REVIEW ARTICLE

Deep Learning for Segmentation and Classification in Mammograms for Breast Cancer Detection: A Systematic Literature Review

Raymond Sutjiadi, MS [a,b], *Siti Sendari, PhD* [a,*], *Heru Wahyu Herwanto, PhD* [a], *Yosi Kristian, PhD* [c]

[a] *Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, East Java, Indonesia;* [b] *Department of Informatics, Faculty of Information Technology, Institut Informatika Indonesia Surabaya, Surabaya, East Java, Indonesia;* [c] *Department of Informatics, Faculty of Science and Technology, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, East Java, Indonesia*

Received November 5, 2023; revision received January 20, 2024; accepted February 17, 2024

Abstract: Integrating machine learning into medical diagnostics has revolutionized the field, particularly enhancing Computer-aided Diagnosis (CAD) systems. These systems assist healthcare professionals by leveraging medical data and machine learning algorithms for more accurate diagnosis and treatment plans. Mammography, an X-ray-based imaging technique, is pivotal in early breast cancer detection, enabling the differentiation between benign and malignant lesions. Recent studies have focused on developing deep learning-enabled mammography CAD systems, which have shown promising results in detecting, segmenting, and classifying anomalies in mammogram images. This comprehensive review presents an innovative system architecture for breast cancer detection, segmentation, and classification using deep learning within mammography CAD systems. It also explores publicly available mammogram datasets and the critical parameters for assessing deep learning system performance. The literature review is meticulously conducted using the PRISMA methodology to evaluate and synthesise novel research findings in this domain. This survey highlights the technological advancements and underlines the potential of deep learning in transforming mammographic analysis for breast cancer detection.

Key words: Computer-aided diagnosis; Deep learning; Mammography; Breast cancer; Detection; Classification

Advanced Ultrasound in Diagnosis and Therapy 2024; 03: 094-105 DOI: 10.37015/AUDT.2024.230051

Breast cancer is the most commonly diagnosed cancer among women worldwide, with an estimated 2.3 million new cases or 11.7 % of all cancer cases in 2020 [1]. A high mortality rate of patients follows this high prevalence. In the United States, from 2017 to 2019, approximately 2.5% of women diagnosed with invasive breast cancer died [2]. One method to reduce the fatality rate is diagnosing breast cancer early for appropriate medical treatment. Breast cancer screening is recommended, especially for those with a high risk of developing breast cancer, such as those exposed to radiation over a certain period, a family history of the disease, and older women [3].

One safe, convenient, and effective breast cancer screening method is mammography. Mammography is a medical imaging tool based on X-rays, which is used to observe the dense tissue in the breast to detect the presence of abnormality. The abnormality detected by mammography is the presence of unwanted mass and calcification, which could lead to early signs of breast cancer [4].

Although mammography is still an effective

[*] *Corresponding author: Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, East Java, Indonesia*
e-mail: siti.sendari.ft@um.ac.id

method for breast cancer screening, the reading of a mammogram needs awareness and thoroughness from the radiologist. Reading mammograms is prone to false positives/negatives in reading and locating abnormal mass or calcification. Radiologists still miss 10%-30% of cancers caused by human and technological limitations [5]. The breast contains complex structures of fibrous, glandular, and fatty tissue layers. A mammogram shows healthy and malignant fibrous and glandular tissues as white regions [6]. In contrast, fatty tissues appear as black regions [6]. Women with dense breast structures find it harder to locate abnormalities because cancer may be disguised under normal tissues. On the other hand, women with more fatty tissues tend to locate benign or malignant lesions more easily. It is challenging for radiologists to establish breast cancer diagnosis based on the reading of mammograms. Sometimes, breast cancer diagnosis requires an additional medical examination, such as histopathology, as the gold standard [7], which is inconvenient and expensive for patients.

Computer-aided diagnosis (CAD) uses computers to help doctors establish a diagnosis based on medical examination results. CAD analyzes imaging and/or non-imaging patient data utilising artificial intelligence or machine learning methods to assess the patient's condition and create an assessment, which can help doctors in their decision-making process [8]. In recent years, deep learning methods have been utilized to process medical imaging, such as Ultrasound Imaging [9], X-ray Imaging [10], Computed Tomography Scan (CT-Scan) Imaging [11], and Magnetic Resonance Imaging (MRI) [12], to detect, segment, and classify abnormalities in human body parts. The result of CAD provides a second medical opinion for the doctors to establish a convincing diagnosis from the medical image interpretation [13].

This review aims to cover the use of deep learning-based CAD in the past six years to interpret mammogram results. Furthermore, this article also contains state-of-the-art deep learning-based CAD research using various architectures and models to search for the optimum system's accuracy and performance. In addition, the paper also describes the different public datasets of mammography that are usually used by researchers and compares the specifications among the datasets.

The following questions are addressed in this paper:

• RQ1: What is the novel system architecture of breast cancer detection and classification using deep learning-based CAD in mammography?

• RQ2: What is the accuracy of each deep learning-based CAD using mammograms as the input to detect and classify breast cancer?

• RQ3: What are the parameters to evaluate the system’s performance in detecting and classifying breast cancer based on mammograms?

• RQ4: What public mammography datasets can the researchers use, and what are their respective specifications?

This paper is organised as follows (as shown in Fig. 1):

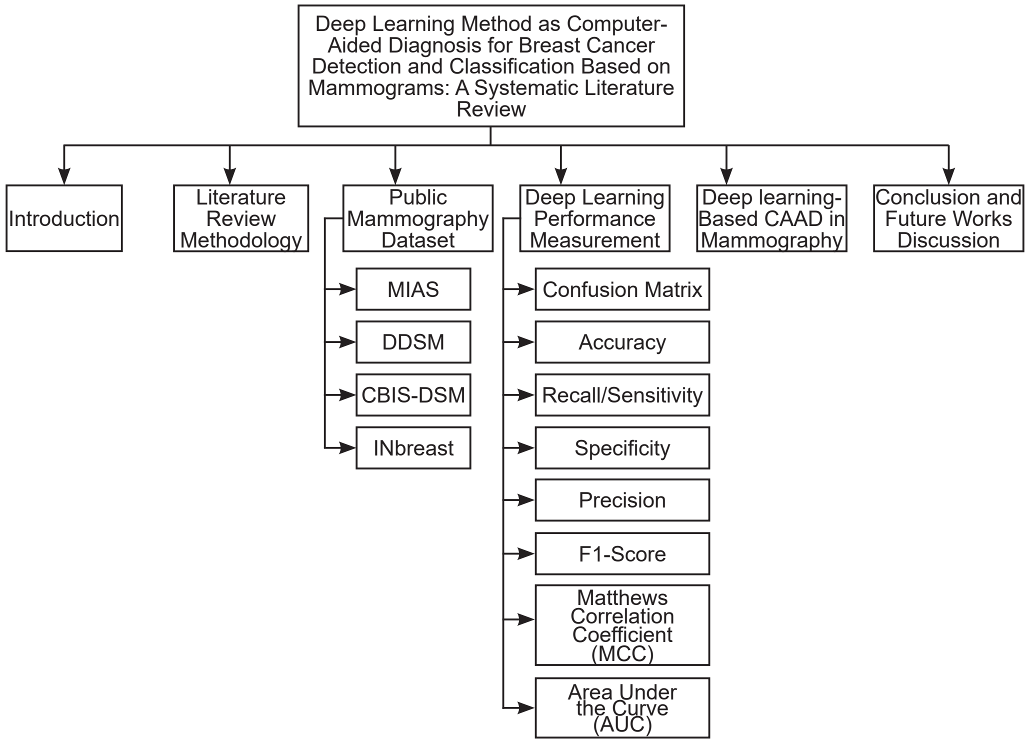

Deep Learning Method as Computer-Aided Diagnosis for Breast Cancer Detection and Classification Based on Mammograms: A Systematic Literature Review
Introduction
Literature Review Methodology
Public Mammography Dataset
Deep Learning Performance Measurement
Deep learning-Based CAAD in Mammography
Conclusion and Future Works Discussion
MIAS
DDSM
CBIS-DSM
INbreast
Confusion Matrix
Accuracy
Recall/Sensitivity
Specificity
Precision
F1-Score
Matthews Correlation Coefficient (MCC)
Area Under the Curve (AUC)

Figure 1 Paper Organisation.

- Section 1 provides the introduction to depict the problem arising in breast cancer diagnosis based on mammograms.
- Section 2 describes the literature review methodology.
- Section 3 presents the variety of public mammography datasets and its specification.
- Section 4 discusses the parameters for measuring the system performance of the deep learning-based CAD in mammography.
- Section 5 explains the novel system architecture of deep learning-based CAD in mammography.
- Section 6 concludes the review and suggests future works to improve deep learning-based CAD in mammography.

Literature Review Methodology

As the source of this literature review, the authors explored several reputable scientific journals and proceedings via Google Scholar, Science Direct, PubMed, and the Institute of Electrical and Electronics Engineers (IEEE) websites. Only literature published during the past six years (2018-2023) is taken and analyze to maintain the novelties. A small portion of cited literature was published more than six years ago because these papers are the original research papers of methodologies or datasets used in this literature review.

Some keywords used to search related papers are "deep learning mammogram", "deep learning breast cancer mammogram", "mammogram breast cancer detection", "mammogram classification detection", and other keyword variations. The authors collect, sort, screen, and analyze the resulting paper according to Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) [14], as shown in Figure 2. Following are some of the exclusion criteria that were used in this review:

- C1: Papers are not scientific research papers.
- C2: Papers are not written in English.
- C3: Papers do not use deep learning methods.
- C4: Papers do not use mammography as the medical image dataset.
- C5: Papers do not present the system's performance measurement.

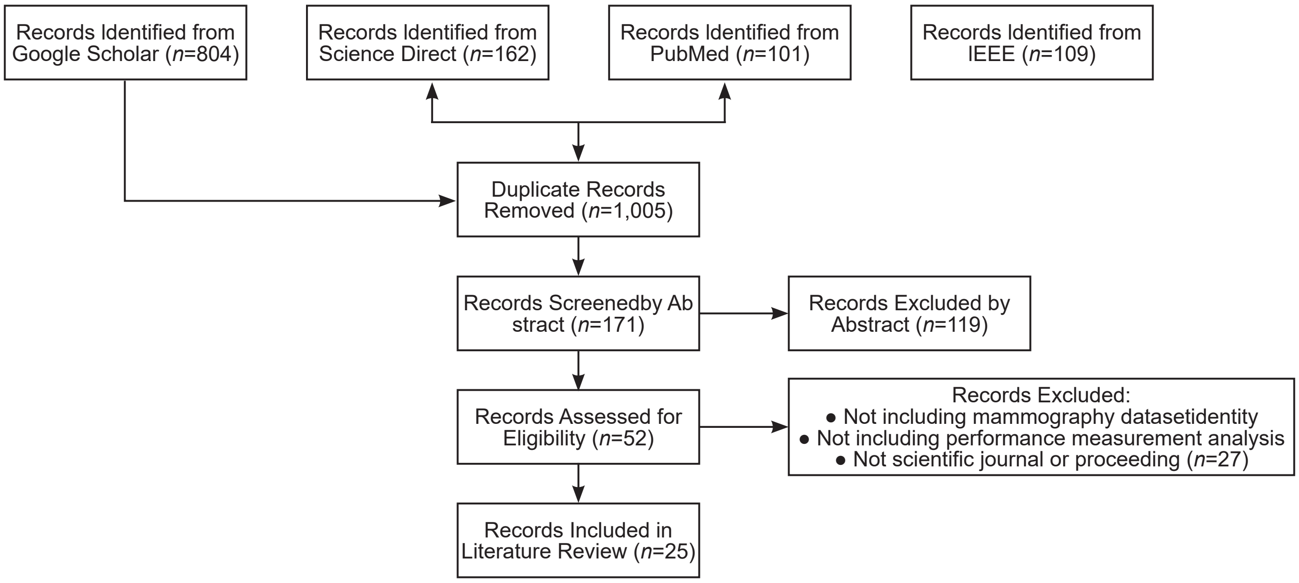

Records Identified from Google Scholar (*n*=804)
Records Identified from Science Direct (*n*=162)
Records Identified from PubMed (*n*=101)
Records Identified from IEEE (*n*=109)
Duplicate Records Removed (*n*=1,005)
Records Screenedby Ab stract (*n*=171)
Records Excluded by Abstract (*n*=119)
Records Assessed for Eligibility (*n*=52)
Records Excluded:
● Not including mammography datasetidentity
● Not including performance measurement analysis
● Not scientific journal or proceeding (*n*=27)
Records Included in Literature Review (*n*=25)

Figure 2 Literature Selection Process According to PRISMA.

Public Mammography Dataset

Deep learning networks are part of the machine learning method [15]. One advantage of deep learning is its availability to learn big data using latent space via feature extraction, which requires lower computational power to process than traditional machine learning by eliminating unnecessary data [16]. In computer vision, deep learning has been utilized to detect, segment, recognise, and classify objects accurately in many areas. The success of deep learning in classifying objects accurately relies on a sufficient dataset to train the model as the source of knowledge [17].

Implementing deep learning in computer-aided diagnosis using medical imaging has a problem of scarcity of datasets to train the deep learning model. This problem arises because of the ethics and confidentiality related to using patient medical records. Luckily, some research organizations have shared anonymised medical imaging for research advancement to contribute to developing intelligent medical imaging technology.

Using deep learning methods to analyze and classify breast cancer via mammography is challenging because of its complexity. Some researchers found that the application of deep learning in mammography results in promising accuracy in the decision-making for the radiologist to diagnose breast cancer. Mammography has two types of image projection, i.e. Mediolateral Oblique (MLO) and Craniocaudal (CC) [18]. The MLO projection is captured from the centre of the chest outward, while the CC projection is captured from above the breast [19], as shown in Figure 3.

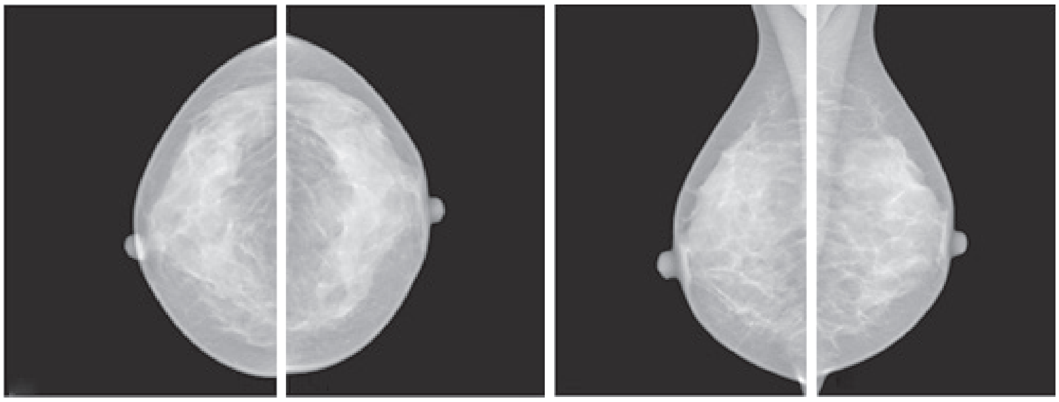


Figure 3 CC and MLO Mammogram Projection [20].

There are public mammogram datasets that the researchers can use for research purposes in the computer vision area. Some of the popular public mammogram datasets are explained below (Table 1):

- Mammographic Image Analysis Society (MIAS) [21]

MIAS is a society in the UK concerned with research in understanding mammogram images. This dataset comprises 322 images of breast imaging from 161 unique patients captured in MLO projection. The format of breast imaging is in Portable Gray Map (PGM) at 50-micron resolution and respective labels (normal, benign, and malignant cases). The total size of the dataset is 1.51 GB in ZIP file format. This dataset is licensed under a CC BY license.

- Digital Database for Screening Mammography (DDSM) [22]

DDSM was compiled by the collaboration of the Massachusetts General Hospital, the University of South Florida, Sandia National Laboratories, Washington University School of Medicine, Wake Forest University School of Medicine, Sacred Heart Hospital, and ISMD Incorporated. Mammograms of 2,620 cases are available in 43 volumes at 42-50-micron resolution. Each volume size varies between 2-6 GB and consists of 4-10 MLO and CC projection mammogram files in lossless JPEG (LJPEG) compression format. Every volume is categorised as normal, cancer, benign, and benign without callback cases.

- Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [23]

CBIS-DDSM is an updated and standardised version of DDSM. This dataset comprises 753 calcification cases and 891 mass cases. The file of mammogram images is in Digital Imaging and Communications in Medicine (DICOM) format, the standard for medical images. Also, this dataset is equipped with metadata in CSV files to inform about the patient's age, the date of image acquisition, the dense tissue category, and other information. Furthermore, instances that have abnormalities include .OVERLAY files that contain the details and types of abnormality (mass or calcification).

Table 1 Mammogram public datasets specification.

Item	MIAS	DDSM	CBIS-DDSM	INbreast
Number of images	322	10,480	10,239	410
Resolution	50 Micron	42-50 Micron	42-50 Micron	70 Micron
Image format	PGM	LJPEG	DICOM	DICOM
Mammograms projection	MLO	MLO and CC	MLO and CC	MLO and CC
Overlay	Yes	Yes	Yes	Yes
Total size	3.16 GB	230.9 GB	163.6 GB	2 GB

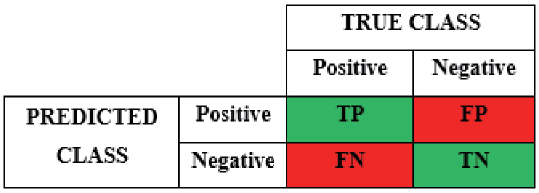
- INbreast [20]

INbreast is a full-field digital (FFD) mammogram dataset that was acquired at the breast centre in Centro Hospitalar de S. João [CHSJ], Portugal. This dataset contains 410 mammogram images at 70-micron resolution from 115 different cases. Furthermore, this dataset is also provided with the information of lesion type (mass, calcification, asymmetry, and distortion) and contours in XML format.

Deep Learning Performance Measurement

Implementing deep learning in a CAD system requires good performance to output a precise diagnosis. Besides huge data samples, deep learning needs parameter optimisation to create robust models [24]. Training deep learning models is time-consuming, expensive, and requires high computational resources. So, the proper performance metric should be utilized to evaluate whether the system is already performing well as expected or still requires further optimisation using the appropriate technique [25].

The result of deep learning-based CAD detection and classification can be assigned into positive or negative classes [26]. When the system returns a positive result and the sample falls under the positive class, the result is labelled as True Positive (TP). If the system returns a negative result and the sample belongs to the negative class, the result is True Negative (TN). Both TP and TN are the correctly classified samples for CAD systems. On the other hand, there is a False Positive (FP) when the system outputs a positive result, but the sample belongs to the negative class. The False Negative (FN) occurs when the system outputs a negative result, but the sample is categorised as a positive class. Neither FP nor FN is expected in CAD systems. To represent the relation among TP, TN, FP, and FN in a deep learning system, the performance of a classification can be visualised using a specialised matrix known as a confusion matrix, also known as an error matrix [27], as shown in Figure 4.



		TRUE CLASS	
		Positive	Negative
PREDICTED CLASS	Positive	TP	FP
	Negative	FN	TN

Figure 4 Confusion Matrix.

The metric to measure the proportion of correctly categorised samples to all samples in the evaluation dataset is called Accuracy (Acc) [28]. The accuracy value range is [0, 1], where 0 denotes incorrectly predicting all positive-negative samples, and 1 denotes correctly predicting all positive-negative samples. The higher accuracy value is expected to indicate the correctness of the system in predicting sample data. This metric is the most commonly used to evaluate the performance of deep learning systems.

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + TN + FP + FN}$$

The other popular metric to measure the performance of deep learning systems is Recall (Rec), sometimes called Sensitivity or True Positive Rate (TPR). This metric shows the ratio between all positive data samples classified correctly and the total data samples predicted as positive class [29]. The recall value resides from 0 to 1 [0, 1], where 0 is the lowest rate denoting incorrectly predicting all positive class samples, and 1 is the highest rate denoting all positive class samples that the systems could classify correctly.

$$Recall\ (Rec) = \frac{TP}{TP + TN}$$

Specificity (Spec) is the opposite of recall. If recall concerns measure positive class samples, the specificity measurement looks up the negative class samples. This measure displays the proportion of all correctly identified negative data samples to the total data samples projected to be in the negative class [30]. The specificity value ranges from 0 to 1 [0, 1], with 0 representing the lowest rate of wrongly predicting all negative class samples and 1 representing the best performance of correctly classifying all negative class samples.

$$Specificity\ (Spec) = \frac{TN}{TN + FP}$$

Precision (Prec) is another metric to measure the ratio between all correctly identified positive data samples and the total data samples identified as positive class by the system classification [29]. The precision value range is [0, 1], where 0 denotes the lowest rate of wrongly predicting all positive class samples, and 1 denotes the highest rate of all positive class samples identified correctly.

$$Precision\ (Prec) = \frac{TP}{TP + FP}$$

Precision and recall are traded off for system performance measurement. The F1 score penalises extreme levels of either precision or recall since they are harmonic means [31]. The F1-score's range is [0, 1], with 0 denoting minimum precision and/or recall and 1 denoting the maximum precision and recall scores.

$$F1 = 2 \times \frac{Prec \times Rec}{Prec + Rec}$$

Matthews Correlation Coefficient (MCC) is a balanced measure of the quality of classifications even if the classes are in different sizes, counting in true and false positives and negatives [32]. MCC can be used to summarise an error or confusion matrix. The MCC value ranges between 1 and -1, where 1 is the best agreement between the predicted and actual value, and -1 is a sign of random prediction according to the actual value.

$$Mcc = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}}$$

Another metric to evaluate the performance of the deep learning system is the Area Under the Curve (AUC). AUC typically refers to the area under the Receiver Operating Characteristic (ROC) curve, as shown in Figure 5. The ROC curve is a graphical representation that illustrates the performance of a binary classification model at various classification thresholds by plotting the sensitivity against specificity for different classification thresholds [33]. The AUC is a metric that quantifies the model's overall performance, considering all possible classification thresholds. It is a useful metric in deep learning for assessing the discriminative power of a model and is commonly used to evaluate the performance of binary classification tasks.

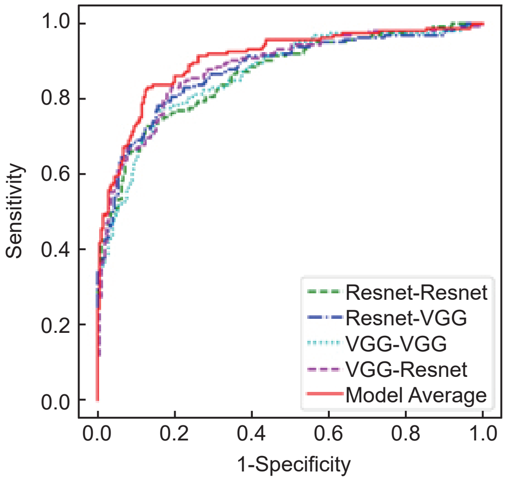


Figure 5 Receiver operating characteristic (ROC) curve [34].

Deep Learning-based CAD in Mammography

Much research in deep learning-based CAD in mammography has been conducted. This literature review scrutinized 25 selected scientific papers which utilized deep learning methods/algorithms. These papers are taken from scientific journals (24 articles) and conference proceedings (1 article) published from 2018-2023.

This paper reviews the various deep learning methods and architectures from the above literature. Also, the literature is classified by the deep learning tasks such as detection, segmentation, classification, or combined tasks. Detection is the implementation of deep learning methods to find objects of interest. Segmentation is a deep learning task to partition the image by building the region to identify the position of the objects of interest. Then, the classification is a task of deep learning to classify the detected objects of interest into some categories.

From Table 2 and Figure 6 above, CNN is positioned as the most used deep learning algorithm in CAD systems based on mammography. 15 papers use CNN, followed by the DCNN used by 5 papers and YOLO used by 2 papers. Then, the other algorithms, i.e., FrCN, Faster R-CNN, TTCNN, Depth-wise CNN, and OMLTS-DLCN, are used only by 1 research paper, respectively. Some papers are classified into multiple categories because they implement more than two deep learning algorithms to build more complex and functional CAD systems. CNN is still the most promising deep learning method to classify objects in digital images. This is because CNN offers the flexibility to capture spatial features from an image, reducing the image dimension to save computation resources, unlike traditional fully connected neural networks.

Table 2 Research papers classified by the deep learning algorithms

No.	Method/Algorithm	Paper
1	Convolutional Neural Network (CNN)	[32], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47]
2	Deep Convolutional Neural Network (DCNN)	[48], [49], [50], [51], [52]
3	Full-resolution Convolutional Networks (FrCN)	[32]
4	Faster R-CNN	[53]
5	Transferable Texture Convolutional Neural Network (TTCNN)	[54]
6	Depth-wise Convolutional Neural Network	[55]
7	Optimal Multi-Level Thresholding-based Segmentation with DL Enabled Capsule Network (OMLTS-DLCN)	[56]
8	You Only Look Once (YOLO)	[32], [57]

Table 3 Research papers classified by the deep learning tasks

No.	Deep learning tasks	Paper
1	Detection	[32], [39], [51], [57]
2	Segmentation	[32], [47]
3	Classification	[32], [34], [35], [36], [37], [38], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [52], [53], [54], [55], [56], [57]

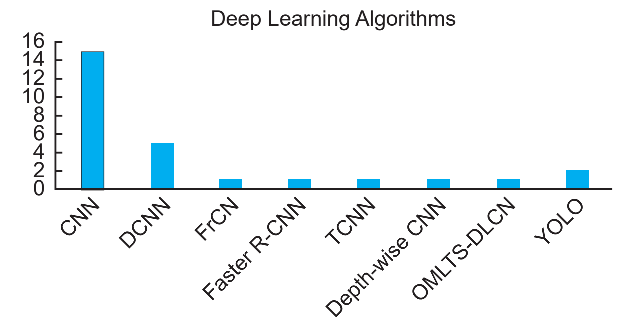


Figure 6 Distribution of deep learning algorithms used by 25 selected papers

From Table 3 and Figure 7 above, most papers (23 papers) utilize deep learning algorithms for classification tasks, classifying mammogram reading by specific categories: normal, benign, or malignant; normal or abnormal; calcification or mass; dense or non-dense; low risk or high risk. In second place, 4 papers use deep learning algorithms only to detect the abnormalities (benign or malignant or detection of architectural distortion) as the object of interest from mammograms. Few papers (2 papers) use a deep learning algorithm to segment the abnormalities from mammograms to show the location that detects abnormalities by drawing the region of interest. Some papers are classified into multiple categories because they implement more than two deep learning tasks to develop complex CAD systems.

Table 4 Research papers classified by the datasets

No.	Dataset	Paper
1	MIAS/Mini-MIAS	[36], [39], [40], [42], [43], [46], [49], [54], [56]
2	DDSM	[38], [39], [41], [49], [52], [54], [55], [56], [57]
3	CBIS-DDSM	[34], [36], [39], [47], [48], [50], [51], [55]
4	INbreast	[32], [34], [38], [39], [40], [44], [45], [51], [53], [54]
5	Private Dataset	[35], [37], [55]

Table 4 and Figure 8 show that the most popular public mammogram dataset is INbreast, which is used by 10 papers. The MIAS and DDSM are used by 9 papers each. CBIS-DDSM is positioned in the next place, used by 8 papers. The other private datasets are used only by 3 papers. INbreast is more popular because the image resolution is the highest compared to the other datasets. The higher resolution provides better clarity, which can improve the CAD system's accuracy in detecting and segmenting malignancies.

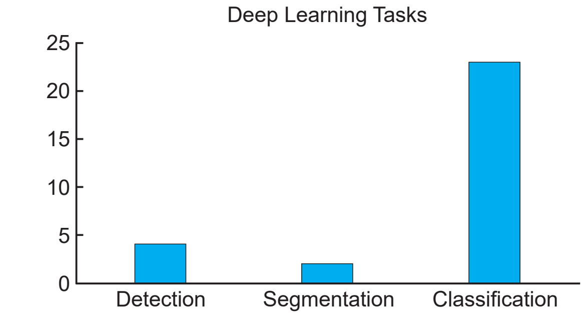


Figure 7 Distribution of deep learning tasks used by 25 selected papers.

Table 5 lists the specification of novel deep learning architectures by examining 25 selected papers. Also, from that table, the system performance metrics for each architecture can be evaluated using several parameters.

The most popular method for developing CAD systems is CNN. This CNN method is usually combined with other deep learning methods for better functionality. In 2018, Al-antari, M.A. et al. utilized the YOLO, FrCN, and CNN methods [32]. YOLO is used for mass detection, FrCN for mass segmentation, and CNN for mass classification functions. The mass detection accuracy is 98.96% and F1-score is 99.24 %; the mass segmentation accuracy is 92.97 % and F1-score is 92.69 %; the mass classification accuracy is 95.64 % and F1-score is 96.84 %. The system accommodates all deep learning tasks to detect, segment, and classify malignancy in mammograms with good value (more than 90 percent) of accuracy and F1-score.

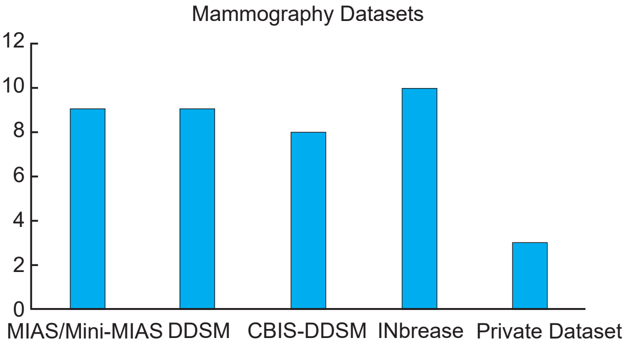


Figure 8 Distribution of mammography datasets used by 25 selected papers

Table 5 Comparison of deep learning architectures for breast cancer detection and performance metrics

No.	Paper	Deep learning architecture	Performance metric
1	Pengcheng, X. et al., 2018 [48]	19-layer VGGNet CNN + 152-layer ResNet (class activation mapping for localising abnormalities)	92.53% Accuracy (CBIS-DDSM Dataset)
2	Jiao, Z. et al., 2018 [49]	10-layer CNN + Metric Learning Layers	97.4% Accuracy (DDSM Dataset) 96.7% Accuracy (MIAS Dataset)
3	Al-masni, M.A. et al., 2018 [57]	26-layer YOLO	99.7% Accuracy of Location Detection and 97% Accuracy of Mass Classification (DDSM Dataset)
4	Al-antari, M.A. et al., 2018 [32]	26-layer YOLO (mass detection) + 16-layer FrCN (mass segmentation) + 7-layer CNN (mass recognition and classification)	Mass Detection 98.96% Accuracy, 99.24% F1-score, 97.62% MCC Mass Segmentation 92.97% Accuracy, 92.69%F1-score, 85.93% MCC Mass Classification 95.64% Accuracy, 96.84% F1-score, 94.78% AUC, 89.91%(INbreast Dataset)
5	Ribli, D. et al., 2018 [53]	Faster R-CNN using 16-layer VGG16 model	95% AUC (INbreast Dataset)
6	Ragab, D.A. et al., 2019 [50]	DCNN using 12-layer AlexNet model	87.2% Accuracy, 94% AUC (CBIS-DDSM Dataset)
7	Shen, L. et al., 2019 [34]	CNN with end-to-end training using combination of 16-layer VGG-16 and 50-layer ResNet50	CBIS-DDSM Dataset 88% AUC (best single model), 91% AUC (four-model averaging) INbreast Dataset 95% AUC (best single model), 98% AUC (four-model averaging)
8	Li, H. et al., 2019 [35]	DenseNet-II Neural Network Model	94.55% Average Accuracy (Private Dataset)
9	Khan, H.N. et al., 2019 [36]	Multi-View Feature Fusion (MVFF): 4-layer Small VGGNet-like using Multi-View ROI as the input	93.2% AUC for mass and calcification classification 84% AUC for malignant and benign classification 93% AUC for normal and abnormal classification (CBIS-DDSM dan Mini-MIAS)
10	Yala, A. et al., 2019 [37]	18-layer ResNet18 and Risk Factor Logistic Regression (RF-LR) model	79% AUC for premenopausal patients & 70% AUC for postmenopausal patients (Private Dataset)
11	Xu, C. et al., 2021 [38]	Multi-Scale Attention Module (MSAM): constructed by stacking multiple MSA bottlenecks.	94.2% AUC (DDSM Dataset), 92.85% AUC (DDSM+INBreast Dataset)
12	Oyelade, O.N. et al., 2021 [39]	12-layer CNN with data augmentation	93.75% Accuracy (DDSM + CBIS, INbreast, and MIAS Dataset), 87.29 % Accuracy (CBIS-DDSM Dataset)
13	AlGhamdi, M. et al. [51]	Dual View-DCNN (DV-DCNN): a 4-layer dense block + neighbourhood patch matching layers with dual view image input	97.5% Accuracy, 95% Sensitivity, 96% Specificity, 98% AUC (CBIS-DDSM) 96% Accuracy, 94% Sensitivity, 95% Specificity, 97% AUC (INbreast)
14	Chouhan, N. et al. [52]	Diverse Features based Breast Cancer Detection (DFeBCD): DCNN (6 highway blocks + 3 fully connected layers) + Support Vector Machine (SVM) / Emotional Learning inspired Ensemble Classifier (ELiEC)	86.1% ROC-AUC (SVM) & 86.5% ROC-AUC (ELiEC) 93.2% PR-AUC (SVM) & 93.4% PR-AUC (ELiEC) 80.5% Accuracy (SVM) & 80.3% Accuracy (ELiEC) (IRMA – DDSM Dataset)
15	El Houby, E.M.F. et. al., 2021 [40]	10-layer CNN with image pre-processing	96.55% Sensitivity, 96.49% Specificity, 96.52% Accuracy, 98% AUC (INbreast Dataset) 98% Sensitivity, 92.6% Specificity, 95.3% Accuracy, 97.4% AUC (MIAS Dataset)
16	Salama, W.M. et al., 2021 [41]	Pre-trained modified U-Net model for segmentation + different deep learning models (InceptionV3, DenseNet121, ResNet50, VGG16, Mobile-NetV2)	98.87% Accuracy, 98.88% AUC, 98.98% Sensitivity, 98.79% Precision, 97.99% F1-Score (MLO DDSM datasets) 99.43% Accuracy, 99.22% AUC, 99.12% Sensitivity, 98.99% Precision, 98.98% F1-Score (MLO and CC DDSM dataset)

No.	Paper	Deep learning architecture	Performance metric
17	Oyelade, O.N. et al., 2022 [42]	Wavelet-CNN-Wavelet with augmented dataset using Generative Adversarial Network (GAN)	99% Accuracy, 99% Recall, 99% Precision, 100% Specificity, 99% F1-Score (MIAS Dataset)
18	Escorcia-Gutierrez, J. et al., 2022 [43]	Automated Deep Learning Based Breast Cancer Diagnosis (ADL-BCD): 34-layer ResNet34	96.07% Accuracy, 95.90% Specificity, 92.15% Recall, 93.54% Precision (MIAS Dataset)
19	Maqsood, S. et al., 2022 [54]	Transferable Texture Convolutional Neural Network (TTCNN) based on deep features of convolutional neural network models (InceptionResNet-V2, Inception-V3, VGG-16, VGG-19, GoogLeNet, ResNet-18, ResNet-50, and ResNet-101)	99.08% Accuracy, 98.96% Specificity, 99.19% Sensitivity (DDSM Dataset) 96.82% Accuracy, 97.68% Specificity, 95.99% Sensitivity (INbreast Dataset) 96.57% Accuracy, 97.03% Specificity, 96.11% Sensitivity (MIAS Dataset)
20	Adedigba, A.P. et al., 2022 [45]	Discriminative Fine-tuning Method using DenseNet & AlexNet	99.8% Accuracy (DenseNet) & 98.8% Accuracy (AlexNet) (INbreast Dataset)
21	Chakravarthy S.R., S. et al., 2022 [44]	18-layer ResNet-18 + Improved Crow-Search Optimized Extreme Learning Machine (ICS-ELM)	97.193% Accuracy (DDSM Dataset), 98.137% Accuracy (MIAS Dataset), 98.266% Accuracy (INbreast Dataset)
22	Rehman, K. et al., 2022 [55]	Depth-wise 2D V-net 64 Convolutional Neural Network	95% Accuracy (PINUM Private Dataset), 97% Accuracy (CBIS-DDSM Dataset), 98% Accuracy (DDSM Dataset)
23	Kavitha, T. et al., 2022 [56]	Optimal Multi-Level Thresholding-based Segmentation with DL-enabled Capsule Network (OMLTS-DLCN): OKMT-SGO (for segmentation) + CapsNet (feature extraction) + BPNN (classification)	98.50% Accuracy (Mini-MIAS Dataset) and 97.55% Accuracy (DDSM Dataset)
24	Elkorany, A.S. et al., 2023 [46]	CNNs (Inception-V3, ResNet50, and AlexNet) + Term Variance (feature selection) + Multiclass SVM (classifier)	97.81% Accuracy (70% training), 98% Accuracy (80% training), 100% Accuracy (90% training) (MIAS Dataset)
25	Bouzar-Benlabiod, L. et al., 2023 [47]	SE-ResNet101 (RoI extraction) + Case-Based Reasoning System/CBR (classification)	86.71% Accuracy, 91.34% Recall (CBIS-DDSM Dataset)

On the other hand, each public mammogram dataset has specific specifications can affect the system's performance. Therefore, choosing the public mammogram dataset is essential for building reliable CAD systems. In 2019, Shen, L. et al. proposed the CNN method equipped with end-to-end training to classify mammogram images, whether the result is normal or cancer detected [34]. This experiment used 2 sources of dataset, i.e., CBIS-DDSM and INbreast. As a result, it reaches an AUC of 88% using the best single model and an AUC of 91% using the four-model averaging (CBIS-DDSM dataset); an AUC of 95 % using the best single model and an AUC of 98% using the four-model averaging (INbreast dataset). As a finding, the INbreast dataset delivers better AUC value than CBIS-DDSM for both the best single model and four-model averaging.

The other experiments implemented multi-view input images to improve the effectiveness and accuracy of the CAD system. This inspired Khan, H.N. et al. (2019) to research the implementation of the CNN algorithm using the Multi-view Feature Fusion (MVFF) technique to combine 4 images taken from 2 projections of mammogram images, MLO and CC, as the input of the CAD system [36]. The output classifies the mammogram reading into three pair of classes: normal-abnormal, mass-calcification, and malignant-benign. According to the experiment result, it achieves 93.2% AUC for mass and calcification classification, 84% AUC for malignant and benign classification, and 93% AUC for normal and abnormal classification using the CBIS-DDSM and MIAS datasets. Similar research conducted by AlGhamdi, M. et al. (2021) utilized Dual-view Deep CNN by combining 2 images of MLO and CC mammogram projection as the input image [51]. This research achieves 98% AUC for CBIS-DDSM and 97% for the INbreast dataset. From these 2 experiments, it can be concluded that the Dual-view Deep CNN method performs better than the CNN method with MVFF.

The deep learning algorithm needs sufficient quantity and variety of datasets to improve the knowledge of the trained model. It inspired Oyelade, O.N. et al. in 2021 to implement the CNN method combined with an augmented dataset technique to detect the presence of architectural distortion in mammogram images [39]. The system accuracy reaches 93.75% using the combination of DDSM+CBIS, INbreast, and MIAS datasets; and 87.29% using CBIS-DDSM only. This research shows that using multiple sources of datasets will provide better system accuracy than using a single dataset only.

Conclusion and Future Works Discussion

The research in deep learning methods CAD has increased in the past few years. This is due to the rapid development of deep learning methods with better accuracy for recognising and classifying images than traditional machine learning. The implementation of deep learning on medical images has a level of accuracy that is good enough to be used as a decision-maker for a radiologist or doctor to diagnose a particular disease or disorder, such as the presence of a tumour, cancer, or other abnormalities.

Breast cancer is one of the leading causes of death among women. Therefore, it is necessary to make a diagnosis as early as possible so that patients can receive immediate treatment and medical action from an early stage. By doing this, it is hoped that the recovery rate for patients will be higher and their life expectancy will increase. One method of diagnosing breast cancer that is easy and comfortable is doing a mammography screening.

On the other hand, reading a mammogram by a radiologist is a complex task. This is because the female breast structure consists of complex tissue. Thus, reading mammograms is prone to misdiagnosis. Breast cancer cannot be detected because it is disguised by healthy tissue. At the same time, healthy breast tissue can be seen as suspicious cancer because it has a similar structure to a malignant mass.

Much research has been conducted to implement CAD using deep learning methods based on mammogram images as a second opinion for the radiologist to establish an accurate diagnosis of whether a patient has benign or malignant breast cancer. Based on existing research, deep learning can accurately diagnose abnormalities in mammogram images based on certain system performance measurement parameters. The success of the deep learning method in detecting and classifying breast cancer in mammogram images is determined by the deep learning model itself. The deep learning method requires a large number of datasets as training data. Even though mammogram datasets are scarce due to confidentiality in protecting patient medical data, some public mammogram datasets can be used by researchers for research purposes.

For future work, synthetic medical image generator algorithms can be used besides the traditional augmented dataset techniques to increase the number and variety of mammogram datasets. The researchers may use promising image generation techniques, like Generative Adversarial Network (GAN) or Denoising Diffusion Probabilistic Model (DDPM), to create synthetic mammogram images. These synthetic medical images can be a solution to increase the performance of deep learning-based mammography CAD systems without violating patient privacy.

Acknowledgment

Not applicable.

Authors' Contributions

Raymond Sutjiadi was responsible for searching the relevant literature and wrote this literature review; Siti Sendari, Heru Wahyu Herwanto, and Yosi Kristian collaborated to complete the literature and suggested valuable revisions. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Data Availability

All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval and consent to participate

Not applicable.

Conflict of interest

The authors have no conflict of interest to declare.

Consent for publication

Not applicable.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021;71:209-249.

[2] Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, et al. Breast Cancer Statistics, 2022. *CA Cancer J Clin* 2022;72:524-541.

[3] Guo F, Kuo YF, Shih YCT, Giordano SH, Berenson AB. Trends in breast cancer mortality by stage at diagnosis among young women in the United States. *Cancer* 2018;124:3500-3509.

[4] Jeun JH, Lee JH, Cho E, Kim SJ, Park EH, Do Byun K. Invasive breast cancer presenting as a mass replaced by calcification on mammography: A report of two cases. *Journal of the Korean Society of Radiology* 2019;80:591-597.

[5] Ekpo EU, Alakhras M, Brennan P. Errors in Mammography Cannot be Solved Through Technology Alone. *Asian Pac J Cancer Prev* 2018;19:291-301.

[6] D. Anyfantis, A. Koutras, G. Apostolopoulos, I. Christoyianni. Breast density transformations using cycleGANs for revealing undetected

findings in mammograms. *Signals* 2023;4:421-438.

[7] Zhang J, Wu J, Zhou XS, Shi F, Shen D. Recent advancements in artificial intelligence for breast cancer: Image augmentation, segmentation, diagnosis, and prognosis approaches. *Semin Cancer Biol* 2023;96:11-25.

[8] Chan HP, Hadjiiski LM, Samala RK. Computer-aided diagnosis in the era of deep learning. *Med Phys* 2020;47:e218-e227.

[9] Huang Q, Zhang F, Li X. Machine Learning in ultrasound computer-aided diagnostic systems: A survey. *Biomed Res In* 2018;2018:5137904.

[10] Visuña L, Yang D, Garcia-Blas J, Carretero J. Computer-aided diagnostic for classifying chest X-ray images using deep ensemble learning. *BMC Med Imaging* 2022;22:178.

[11] Wahab Sait AR, Dutta AK. Developing a deep-learning-based coronary artery disease detection technique using computer tomography images. *Diagnostics (Basel)* 2023;13:1312.

[12] He M, Cao Y, Chi C, Yang X, Ramin R, Wang S, et al. Research progress on deep learning in magnetic resonance imaging-based diagnosis and treatment of prostate cancer: a review on the current status and perspectives. *Front Oncol* 2023;13:1189370.

[13] Kadhim YA, Khan MU, Mishra A. Deep learning-based computer-aided diagnosis (CAD): applications for medical image datasets. *Sensors (Basel)* 2022;22:8999.

[14] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.

[15] Pranolo, A, Mao, Y, Wibawa, AP, Utama, ABP, Dwiyanto, FA. Optimized three deep learning models based-PSO hyperparameters for Beijing PM2.5 prediction. Knowledge *Engineering and Data Science* 2022.

[16] Li H, Shen HW. Local latent representation based on geometric convolution for particle data feature exploration. *IEEE Trans Vis Comput Graph* 2023;29:3354-3367.

[17] Taye, MM. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers* 2023;12.

[18] Sweeney RI, Lewis SJ, Hogg P, McEntee MF. A review of mammographic positioning image quality criteria for the craniocaudal projection. *Br J Radiol* 2018;91:20170611.

[19] Mohamed AA, Luo Y, Peng H, Jankowitz RC, Wu S. Understanding clinical mammographic breast density assessment: a deep learning perspective. *J Digit Imaging* 2018;31:387-392.

[20] Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. *Acad Radiol* 2012;19:236-248.

[21] Suckling J.Mammographic Image Analysis Society (MIAS) database v1.21. [Dataset]. *Apollo-University of Cambridge Repository* 2015.

[22] Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer P. The digital database for screening mammography. *5th International Workshop on Digital Mammography Toronto* 2001:212-218.

[23] Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* 2017;4:170177.

[24] Cui R, Wang L, Lin L, Li J, Lu R, Liu S, et al. Deep learning in barrett's esophagus diagnosis: current status and future directions. *Bioengineering (Basel)* 2023;10:1239.

[25] Dewandra ARF, Wibawa AP, Pujianto U, Utama ABP, Nafalski A. Journal unique visitors forecasting based on multivariate attributes using CNN. *International Journal of Artificial Intelligence Research* 2022;6.

[26] Orozco-Arias S, Piña JS, Tabares-Soto R, Castillo-Ossa LF, Guyot R, Isaza G. Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. *Processes* 2020;8.

[27] Duntsch I, Gediga G. Confusion matrices and rough set data analysis. *J Phys Conf Ser* 2019;1229.

[28] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 2021;8:53.

[29] Fränti P, Mariescu-Istodor R. Soft precision and recall. *Pattern Recognit Lett* 2023;167:115-121.

[30] Steiner JM, Morse C, Lee RY, Curtis JR, Engelberg RA. Sensitivity and specificity of a machine learning algorithm to identify goals-of-care documentation for adults with congenital heart disease at the end of life. *J Pain Symptom Manage* 2020;60:e33-e36.

[31] Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep* 2022;12:5979.

[32] 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.

[33] Mrudula Devi K, Venkata Ramakrishna S, Rama Koteswara Rao G, Prasad C. Gradient-based optimization of the area under the minimum of false positive and false negative functions. *2nd International Conference on Smart Electronics and Communication (ICOSEC)* 2021;779-785.

[34] Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. *Sci Rep* 2019;9:12495.

[35] Li H, Zhuang SS, Li DA, Zhao JM, Ma YY. Benign and malignant classification of mammogram images based on deep learning. *Biomed Signal Process Control* 2019;51:347-354.

[36] Khan HN, Shahid AR, Raza B, Dar AH, Alquhayz H. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access* 2019;7:165724-165733.

[37] Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 2019;292:60-66.

[38] Xu CB, Lou M, Qi YL, Wang YM, Pi JD, Ma YD. Multi-Scale Attention-Guided Network for mammograms classification. *Biomed Signal Process Control* 2021;68.

[39] Oyelade ON, Ezugwu AE. A deep learning model using data augmentation for detection of architectural distortion in whole and patches of images. *Biomed Signal Process Control* 2021;65.

[40] El Houby EMF, Yassin NIR. Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks. *Biomed Signal Process Control* 2021;70.

[41] Salama WM, Aly MH. Deep learning in mammography images segmentation and classification: Automated CNN approach. *Alexandria Engineering Journal* 2021;60:4701-4709.

[42] Oyelade ON, Ezugwu AE. A novel wavelet decomposition and transformation convolutional neural network with data augmentation for breast cancer detection using digital mammogram. *Sci Rep* 2022;12:5913.

[43] Escorcia-Gutierrez J, Mansour RF, Beleño K, Jiménez-Cabas J, Pérez M, Madera N, et al. Automated deep learning empowered breast cancer diagnosis using biomedical mammogram images. *Computers, Materials & Continua* 2022;71:4221-4235.

[44] Chakravarthy SRS, Rajaguru H. Automatic detection and classification of mammograms using improved extreme learning machine with deep learning. *IRBM* 2022;43:49-61.

[45] Adedigba AP, Adeshina SA, Aibinu AM. Performance evaluation

of deep learning models on mammogram classification using small dataset. *Bioengineering* 2022;9:161.

[46] Elkorany AS, Elsharkawy ZF. Efficient breast cancer mammograms diagnosis using three deep neural networks and term variance. *Sci Rep* 2023;13:2663.

[47] Bouzar-Benlabiod L, Harrar K, Yamoun L, Khodja MY, Akhloufi MA. A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification. *Comput Biol Med* 2023;163:107133.

[48] Xi PC, Shu C, Goubran R. Abnormality Detection in Mammography using Deep Convolutional Neural Networks. *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* 2018;1-6.

[49] Jiao ZC, Gao XB, Wang Y, Li J. A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognit* 2018;75:292-301.

[50] Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ* 2019;7:e6201.

[51] AlGhamdi M, Abdel-Mottaleb M. DV-DCNN: Dual-view deep convolutional neural network for matching detected masses in mammograms. *Comput Methods Programs Biomed* 2021;207:106152.

[52] Chouhan N, Khan A, Shah JZ, Hussnain M, Khan MW. Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography. *Comput Biol Med* 2021;132:104318.

[53] Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. *Sci Rep* 2018;8:4165.

[54] Maqsood S, Damasevicius R, Maskeliunas R. TTCNN: A breast cancer detection and classification towards computer-aided diagnosis using digital mammography in early stages. *Applied Sciences* 2022;12:3273.

[55] Rehman KU, Li J, Pei Y, Yasin A, Ali S, Saeed Y. Architectural distortion-based digital mammograms classification using depth wise convolutional neural network. *Biology (Basel)* 2021;11:15.

[56] Kavitha T, Mathai PP, Karthikeyan C, Ashok M, Kohar R, Avanija J, et al. Deep learning based capsule neural network model for breast cancer diagnosis using mammogram images. *Interdiscip Sci* 2022;14:113-129.

[57] Al-Masni MA, Al-Antari MA, Park JM, Gi G, Kim TY, Rivera P, 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.