



## Artificial intelligence and convolution neural networks assessing mammographic images: a narrative literature review

Dennis Jay Wong, MDR, BMSc (Hons),  Ziba Gandomkar, PhD, MSc, BSc, Wan-Jing Wu, MDR, Guijing Zhang, MDR, Wushuang Gao, MDR, Xiaoying He, MDR, Yunuo Wang, MDR, & Warren Reed, PhD, Grad Cert (T&L), Bsc (Hons) 

Discipline of Medical Imaging Sciences, The University of Sydney, Lidcombe, New South Wales, Australia

### Keywords

Artificial intelligence, breast cancer, breast density, convolutional neural network, mammography

### Correspondence

Dennis Jay Wong, M504, Block M, 75 East Street, Lidcombe, 2141, New South Wales, Australia. Tel: +61 02 9351 9586; E-mail: [dwon6926@uni.sydney.edu.au](mailto:dwon6926@uni.sydney.edu.au)

Received: 3 October 2019; Revised: 18 January 2020; Accepted: 11 February 2020

*J Med Radiat Sci* **67** (2020) 134–142

doi: 10.1002/jmrs.385

### Abstract

Studies have shown that the use of artificial intelligence can reduce errors in medical image assessment. The diagnosis of breast cancer is an essential task; however, diagnosis can include ‘detection’ and ‘interpretation’ errors. Studies to reduce these errors have shown the feasibility of using convolution neural networks (CNNs). This narrative review presents recent studies in diagnosing mammographic malignancy investigating the accuracy and reliability of these CNNs. Databases including ScienceDirect, PubMed, MEDLINE, British Medical Journal and Medscape were searched using the terms ‘convolutional neural network or artificial intelligence’, ‘breast neoplasms [MeSH] or breast cancer or breast carcinoma’ and ‘mammography [MeSH Terms]’. Articles collected were screened under the inclusion and exclusion criteria, accounting for the publication date and exclusive use of mammography images, and included only literature in English. After extracting data, results were compared and discussed. This review included 33 studies and identified four recurring categories of studies: the differentiation of benign and malignant masses, the localisation of masses, cancer-containing and cancer-free breast tissue differentiation and breast classification based on breast density. CNN’s application in detecting malignancy in mammography appears promising but requires further standardised investigations before potentially becoming an integral part of the diagnostic routine in mammography.

### Introduction

Screening mammography is the recommended tool for the early detection of breast cancer in women experiencing no symptoms and has been shown to decrease breast cancer mortality rate by 40–63%.<sup>1</sup> However, current diagnostic frameworks for assessing mammograms such as the breast imaging reporting and data system (BI-RADS), developed by the American College of Radiology (ACR) used to report findings into a number of well-defined categories, can be limited by ‘detection errors’ (pathology missed) and ‘interpretation errors’ (pathology misinterpreted).<sup>2</sup> One possible solution to reduce the errors is utilising artificial intelligence (AI) tools, which automatically find abnormalities (detection

and categorise them either as normal or as abnormal (interpretation).

In general, AI can be defined as a sophisticated computer program able to carry out tasks usually requiring human intellect in areas such as visual perception and decision-making. Over the past few years, deep learning has dramatically transformed the AI industry. Deep learning is a subset of machine learning methods based on artificial neural networks. A convolutional neural network (CNN) is a class of deep neural networks, commonly applied to analyse image data. Similar to conventional machine learning methods, CNN allows the machine to perform specific tasks by relying on the inference of decision boundaries rather than explicit instructions by a user.

Furthermore, CNNs learn relevant features from the data, with limited input from a human expert, and make predictions via heuristics when new data are entered.<sup>3</sup> CNNs can be used to assist with the detection of breast cancer on mammographic images with a high degree of accuracy as measured by sensitivity and specificity.<sup>4</sup> It can also reduce the time required to assess a large volume of mammograms.<sup>5</sup> These potential benefits have emphasised the importance of exploring CNN's ability to accurately diagnose abnormalities on mammograms promptly.<sup>4,5</sup> CNN is faster with similar accuracy when compared to radiologists without other limiting factors such as a lack of qualified personnel.<sup>6,7</sup>

Considering recent advances in AI tools in mammography, it is essential for the imaging and the wider health community to better understand current trends and developments. However, there is a lack of comparisons in the literature between similar studies that this literature review aims to address. The purpose of this study therefore was to compare the recent developments of CNN in the diagnosis of mammographic images to enable medical imaging professionals to gain greater insight on how this technology can assist our patients in future.

## Methodology

The focus of this review was to identify the current developments and uses of CNN in mammography diagnostics. Due to the review's narrative nature, results were described by description and exploration with qualitative elements rather than systematically.

Using the keywords listed in Table 1, ScienceDirect, PubMed, MEDLINE, British Medical Journal and Medscape databases were searched for original English research studies published between 2011 and 2019. Studies which used imaging modalities other than mammography were excluded. A total of 33 relevant peer-reviewed articles were found.<sup>8–40</sup> To compare the results among studies, relevant features were extracted: (1) author and publication year; (2) study aim; (3) utilised database; (4) number of images used; and (5) performance measures including sensitivity, specificity and area under the curve (AUC).

## Results

A total of 33 out of 65 articles were selected under the inclusion criteria; 32 studies were rejected due to lack of direct relevance. In the accepted literature, there were various recurring themes, so articles were further categorised into four different salient categories:

Category 1. Differentiation between benign and malignant tissues;

**Table 1.** List of keywords used to search for relevant topics related to the review.

Topic	Keywords
Artificial Intelligence	Artificial Intelligence; Deep learning; Artificial Neural Networks; Convolutional Neural Network
Anatomy; Pathology of Interest	Breast; Breast Cancer; Breast Lesion; Breast Carcinoma; Breast Neoplasm
Imaging Modality	Mammogram; Mammography
Others	Image Reading; Diagnostics; BI-RADS; Breast Screening

Category 2. Localisation of mass in breast tissue;

Category 3. 'Cancer-containing' and 'cancer-free' breast tissue differentiation;

Category 4. Breast classification based on breast density.

A total of 12 studies focused exclusively on Category 1, primarily testing CNN's ability to distinguish between benign and malignant masses. Category 2 had seven studies that focused on mass localisation. Category 3 had seven studies which focused on screening images as 'cancer-free' and 'cancer-containing'; Category 4 had seven studies that used CNN to categorise mammograms by breast density.

Twenty-two of the 33 articles used the ACR's BI-RADS system to report mammograms assigning breast imaging studies to one of six assessment categories (0 – incomplete study, 1 – negative, 2 – benign, 3 – probably benign, 4 – suspicious for malignancy, 5 – highly suggestive of malignancy and 6 – known biopsy-proven malignancy).<sup>41</sup>

In all 33 articles, four primary image databases were used:

1. Digital Database for Screening Mammography (DDSM) (Digitised Images): This database from the University of South Florida contains 2500 studies, with each study containing two images of each breast and patient information such as age, when the image biopsy proven, and ACR breast density classification. All cases are female.<sup>42</sup>
2. MIAS Mini Mammographic Database (Digital Images): The mini-MIAS provided by the Mammographic Image Analysis Society contains 161 pairs of films and has commonly encountered pathologies and normal cases.<sup>43</sup>
3. INbreast database (Digital images): This database has 410 images from 115 cases: 90 cases have images of both breasts and 25 cases from mastectomy patients. All patients are females. Different pathologies include masses, calcification, asymmetries and distortions.<sup>44</sup>
4. Breast Cancer Digital Repository (BCDR) (Digital and Digitised images): BCDR is a public repository

composed of breast cancer patient's case studies hosted by the IMED Project in March 2009.<sup>8</sup> Data classification uses BI-RADS and is subdivided into two repositories: the Film Mammography Repository and Full Field Digital Mammography Repository. The BCDR has been used to train CAD schemes and is still in development.<sup>45</sup> Digitised images contain 1010 patients (998 female and 12 male) ranging from ages 20 to 90, 1125 studies and a total of 3703 images. The digital database is in development with 724 patients (723 female and 1 male) ranging from ages 27 to 92, 1042 studies and a total of 3612 images.<sup>46</sup>

## Discussion

### Category 1: Differentiation between a benign and malignant mass

The differentiation between benign and malignant masses in CNN breast screening was the most prominent category of study. Akselrod-Ballin's CNN accurately diagnosed mammograms classified as BI-RADS 1; however, it found difficulty diagnosing higher graded images.<sup>19</sup> Importantly, they noted that CNN systems have issues differentiating benign and malignant masses in BI-RADS 3 images.<sup>19</sup> When BI-RADS 3 images were excluded, their CNN system improved detection AUC from 0.60 to 0.72. Aside from this study, other studies did not mention BI-RADS 3 exclusion.

Furthermore, Carneiro suggested that their CNN's inability to diagnose BI-RADS images >1 was due to training images being annotated for either microcalcifications or masses, but not for both in a single image, potentially affecting heuristics.<sup>15</sup> Additionally, automated segmentation made by CNN generally resulted in lower accuracy, since detection of false-positive masses impacts the system's performance in mass classification.<sup>15</sup>

### Category 2: Localisation of masses in breast tissue

Although many studies focused on benign and malignant differentiation, tumour localisation was another common topic. Five studies showed detecting a mass in dense breast regions and pectoral muscles to be difficult due to high intensities.<sup>21–24,36</sup>

Al-Masni *et al.*<sup>47</sup> employed a You Only Look Once (YOLO) system that uses a singular network for diagnostics by dividing the input image into multiple subregions, compared to traditional neural networks requiring multiple networks for each extracted region. As a result, their system operated quicker compared to traditional neural networks and successfully identified masses in pectoral muscles and

dense tissues; they suggested that its ability to localise masses in radiopaque areas was due to augmented training by digitally manipulating images to produce more training images.<sup>47</sup>

Hwang *et al.*<sup>21</sup> experimented with CNN's ability to self-learn via a weakly supervised CNN system, meaning that images used for training were minimally annotated with a weak description. Compared to other studies, the results were poor for AUC, sensitivity, specificity and accuracy with no results exceeding 0.77.<sup>21</sup> This finding suggests that current CNN's need highly annotated images for training for accurate results.

Overall, the results from the studies in Category 2 appear promising in highly supervised training methods, but weak in lowly supervised systems.

### Category 3: 'Cancer-containing' and 'cancer-free' breast tissue differentiation

In 2018, a study using a CNN system was designed to differentiate between normal and abnormal images via biomarkers such as breast density, presence of mass and microcalcification.<sup>28</sup> Their results indicated that abnormal images were more accurately identified when microcalcifications and benign cases were excluded. The authors suggested that due to the training data having an uneven number of cases of normal and cancer cases, this caused overfitting and database dependency. This creates difficulty distinguishing between benign and malignant cases due to overreliance on heuristics and an inability to learn and adapt from new cases.<sup>28</sup>

Al-Antari modified Al-Masni *et al.*'s YOLO system to differentiate between cancer-containing and cancer-free images, their study demonstrated an accuracy of 95.64% and noted that a robust CNN system heavily relies on the deep learning model in segmenting specific region of masses to reduce potential false positives from surrounding tissues.<sup>30</sup>

### Category 4: Breast classification based on breast density

Breast tissue density may increase the likelihood of developing breast cancer, as increased breast density and glandular tissue may make it difficult to detect early signs of breast cancer such as microcalcifications.<sup>47</sup> The importance of classifying breast density for breast cancer risk must be stressed as the relative risk of developing breast cancer in women with heterogeneously dense breasts is 1.2 times more compared to the average women, and for women with extremely dense breasts, the relative risk is 2.1 times higher. This is likely due to the masking effect, excess glandular tissue and breast density as an independent risk factor.<sup>48</sup>

Between 2018 and 2019, there were studies that tested CNN's capabilities in categorising mammograms by breast density.<sup>37,38,40,49</sup> Ha's preliminary system demonstrated an accuracy of 72% in predicting 'high'-risk and 'low'-risk mammograms by generating heat maps on the mammograms where 'red' areas had overlapping mammographic features suggesting high density, high and low risk defined as dense and non-dense, respectively.<sup>37</sup> They noted that while the red areas on the heatmaps suggest areas of high density, the pixel map generated does not necessarily indicate specific areas where breast cancer may develop.<sup>37</sup> Notably, Mohammed's study demonstrated a high accuracy on differentiating between BI-RADS category 'scattered density' and 'heterogeneously dense' images, however, noted that their study was a retrospective study and the images used had little variation in mammographic vendors and imaging protocol, suggesting that their system may not guarantee the same results from different manufacturers and imaging standards.<sup>38</sup>

Fonseca *et al* conducted a preliminary study to diagnose breast density using the ACR's breast composition categories (a – entirely fatty; b – scattered areas of fibro-glandular density; c – heterogeneously dense tissue, potentially obscuring small masses; and d – breasts extremely dense, lowering mammography sensitivity)<sup>41</sup> and achieved a mean accuracy of 0.7305, similar to a radiologist when categorising breast density.<sup>35</sup> The study's weakness was that the images used were sourced locally and not from a mainstream database. Additionally, a study found that inter-reader agreement regarding breast density classification was only at 49% and agreement was usually on either fatty breasts or extremely dense breasts.<sup>50</sup> Therefore, there is a great potential in improving breast density classification in assessing breast cancer risk using CNN's capabilities.

Diniz in 2018<sup>36</sup> classified breast tissue by density and regions with or without a mass; by combining them, they localised and detected masses in dense (Cooper's ligaments, mamillary glands and ducts) and non-dense tissue (adipose tissue). Sensitivity was slightly higher in non-dense tissue (0.9156 vs. 0.9036), yet dense tissue had significantly higher specificity (0.9635 vs. 0.9073), meaning that their CNN was better at recognising negative non-dense images but identified abnormalities more accurately in dense tissues.<sup>36</sup>

## Sensitivity and specificity, total accuracy and AUC of CNNs

Basic statistical analyses are employed for simple comparisons between a study's AUC, sensitivity, specificity and total accuracy gathered from Table 2. The

overall mean of CNN's AUC, specificity, sensitivity and total accuracy is listed in Table 3. It is important to note that data sets used in the radiologist studies are not standardised and images are from patients with vastly different clinical history.

For sensitivity, minimum and maximum value disparity is greater for CNN, but the lowest and highest sensitivity range is more significant. The lowest sensitivity and specificity are from a study by Hwang and Kim; their results are lowest across all parameters, reinforcing that a highly supervised learning requirement is required. They argued that the training database (DDSM) contained low-quality images with many artefacts. However, this issue did not occur for other studies.

Of note, some studies omitted sensitivity and specificity results, so performance comparisons between the CNN and the radiologist are rudimentary. Furthermore, every study used a different number of images to train and test their systems, making exact comparisons difficult.

From these four parameters, the current development of CNN appears promising in mammographic diagnostics but requires highly supervised learning (images that are highly annotated with regions of interests to learn and be segmented), sufficient images for training and adjustment of learning algorithms to ensure overfitting and database dependency do not occur to achieve good diagnostic results.

In 2018, a retrospective study compared the performance of a CNN system to 101 radiologists and found that the CNN system slightly outperformed the radiologists with an AUC of 0.840 and 0.814, respectively.<sup>51</sup> However, due to the study's retrospective nature, results may suffer from the 'laboratory effect' and may not be applicable in a clinical setting.<sup>51</sup>

## Limitations of this review and current literature

Although some studies have used the same databases for training, many have used a different combination of databases and various numbers of images for training. Hence, valuable comparisons between studies are difficult to make as each database uses mammograms from different hospitals, image parameters and manufacturers, resulting in a lack of reference standards between data sets. Since the study outcomes depend significantly on the database used for training, the variability between these databases can affect each AI's performance algorithm. This variability is called 'inter-database variability', causing optimisation of algorithm and standardised comparison between studies extremely challenging.<sup>52</sup>

Additionally, there are several 'cases' per database and studies may not use the same cases. For example, Rouhi

**Table 2.** 33 studies listed by category, author, year, database used, number of images, AUC, specificity, sensitivity and accuracy.<sup>8–40</sup>

Author	Year	Database	#Cases (Images)	AUC	Specificity	Sensitivity	Accuracy
<i>Category 1 studies</i>							
Ramos-Pollán et al. <sup>8</sup>	2011	BCDR	286	0.996	0.77	0.95	0.97
Rouhi et al. <sup>9</sup>	2014	DDSM	170 (170)	0.951	0.96	0.97	0.96
Arevalo et al. <sup>10</sup>	2015	BCDR	344 (736)	0.86	–	–	–
Arevalo et al. <sup>11</sup>	2016	BCDR	344 (736)	0.7	–	–	–
Dhungel et al. <sup>12</sup>	2016	INbreast	115 (410)	0.87	–	–	0.91
Huynh et al. <sup>13</sup>	2016	Custom	– (–)	0.81	–	–	–
Jiao et al. <sup>14</sup>	2016	DDSM	– (–)	–	–	–	0.97
Carneiro et al. <sup>15</sup>	2017	INbreast	115 (410)	0.72	0.66	0.69	–
Carneiro et al. <sup>15</sup>	2017	INbreast	115 (410)	0.87	0.92	0.69	–
Carneiro et al. <sup>15</sup>	2017	DDSM	172 (680)	0.91	0.97	0.94	–
Kooi et al. <sup>27</sup>	2017	Custom	956 (1804)	0.8	–	–	–
Teare et al. <sup>17</sup>	2017	DDSM	0	0.92	0.91	0.8	–
Chougrad et al. <sup>18</sup>	2018	DDSM	1329 (5316)	0.98	–	–	0.97
Chougrad et al. <sup>18</sup>	2018	INbreast	50 (200)	0.97	–	–	0.96
Chougrad et al. <sup>18</sup>	2018	BCDR	300 (600)	0.96	–	–	0.97
Chougrad et al. <sup>18</sup>	2018	MD***	1529 (6116)	0.99	–	–	0.99
Chougrad et al. <sup>18</sup>	2018	MIAS	113 (113)	0.99	–	–	0.98
Akselrod-Ballin et al. <sup>*19</sup>	2016	DDSM, INbreast	850 (850)	–	–	–	0.78
Akselrod-Ballin et al. <sup>**19</sup>	2016	DDSM	850 (850)	–	–	–	0.77
	2016	INbreast (without BI–RADS 3)	– (–)	–	–	–	–
Akselrod-Ballin et al. <sup>*19</sup>	2016	DDSM, INbreast	–850	0.6	–	–	–
Akselrod-Ballin et al. <sup>**19</sup>	2016	DDSM, INbreast (without BI–RADS 3)	–850	0.72	–	–	–
Ribli et al. <sup>20</sup>	2018	INbreast	115 (115)	0.85	0.9	–	–
<i>Category 2 studies</i>							
Hwang et al. <sup>21</sup>	2016	DDSM and MIAS	10,363 (322)	0.68	0.76	0.58	0.7
Hwang et al. <sup>21</sup>	2016	DDSM and MIAS	10,363 (322)	0.54	0.66	0.44	0.66
Al-masni et al. <sup>47</sup>	2017	DDSM	–600	–	–	–	0.9633
Sampaio <sup>23</sup>	2011	DDSM	566 (566)	0.87	0.86	0.8	0.85
Sun et al. <sup>24</sup>	2016	Custom	–1874	0.8818	0.72	0.81	0.8243
Shen et al. <sup>25</sup>	2019	DDSM	2223	0.922	–	0.9643	–
Savelli et al. <sup>26</sup>	2019	INbreast	115 (410)	–	–	0.763	–
Kooi et al. <sup>16</sup>	2017	Custom	44,090	0.941	–	–	–
<i>Category 3 studies</i>							
Kim et al. <sup>28</sup>	2018	Custom	29,107	0.91	0.89	0.76	–
Jadoon et al. <sup>29</sup>	2017	DDSM and MIAS	– (–)	0.85	0.82	0.88	0.82
Al-antari et al. <sup>30</sup>	2018	INbreast	115 (410)	0.9478	0.9241	0.9714	0.9564
Kaur et al. <sup>31</sup>	2019	MIAS	20	–	0.88	0.9	0.9
Anitha et al. <sup>32</sup>	2017	DDSM	300	–	–	0.925	–
Anitha et al. <sup>32</sup>	2017	MIAS	170	–	–	0.935	–
Wichakam et al. <sup>33</sup>	2016	INbreast	216	–	–	–	0.9727
Suzuki et al. <sup>34</sup>	2016	DDSM	198	–	–	0.899	–
<i>Category 4 studies</i>							
Fonseca et al. <sup>53</sup>	2015	Custom	–1157	–	–	–	0.73
Bandeira Diniz et al. <sup>36</sup>	2018	DDSM	Non-dense; – (1004)	–	0.91	0.92	0.91
Bandeira Diniz et al. <sup>36</sup>	2018	DDSM	Dense; – (1482)	–	0.96	0.9	0.95
Ha et al. <sup>37</sup>	2019	Custom	1474	–	–	–	0.72
Mohamed et al. <sup>49</sup>	2018	Custom	15,415	0.95	–	–	–
Ionescu et al. <sup>39</sup>	2019	Custom	67,520	–	–	–	–
Mohamed et al. <sup>38</sup>	2018	Custom	22,000	0.926	–	–	–
Trivizakis et al. <sup>40</sup>	2019	DDSM	2500 (10,239)	0.548	–	–	–
Trivizakis et al. <sup>40</sup>	2019	MIAS	161 (322)	0.798	–	–	–

\*Exp1.

\*\*Exp2.

\*\*\*MD = DDSM+ INbreast + BCDR.

**Table 3.** Comparison between mean, median, minimum and maximum values of AUC, specificity, sensitivity and total accuracy from the 33 CNN studies<sup>8–40</sup> and four radiologists studies.<sup>54–57</sup>

	Mean		Median		Minimum		Maximum	
	CNN	Radiologist	CNN	Radiologist	CNN	Radiologist	CNN	Radiologist
AUC	0.851	–	0.876	–	0.540	–	0.996	–
Specificity	0.851	0.905	0.890	0.808	0.660	0.889	0.970	0.988
Sensitivity	0.833	0.795	0.899	0.916	0.440	0.600	0.971	0.899
Total accuracy	0.883	–	0.930	–	0.660	–	0.990	–

et al.'s<sup>9</sup> model aimed to differentiate malignant or benign lesions, while Diniz et al.'s<sup>36</sup> design aimed to localise mass and categorise images based on breast tissue density. Although both used the same database, the cases used were different, meaning that direct comparisons are difficult. Furthermore, current data sets are archaic and lack realism; many images still use digitised rather than digital images; the data sets lack clinical history such as risk factors, ethnicity and age make thorough analysis difficult. Also, in real-world screening approximately 0.5–0.8% cases contain cancer, meaning that data sets used had a high number of malignancies.

Likewise, the number of images used to pre-train and test the CNN varies between studies and this potentially affects its calibration. A CNN extensively trained with small training schemes may lead to overfitting and database dependency which causes the CNN to be over-reliant on heuristics taught in training, resulting in weak future predictions, analysis and adaptation to new images.<sup>3</sup> To improve comparisons between CNN and radiologists, their performance must be examined on the same data set and investigations to identify overlapping errors between AI and humans are recommended.

To solve these issues, mainstream databases should be selected and standardised with the same cases and images used so meaningful statistical analysis can be performed between studies such as ROC curve analysis and homogeneity.

## Conclusion

CNN diagnosis in mammography has been demonstrated as an expanding field in artificial intelligence diagnostics. The most significant limitation of the current studies is the lack of standardisation between studies. Despite this limitation, the mean accuracy across the studies is promising, showing the potential reliability of CNN in mammography diagnostics.

Current innovation in CNN diagnostics for mammography includes the differentiation of benign and malignant masses, the localisation of masses, cancer-containing and cancer-free breast tissues differentiation

and breast classification based on density. Literature indicates that the differentiation of benign and malignant masses has the most developments, and the consensus suggests that CNN is comparable in performance to a radiologist. However, there are still some significant weaknesses within the 33 studies due to the inter-database variability, inadequate original data reports and other aspects of AI training such as unsupervised learning and training schemes. Therefore, as promising as the introduction of a CNN application in detecting malignancy in breasts seems, it does not currently appear ready for the real clinical environment and requires further standardised investigations before it can be a part of the diagnostic routine for mammography imaging. With the current developments, instead of solely investigating CNN performance versus radiologists, the real opportunity is to examine CNN's potential to support radiologist's performance both for training purposes and as a computer aid to help them to further improve medical image diagnosis.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

1. Tabár L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001; **91**(9): 1724–31.
2. Boyer B, Canale S, Arfi-Rouche J, Monzani Q, Khaled W, Balleyguier C. Variability and errors when applying the BIRADS mammography classification. *Eur J Radiol* 2013; **82**(3): 388–97.
3. Lee J-G, Jun S, Cho Y-W, et al. Deep learning in medical imaging: general overview. *Korean J Radiol* 2017; **18**(4): 570.
4. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; **2**(4): 230–43.
5. Krittawong C. The rise of artificial intelligence and the uncertain future for physicians. *Eur J Intern Med* 2018; **48**: e13–4.

6. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018; **15**(11): e1002686.
7. Hanhirova J, Kämäräinen T, Seppälä S, Siekkinen M, Hirvisalo V, Ylä-Jääski A. Latency and throughput characterization of convolutional neural networks for mobile computer vision. In: Proceedings of the 9th ACM Multimedia Systems Conference on - MMSys '18 [Internet]. ACM Press, New York, 2018 [cited 2019 May 5]; 204–15. Available from: <http://dl.acm.org/citation.cfm?doid=3204949.3204975>.
8. Ramos-Pollán R, Guevara-López MA, Suárez-Ortega C, et al. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. *J Med Syst* 2012; **36** (4): 2259–69.
9. Rouhi R, Jafari M, Kasaei S, Keshavarzian P. Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Syst Appl* 2015; **42** (3): 990–1002.
10. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Convolutional neural networks for mammography mass lesion classification. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. 2015.
11. Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput Methods Programs Biomed* 2016; **127**: 248–57.
12. Dhungel N, Carneiro G, Bradley AP. The automated learning of deep features for breast mass classification from mammograms [Internet], 2019 [cited 2019 May 5]. Available from: [https://cs.adelaide.edu.au/~neeraj/mass\\_classification\\_miccai\\_16.pdf](https://cs.adelaide.edu.au/~neeraj/mass_classification_miccai_16.pdf).
13. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging* 2016; **3**(3): 34501.
14. Jiao Z, Gao X, Wang Y, Li J. A deep feature based framework for breast masses classification. *Neurocomputing* 2016; **197**: 221–31.
15. Carneiro G, Nascimento J, Bradley AP. Automated analysis of unregistered multi-view mammograms with deep learning [Internet]. 2017 [cited 2019 May 5]. Available from: [https://cs.adelaide.edu.au/~carneiro/publications/tmi\\_multimodal.pdf](https://cs.adelaide.edu.au/~carneiro/publications/tmi_multimodal.pdf).
16. 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**: 303–12.
17. Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. 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.
18. Chougrad H, Zouaki H, Alheyane O. Deep convolutional neural networks for breast cancer screening. *Comput Methods Programs Biomed* 2018; **157**: 19–30.
19. Akselrod-Ballin A, Karlinsky L, Alpert S, Hasoul S, Ben-Ari R, Barkan E. A region based convolutional network for tumor detection and classification in breast mammography, 2016 [cited 2019 May 5]; 197–205. Available from: [http://link.springer.com/10.1007/978-3-319-46976-8\\_21](http://link.springer.com/10.1007/978-3-319-46976-8_21).
20. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. *Sci Rep* 2018; **8**(1): 4165.
21. Hwang S, Kim H-E. Self-transfer learning for fully weakly supervised object localization, 2016 Feb 4 [cited 2019 May 5]. Available from: <http://arxiv.org/abs/1602.01625>.
22. Al-masni MA, Al-antari MA, Park JM, et al. Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [Internet]. IEEE; 2017 [cited 2019 May 5]; 1230–3. Available from: <https://ieeexplore.ieee.org/document/8037053/>.
23. Borges Sampaio W, Moraes Diniz E, Corrêa Silva A, Cardoso de Paiva A, Gattass M. Detection of masses in mammogram images using CNN, geostatistic functions and SVM. *Comput Biol Med* 2011; **41**(8): 653–64.
24. Sun W, Tseng T-L, Zhang J, Qian W. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput Med Imaging Graph* 2017; **57**: 4–9.
25. Shen R, Yan K, Tian K, Jiang C, Zhou K. Breast mass detection from the digitized X-ray mammograms based on the combination of deep active learning and self-paced learning. *Futur Gener Comput Syst* 2019; **101**: 668–79.
26. Savelli B, Bria A, Molinara M, Marrocco C, Tortorella F. A multi-context CNN ensemble for small lesion detection. *Artif Intell Med* 2019; **103**: 101749.
27. Kooi T, van Ginneken B, Karssemeijer N, den Heeten A. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Med Phys* 2017; **44**(3): 1017–27.
28. Kim E-K, Kim H-E, Han K, et al. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. *Sci Rep* 2018; **8**(1): 2762.
29. Jadoon MM, Zhang Q, Haq IU, Butt S, Jadoon A. Three-class mammogram classification based on descriptive CNN features. *Biomed Res Int* 2017; **2017**: 1–11.
30. Al-antari MA, Al-masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection,

- segmentation, and classification. *Int J Med Inform* 2018; **117**: 44–54.
31. Kaur P, Singh G, Kaur P. Erratum: Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification. *Informat Med Unlocked* 2019; **16**: 100151.
  32. Anitha J, Dinesh Peter J, Immanuel Alex Pandian S. A dual stage adaptive thresholding (DuSAT) for automatic mass detection in mammograms. *Comput Methods Programs Biomed* 2017; **138**: 93–104.
  33. Wichakam I, Vateekul P. Combining deep convolutional networks and SVMs for mass detection on digital mammograms. In 2016 8th Int Conf Knowl Smart Technol KST 2016, 2016; 239–44.
  34. Suzuki S, Zhang X, Homma N, et al. Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis. In 2016 55th Annu Conf Soc Instrum Control Eng Japan, SICE 2016. 2016;1382–6.
  35. Fonseca P, Mendoza J, Wainer J, Ferrer J, Pinto J, Guerrero J, Castaneda B. Automatic breast density classification using a convolutional neural network architecture search procedure. In: *Proc. SPIE 9414, Medical Imaging 2015: Computer-Aided Diagnosis*, 941428. 2015 <https://doi.org/10.1117/12.2081576>
  36. Bandeira Diniz JO, Bandeira Diniz PH, Azevedo Valente TL, Corrêa Silva A, de Paiva AC, Gattass M. 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.
  37. Ha R, Chang P, Karcich J, et al. Convolutional neural network based breast cancer risk stratification using a mammographic dataset. *Acad Radiol* 2018; **26**(4): 544–9.
  38. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. *Med Phys* 2018; **45**(1): 314–21.
  39. Ionescu GV, Fergie M, Berks M, et al. Prediction of reader estimates of mammographic density using convolutional neural networks. *J Med Imaging* 2019; **6**(3): 1.
  40. Trivizakis E, Ioannidis GS, Melissianos VD, et al. A novel deep learning architecture outperforming ‘off-the-shelf’ transfer learning and feature-based methods in the automated assessment of mammographic breast density. *Oncol Rep* 2019; **42**(5): 2009–15.
  41. Sickles E, D’Orsi C, Bassett L. ACR BI-RADS<sup>®</sup> Atlas, Breast Imaging Reporting and Data System, 5th edn. Reston, VA: American College of Radiology, 2013.
  42. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. In: Yaffe MJ (ed). *Digital Mammography 2000: Proceedings of the Fifth International Workshop on Digital Mammography*. Medical Physics, Madison, 2000; 431–434.
  43. Suckling J, Parker J, Dance D, et al. Mammographic Image Analysis Society (MIAS) database v1.21. 2015 Aug 28 [cited 2019 May 5]. Available from: <https://www.repository.cam.ac.uk/handle/1810/250394>.
  44. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. *Acad Radiol* 2012; **19**(2): 236–48.
  45. Suarez Ortega C, Es CS, Franco Valiente JM, et al. Improving the breast cancer diagnosis using digital repositories [Internet]. 2019. [cited 2019 May 5]. Available from: [http://iwbbio.ugr.es/papers/iwbbio\\_094.pdf](http://iwbbio.ugr.es/papers/iwbbio_094.pdf).
  46. BOUME. Breast Cancer digital mammography [Internet]. [cited 2019 Jul 3]. Available from: <https://bcdr.eu/infomation/about>.
  47. Al-masni MA, Al-antari MA, Park JM, et al. Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [Internet]. IEEE; 2017 [cited 2019 May 5]; 1230–3. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29060098>.
  48. Freer PE. Mammographic breast density: impact on breast cancer risk and implications for screening. *Radiographics* 2015; **35**(2): 302–15.
  49. Mohamed AA, Luo Y, Peng H, Jankowitz RC, Wu S. Understanding clinical mammographic breast density assessment: a deep learning perspective. *J Digit Imaging* 2018; **31**(4): 387–92.
  50. Nicholson BT, LoRusso AP, Smolkin M, Bovbjerg VE, Petroni GR, Harvey JA. Accuracy of assigned BI-RADS breast density category definitions. *Acad Radiol* 2006; **13**(9): 1143–9.
  51. 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): 916–22.
  52. Wirth M, Lyon J, Fraschini M, Nikitenko D. The effect of mammogram databases on algorithm performance. In: Proceedings 17th IEEE Symposium on Computer-Based Medical Systems [Internet]. IEEE Comput. Soc [cited 2019 May 5]; 15–20. Available from: <http://ieeexplore.ieee.org/document/1311684/>.
  53. Fonseca P, Mendoza J, Wainer J, et al. Automatic breast density classification using a convolutional neural network architecture search procedure [Internet], 2019 [cited 2019 May 5]. Available from: [http://metis.pucp.edu.pe/~pfonseca/cnn\\_breast\\_density\\_pfonseca.pdf](http://metis.pucp.edu.pe/~pfonseca/cnn_breast_density_pfonseca.pdf)
  54. Kavanagh AM, Giles GG, Mitchell H, Cawson JN. The sensitivity, specificity, and positive predictive value of screening mammography and symptomatic status. *J Med Screen* 2000; **7**(2): 105–10.



55. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015; **175**(11): 1828.
56. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology* 2002; **225**(1): 165–75.
57. Kerlikowske K, Hubbard RA, Miglioretti DL, et al. Comparative effectiveness of digital versus film-screen mammography in community practice in the United States. *Ann Intern Med* 2011; **155**(8): 493.