologic concern that raises substantial uncertainty regarding the quality of the imaging data used to train AI models in terms of limited applicability (external validity) beyond the reported experiment. The over-representation of malignant lesions in the imaging samples would be expected to affect reported measures of AI model performance potentially over-estimating accuracy. Second, the majority of studies did not undertake validation of the developed AI model using an independent external data-set (and the few that did so used small selected datasets), raising more uncertainty regarding transferability of the model's performance to breast cancer screening.

We found that studies were mostly focused on describing, refining, enhancing, and diversifying the AI techniques and algorithms, with little attention given to whether (or how) the imaging data sets used to train and test the AI models were representative of images encountered routinely in the breast screening context, and whether AI models were capable of recognizing the common 'normality' inherent in the screening scenario. AI algorithms may perform differently in different patient populations given heterogeneity of breast cancer risk factors and potentially imaging features between populations. This limitation suggests that larger validation datasets, preferably in diverse screening environments and population, are required in order for promising AI algorithms to progress to the next step of clinical development. As evidenced by the high proportion of cancerous images in the data sources used thus far, the imaging data may not be representative of the real-world screening setting and may additionally be biased due to deviation from the spectrum of findings usually seen in breast screening. Bias in datasets used to train AI algorithms is likely to lead to similar bias when applied in screening practice or may lead to non-robust models not due to poor algorithmic science but due to unbalanced imaging datasets. This problem may be magnified by the small sample sizes of imaging datasets in most of the studies, with the exception of two studies that trained AI models using larger datasets [18,31].

Third, there were limited data on AI versus human interpretation of breast screening examinations. Only five studies reported comparative estimates of accuracy for AI and radiologists [9,14,16,18,31], and those studies generally showed that the AI models achieved accuracy measures that approximate those of radiologists (Table 2). One of the largest studies based on imaging sets from a Dutch screening program, from Kooi et al. [18], showed a high AUC for the AI model (AUC 92.9%); however, this estimate was significantly lower at the testing phase (AUC 85.2%) than the mean AUC for radiologists (AUC 91.1%). Future studies should compare AI algorithms to radiologists' performance in *unselected* screening examinations, or report the incremental improvement for AI algorithms in combination with radiologist interpretive performance. It may be that AI algorithms are detecting different findings than human interpreters, and vice versa, but this cannot be determined from the currently available studies. We also searched the eligible studies for clinical outcome measures or conventional breast screening metrics (such as cancer detection rates or recall rates) but did not identify any data on these outcomes, and none of the studies attempted to canvass women's or societal perspectives on the acceptability of AI. We also noted that none of the abstracts retrieved in the literature search addressed the latter issues. It is likely that research into women's or societal perspectives is beyond the scope of studies evaluating AI models for breast cancer detection, however, these issues merit consideration in future AI research.

Finally, all the currently published studies meeting our inclusion criteria developed AI models using data from screen-film or digital mammography. However, as DBT is progressively becoming the breast screening modality of choice, future AI studies should include imaging data from DBT screening. AI algorithms that are only developed and validated using conventional (2D) mammography data may be outdated by the time of clinical adoption, as more than half of screening facilities in the USA now have DBT capability [48]. Moreover, DBT represents volumetric data from multiple summed 2D imaging slices, with the prospect of providing a much larger amount of quantitative imaging data that could further improve AI algorithm performance. Therefore, future testing and validation imaging sets should include DBT screening examinations linked to radiologist performance and cancer outcomes data.

There are limitations to our scoping review; we focused on published studies from 2010 onwards to factor in advances in AI capabilities, such as deep learning, therefore we did not review older studies that paved the way for more recent AI studies. We did not attempt to detail the AI techniques or computational methods reported in the eligible studies beyond the basic details shown in Table 2, we recognize the heterogeneity of AI systems and that this impacts model performance but this was beyond the scope of our review. We were aiming to gauge the readiness of AI for screening application, rather than describing the highly detailed techniques of AI models. Finally, as for any structured review, we had pre-specified inclusion criteria, hence some studies (such as those restricted solely to cases who had biopsy) were not eligible for inclusion in this scoping review.

4.1. Conclusions

Our scoping review of studies of AI for breast cancer detection showed predominantly retrospective studies based on relatively small and highly selected image datasets and has identified methodologic limitations that detract from the applicability of AI systems in the breast screening setting. Although the reviewed studies used novel techniques and reported encouraging results for AI model accuracy, the methodologic issues highlighted in our work (such as use of imaging data that may not represent the screening setting, the potential for bias in model training, and the lack of comparative data) can inform future studies and improve the translation of AI systems into breast cancer screening practice.

4.2. Expert opinion

We foresee that several factors, in combination, are driving a growing interest in the development of AI approaches for routine breast cancer screening. These factors include advances in AI sciences, including increased computing power and cloud storage of large amounts of imaging data, as well as a genuine need to improve breast screening outcomes, such as reducing false-positive mammography screening results. Moreover, AI approaches that can help decrease human workload would improve screening practices in resource-limited screening settings or in population breast screening programs that currently rely on double-reading.

Beyond improved techniques for training and validating dedicated AI models for mammography screening, large prospective studies will be needed to evaluate developed AI models using a mix of screening examinations that represent real-world screening scenarios (in terms of a spectrum of positive and negative imaging findings, and cancer prevalence in populations). Ideally, these should be validated using independent large screening datasets from diverse populations, with input from imaging experts and those working in the screening environment, to ensure relevance and timely translation. Currently, these data exist in closed, national screening programs with complete cancer capture. Ideally, these datasets linked to ground truth could be used to validate the many commercial AI algorithms that are potentially likely to gain approval for direct consumer marketing over the next five years.

We believe that well-designed studies should be developed to compare AI algorithms to radiologists' performance or to estimate the incremental improvement (or change) in accuracy when AI algorithms are combined with radiologists' interpretations or substituted for one of two screen-readers. These studies should factor in the unexplored interaction between the AI algorithm output and the radiologists' use of this additional information to arrive at an ultimate recommendation. The incremental improvement of AI in combination with human interpretation will be critical to organized screening programs that use double-reading, as an effective AI system could be a solution to radiologist shortages by creating a single-reader model with AI support.

We also anticipate that future studies will soon develop and test models for the interpretation of DBT to improve detection metrics and to ensure relevance for future population breast cancer screening practice. As DBT becomes a screening modality of choice in some programs, AI algorithms will have to adapt to the new imaging modality. However, in contrast, some screening programs may have just recently adopted digital mammography and changing hardware to DBT systems may be cost-prohibitive. Addition of a cost-effective AI algorithm in combination with DM may demonstrate an incremental improvement to screening accuracy that could approximate DBT performance and be a more cost-effective solution for these programs.

Finally, we expect that future research in AI development and evaluation will progress in parallel with qualitative research that addresses the major knowledge gaps around the acceptability of using AI in breast cancer screening services, and the many ethical, social, and legal implications of their use in healthcare. In addition, from a big picture perspective, if AI is adopted in breast screening practice, the benefits and harms trade-off inherent in population breast cancer screening will need to be reassessed to factor in the incremental benefits and harms including the unintended consequences from using AI in lieu of human image interpretation.

Funding

This work was supported by a National Breast Cancer Foundation (NBCF Australia) Breast Cancer Research Leadership Fellowship to N Houssami; CI Lee was supported in part by the Safeway Foundation; G Kirkpatrick-Jones was supported by a Sydney Summer Scholarship from the Faculty of Medicine & Health, University of Sydney.

Declaration of interest

CI Lee declares research grant funding (to institution) from GE Healthcare unrelated to the work in this manuscript. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. The national artificial intelligence research and development strategic plan. In: Council NSaT, editor. United States: networking and Information Technology Research and Development Subcommittee; US: National Science and Technology Council; 2016. p. 1–40.

- **Comprehensive AI research and development report.**

2. Houssami N, Lee CI, Buist DSM, et al. Artificial intelligence for breast cancer screening: opportunity or hype? Breast. 2017;36:31–33.
3. Trister AD, Buist D, Lee CI. Will machine learning tip the balance in breast cancer screening? JAMA Oncol. 2017;3:1463.

- **Insightful commentary on AI for mammography screening.**

4. Wang D, Khosla A, Gargeya R, et al. Deep learning for identifying metastatic breast cancer. Beth Israel Deaconess Medical Center, Harvard Medical School; 2016. p. 1–6.
5. Peters MD, Godfrey C, Khalil H, et al. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc. 2015 Sep;13(3):141–146.
6. Colquhoun HL, Levac D, O'Brien KK, et al. Scoping reviews: time for clarity in definition, methods, and reporting. J Clin Epidemiol. 2014;67(12):1.
7. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol. 2005;8(1):19–32.
8. Tricco AC MSc, Lillie E MSc, Zarin W MPH, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018 Sep 4;169(7):467–473.
9. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst. 2019;111(9):djy222.
10. Al-Masni MA, Al-Antari MA, Park J-M, et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Comput Methods Programs Biomed. 2018;157:85–94.
11. Ribli D, Horváth A, Unger Z, et al. Detecting and classifying lesions in mammograms with deep learning. Sci Rep. 2018;8(1):4165.
12. Chougrad H, Zouaki H, Alheyane O. Deep convolutional neural networks for breast cancer screening. Comput Methods Programs Biomed. 2018 Apr 01;157: 19–30.
13. Bandeira Diniz JO, Bandeira Diniz PH, Azevedo Valente TL, et al. Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks. Comput Methods Programs Biomed. 2018;156:191–207.
14. Becker AS, Mueller M, Stoffel E, et al. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. Br J Radiol. 2018;91(1083):20170576. PubMed PMID: 29215311.
15. Lotter W, Sorensen G, Cox D. A multi-scale CNN and curriculum learning strategy for mammogram classification. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer; 2017. p. 169–177.
16. Becker AS, Marcon M, Ghafoor S, et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol. 2017;52(7):434–440. PubMed PMID: 00004424-201707000-00007.
17. de Oliveira Silva LC, Barros AK, Lopes MV. Detecting masses in dense breast using independent component analysis. Artif Intell Med. 2017;80:29–38.
18. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2017;35:303–312.
19. Samala RK, Chan H-P, Hadjiiski LM, et al. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. Phys Med Biol. 2017;62(23):8894–8908. PubMed PMID: 29035873.
20. Carneiro G, Nascimento J, Bradley AP. Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans Med Imaging. 2017;36(11):2355–2365.
21. Teare P, Fishman M, Benzaquen O, et al. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J Digit Imaging. 2017;30(4):499–505. PubMed PMID: 28656455.
22. Dhungel N, Carneiro G, Bradley AP. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med Image Anal. 2017 Apr 01;37: 114–128.
23. Sun W, Tseng T-L, Zhang J, et al. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Computerized Med Imaging Graphics. 2017;57:4–9.
24. Saraswathi D, Srinivasan E. A CAD system to analyse mammogram images using fully complex-valued relaxation neural network ensembled classifier. J Med Eng Technol. 2014 Oct 01;38(7):359–366.
25. Velikova M, Lucas PJF, Samulski M, et al. On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks. Artif Intell Med. 2013 Jan 01;57(1):73–86.
26. Dheeba J, Tamil Selvi S. An improved decision support system for detection of lesions in mammograms using differential evolution optimized wavelet neural network [journal article]. J Med Syst. 2012 Oct 01;36(5):3223–3232.
27. Dheeba J, Selvi ST. A swarm optimized neural network system for classification of microcalcification in mammograms [journal article]. J Med Syst. 2012 Oct 01;36(5):3051–3061.
28. Parmeggiani D, Avenia N, Sanguinetti A, et al. Artificial intelligence against breast cancer (A.N.N.E.S-B.C.-Project). Ann Ital Chir. 2012 Jan-Feb;83(1):1–5. PubMed PMID: 22352208; eng.
29. Lesniak JM, Hupse R, Blanc R, et al. Comparative evaluation of support vector machine classification for computer aided detection of breast masses in mammography. Phys Med Biol. 2012;57(16):5295.
30. Huang M-L, Hung Y-H, Lee W-M, et al. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis [journal article]. J Med Syst. 2012 Apr 01;36(2):407–414.
31. Ayer T, Alagoz O, Chhatwal J, et al. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer . 2010;116(14):3310–3321. PubMed PMID: 20564067.

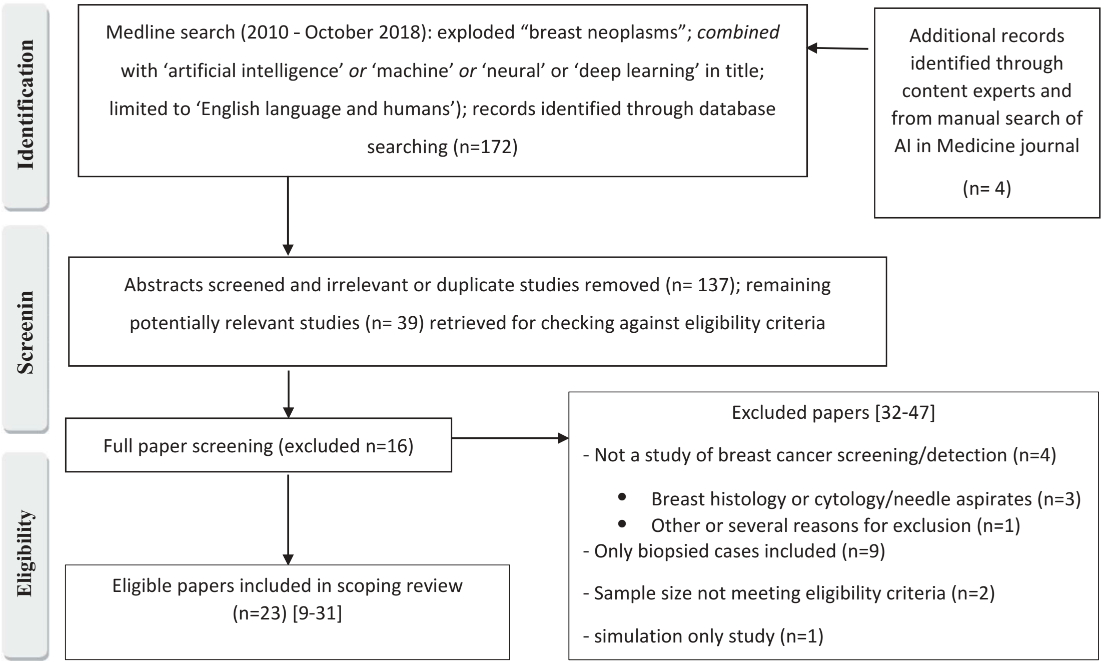
- **A relevant study using a large mammography dataset.**

32. Cai H, Peng Y, Ou C, et al. Diagnosis of breast masses from dynamic contrast-enhanced and diffusion-weighted MR: a machine learning approach. PLOS ONE. 2014;9(1):e87387.
33. Chen H-L, Yang B, Wang G, et al. Support vector machine based diagnostic system for breast cancer using swarm intelligence [journal article]. J Med Syst. 2012 Aug 01;36(4):2505–2519.
34. Dheeba J, Albert Singh N, Tamil Selvi S. Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach. J Biomed Inform. 2014 Jun 01;49: 45–52.
35. Hsieh S-L, Hsieh S-H, Cheng P-H, et al. Design ensemble machine learning model for breast cancer diagnosis [journal article]. J Med Syst. 2012 Oct 01;36(5):2841–2847.
36. Huang M-L, Hung Y-H, Chen W-Y. Neural network classifier with entropy based feature selection on breast cancer diagnosis [journal article]. J Med Syst. 2010 Oct 01;34(5):865–873.
37. Kamra A, Jain VK, Singh S, et al. Characterization of architectural distortion in mammograms based on texture analysis using support vector machine classifier with clinical evaluation. J Digit Imaging. 2016;29(1):104–114. PubMed PMID: 26138756.
38. Liu B, Jiang Y. A multitarget training method for artificial neural network with application to computer-aided diagnosis. Med Phys. 2013;40(1):011908. PubMed PMID: 23298099.
39. Qiu Y, Yan S, Gundreddy RR, et al. A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. J Xray Sci Technol. 2017;25(5):751–763. PubMed PMID: 28436410.
40. Seokmin H, Ho-Kyung K, Ja-Yeon J, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol. 2017;62(19):7714.
41. Tan M, Pu J, Zheng B. Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support

vector machine (SVM) model. Int J Comput Assist Radiol Surg. 2014;9(6):1005–1020. PubMed PMID: 24664267.

42. Venkatesh SS, Levenback BJ, Sultan LR, et al. Going beyond a first reader: a machine learning methodology for optimizing cost and performance in breast ultrasound diagnosis. Ultrasound Med Biol. 2015 Dec 01;41(12):3148–3162.
43. Wang J, Yang X, Cai H, et al. Discrimination of breast cancer with microcalcifications on mammography by deep learning [Article]. Sci Rep. 2016 Jun 07;6: 27327. online. Available from: https://www.nature.com/articles/srep27327#supplementary-information
44. Wu W-J, Lin S-W, Moon WK. Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. Computerized Med Imaging Graphics. 2012 Dec 01;36 (8):627–633.
45. Wu W-J, Lin S-W, Moon WK. An artificial immune system-based support vector machine approach for classifying ultrasound breast tumor images [journal article]. J Digit Imaging. 2015 Oct 01;28 (5):576–585.
46. Zhang Q, Xiao Y, Dai W, et al. Deep learning based classification of breast tumors with shear-wave elastography. Ultrasonics. 2016;72:150–157.
47. Kooi T, Ginneken B, Karssemeijer N, et al. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. Medical Physics. 2017;44 (3):1017-1027. doi:10.1002/mp.12110.
48. Houssami N, Miglioretti DL. Digital breast tomosynthesis: a brave new world of mammography screening. JAMA Oncol. 2016;2(6):725–727.
    - **Concise overview on tomosynthesis for breast cancer screening.**

Appendices


Identification
Screenin
Eligibility
Medline search (2010 - October 2018): exploded "breast neoplasms"; *combined* with 'artificial intelligence' *or* 'machine' *or* 'neural' or 'deep learning' in title; limited to 'English language and humans'); records identified through database searching (n=172)
Additional records identified through content experts and from manual search of AI in Medicine journal (n= 4)
Abstracts screened and irrelevant or duplicate studies removed (n= 137); remaining potentially relevant studies (n= 39) retrieved for checking against eligibility criteria
Full paper screening (excluded n=16)
Excluded papers [32-47]
- Not a study of breast cancer screening/detection (n=4)
• Breast histology or cytology/needle aspirates (n=3)
• Other or several reasons for exclusion (n=1)
- Only biopsied cases included (n=9)
- Sample size not meeting eligibility criteria (n=2)
- simulation only study (n=1)
Eligible papers included in scoping review (n=23) [9-31]

Appendix 1. Literature search and study identification strategy – Artificial Intelligence (AI) for breast cancer detection

Appendix 2. Database search terms

Database: Ovid MEDLINE(R) <1946 to October Week 4 2018>
1 exp Breast Neoplasms/ (268387)
2 limit 1 to (English language and humans and yr = "2010 -Current") (90643)
3 artificial intelligence.m_titl. (659)
4 machine.m_titl. (8965)
5 neural.m_titl. (61500)
6 deep learning.m_titl. (366)
7 3 or 4 or 5 or 6 (71321)
8 2 and 7 (172)