Computers in Biology and Medicine 131 (2021) 104248

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Convolutional neural networks for breast cancer detection in mammography: A survey

Leila Abdelrahman [a], Manal Al Ghamdi [b], Fernando Collado-Mesa [c], Mohamed Abdel-Mottaleb [a,*]

[a] University of Miami, Department of Electrical and Computer Engineering, Memorial Dr, Coral Gables, FL, 33146, USA
[b] Umm Al-Qura University, Department of Computer Science, Alawali, Mecca, 24381, Saudi Arabia
[c] University of Miami Miller School of Medicine, Department of Radiology, 1115 NW 14th Street Miami, FL, 33136, USA

ARTICLE INFO

Keywords:
Mammography
Computer-aided detection
Deep learning
Convolutional neural networks

ABSTRACT

Despite its proven record as a breast cancer screening tool, mammography remains labor-intensive and has recognized limitations, including low sensitivity in women with dense breast tissue. In the last ten years, Neural Network advances have been applied to mammography to help radiologists increase their efficiency and accuracy. This survey aims to present, in an organized and structured manner, the current knowledge base of convolutional neural networks (CNNs) in mammography. The survey first discusses traditional Computer Assisted Detection (CAD) and more recently developed CNN-based models for computer vision in mammography. It then presents and discusses the literature on available mammography training datasets. The survey then presents and discusses current literature on CNNs for four distinct mammography tasks: (1) breast density classification, (2) breast asymmetry detection and classification, (3) calcification detection and classification, and (4) mass detection and classification, including presenting and comparing the reported quantitative results for each task and the pros and cons of the different CNN-based approaches. Then, it offers real-world applications of CNN CAD algorithms by discussing current Food and Drug Administration (FDA) approved models. Finally, this survey highlights the potential opportunities for future work in this field. The material presented and discussed in this survey could serve as a road map for developing CNN-based solutions to improve mammographic detection of breast cancer further.

1. Introduction

In the US, approximately one in eight women are diagnosed with breast cancer in their lifetime. Breast cancer is the most common non-cutaneous cancer diagnosed in women: in 2020, researchers predict an estimated 276,480 women are diagnosed, of whom they expect 42,170 to die [1]. Worldwide, breast cancer is the most common non-cutaneous cancer in women, with over two million annual diagnoses [2] .

Scientists developed screening mammography to provide early detection and to save lives from breast cancer. Observational studies show a mortality reduction of about 40% [3,4] after mammography screening. Screening mammography does, however, have limitations. Clinicians will recommend about 15 in 1000 women screened with mammography for a needle biopsy, and 10 to 13 of those biopsies will show that cancer is not present (false positives) [5]. Kuhl et al. [6] highlight one major limitation of mammography screening: prognostically significant breast cancers are underdiagnosed.

Researchers and clinicians have implemented several strategies to improve screening mammography's performance, including double-reading, screening at yearly interval [7], obtaining two views per breast [8], and analyzing prior mammograms for comparison [9]. Radiologists aim to detect critical features like microcalcifications (MCs), architectural distortions (ADs), and asymmetries as biomarkers for cancer or cancer risk. Manually detecting these features leads to additional economic costs and strains on an already scarce breast imaging radiologist workforce [10]. Computer-aided detection systems (CAD) emerged in the 1990s to automatically detect and classify breast lesions in mammograms. Still, these traditional CAD systems fail to significantly improve screening performance, mainly due to their low specificity [11,12]. Specificity connects to how well algorithms detect and classify abnormalities in mammograms. This differs from diagnosis, which involves causal inference regarding an abnormality's origin. Detecting

* Corresponding author.
E-mail address: mottaleb@miami.edu (M. Abdel-Mottaleb).

https://doi.org/10.1016/j.compbiomed.2021.104248
Received 20 November 2020; Received in revised form 8 January 2021; Accepted 25 January 2021
Available online 9 February 2021

anomalies in mammographs is essential.

Recently, Trister et al. [13] report that novel algorithms based on Convolutional Neural Networks (CNNs) improve screening mammography's performance and increase breast imaging radiologists' efficiency. Researchers have developed several CNN-based algorithms for automated mammographic analysis, some of which the US Food and Drug Administration (FDA) has approved [14].

1.1. Survey objectives

This article aims to present a comprehensive and critical analysis of recent literature on CNN applications to mammography. Prior surveys broadly scope Deep Learning CAD for breast cancer [15–18]. However, our survey reevaluates the present landscape with a focused structure on mammography; it examines the background of foundational CNN architectures before highlighting recent CNN-based methods on mammographic data. This paper further offers a systematic analysis of the latest advances in supervised, unsupervised, and semi-supervised learning by first reviewing fundamental and commonly-cited computational concepts in the CNN CAD-mammography domain. Previous surveys [16] have aggregated modalities ranging from MRIs to histology samples or focus broadly on statistical methods [17]; our survey presents CNN applications focused on mammography literature. It is the first survey to break down and evaluate papers as they address four distinct mammographic tasks: (1) density classification and (2) asymmetry detection and classification for risk assessment and (3) calcification detection and classification, and (4) mass detection and classification for cancer diagnosis. The survey explicitly lists recent CAD challenges for mammography and calls for future scientists to address these issues through open areas for research.

After analyzing completed works encompassing these four tasks, this paper presents a quantitative comparison of reported results from the literature; this analysis allows us to point out open areas for additional research. This article further shows how many of the CNN insights in academia have translated into commercial, FDA-approved devices used by thousands of radiologists in the clinic.

1.2. Article selection criteria

We screen for articles published in English between 2011 and 2020 on developing CNNs for mammography, including those presenting available datasets or composing CNNs for mammography. We search for articles through PubMed, Institute of Electrical and Electronics Engineers (IEEE), and Google Scholar databases. Moreover, we reference articles on the subject that were published in English before 2011 for background context.

1.3. Survey topology

First, traditional CAD systems' history is presented in Section 1. CNN architecture and core concepts for training CNNs are discussed in Section 2. In Section 3, available mammography datasets are presented. Then, algorithms developed for the four tasks mentioned above are critically analyzed in Section 4. We discuss quantitative results for these in Section 5. The pros and cons of the CNN-based approaches are discussed in Section 6. How these findings translate into FDA-cleared Deep Learning-based CAD systems is presented in Section 7. Finally, open areas for research on these subjects are presented in Section 8.

2. Traditional CAD systems

Researchers first implemented CAD systems in the 1990s based on conventional machine learning models to find or localize abnormal or suspicious regions and alert clinicians for attention [19]. CAD's primary goal is to increase the detection rate of diseased regions while reducing the false-negative rate, possibly due to observers' mistakes or fatigue.

CAD for breast cancer detection uses different types of clinical medical imaging. Images range from X-ray based mammography to ultrasounds and magnetic resonance imaging (MRI). For this article, we choose mammography for extensive review. Mammography is the most common type of breast cancer screening method: over 65% of women over 40 have had at least one mammogram within two years after 2017 [20]. After doctors find anomalies in a mammogram, they order different imaging techniques, such as an MRI [21]. For high-risk individuals or those with apparent masses in their radiograms, physicians call for a more invasive histological image analysis [21]. Mammograms are, in comparison, minorly-invasive. Because it is a first-line screening method, mammographic analysis merits significant improvement with CAD methods.

The traditional CAD pipeline, as Fig. 1 shows, is: **1)** Algorithms first detect candidate regions through image processing techniques that often rely on an exhaustive set of handcrafted features (texture, shape, gray-level intensity) that experts spend painstaking hours to extract (Fig. 1 A). **2)** They then construct a morphological or statistical set of features representing the candidate regions (Fig. 1 B). **3)** A statistical classifier uses the engineered features to output a probability of, or predict, a disease state (Fig. 1C and D). Recently, researchers have exchanged traditional CAD for CNN-based methods [22]. Before Deep Learning and CNNs' popularity, Support Vector Machine (SVM), Random Forest (RF), and other statistical methods [23,24] were popular, yet their reliance on time-intensive feature engineering has propelled researchers towards CNN-based methods, or hybrid methods combining statistical models with CNN models [25].

3. Deep convolutional neural networks

CNNs are powerful for analyzing images because pattern-specific filters can preserve the picture's spatial features. CNNs often outperform dense neural network methods because fully connected architectures flatten input images, overlooking vital spatial patterns. Fig. 2 details the CNN architectures, and Fig. 2 A) highlights how CNN-based methods rely heavily on end-to-end learning. CNNs emphasize raw input data with minimal feature engineering for mammographic classification, among other tasks. Pattern-specific feature maps generate feature maps; shallow layers in the network detect features resembling the input image, but deeper layers detect abstract patterns, as shown in Fig. 2 B). One filter may pick up vertical lines in a picture, while another specializes in seeing horizontal lines. Depending on the task, feature maps can be flattened into dense (fully-connected) layers for classification tasks. AlexNet [26] popularized CNNs by applying them to the ImageNet challenge, reaffirming their strong performance on large, high-resolution image datasets even at low numbers of parameters. Since then, researchers successfully use these base architectures for mammographic image segmentation and classification. The following sections detail key CNN components and their applications to mammography CAD.

3.1. Optimizers

In order to numerically minimize model loss functions with backpropagation, research rely on optimizers. Backpropagation (as shown in Fig. 2 A) updates model weights based on errors relative to ground truth images (the Loss in Fig. 2 C). Researchers have used various optimizers, like stochastic gradient descent (SGD) [27], the adaptive gradient algorithm (AdaGrad) [28], and adaptive moment (Adam) [29]. In practice, researchers [30–34] still use the classic SGD, although dynamic optimizers like Adam are becoming increasingly popular [31,33,35–40]. Adam dynamically updates the learning rate using gradient momentum.

3.2. Max and average pooling

Large feature map channels with basic convolutional models yield

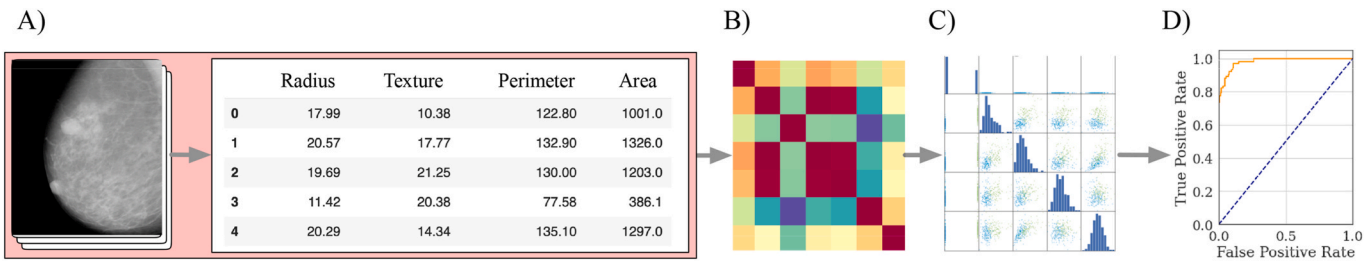

Fig. 1. The Traditional CAD process. A) Researchers handcraft features describing the image. B) Correlation matrices and other filtering is used to select relevant features. C) A statistical model uses the cleaned features to then D), make inferences, as in the receiver operator curve. Best in color.

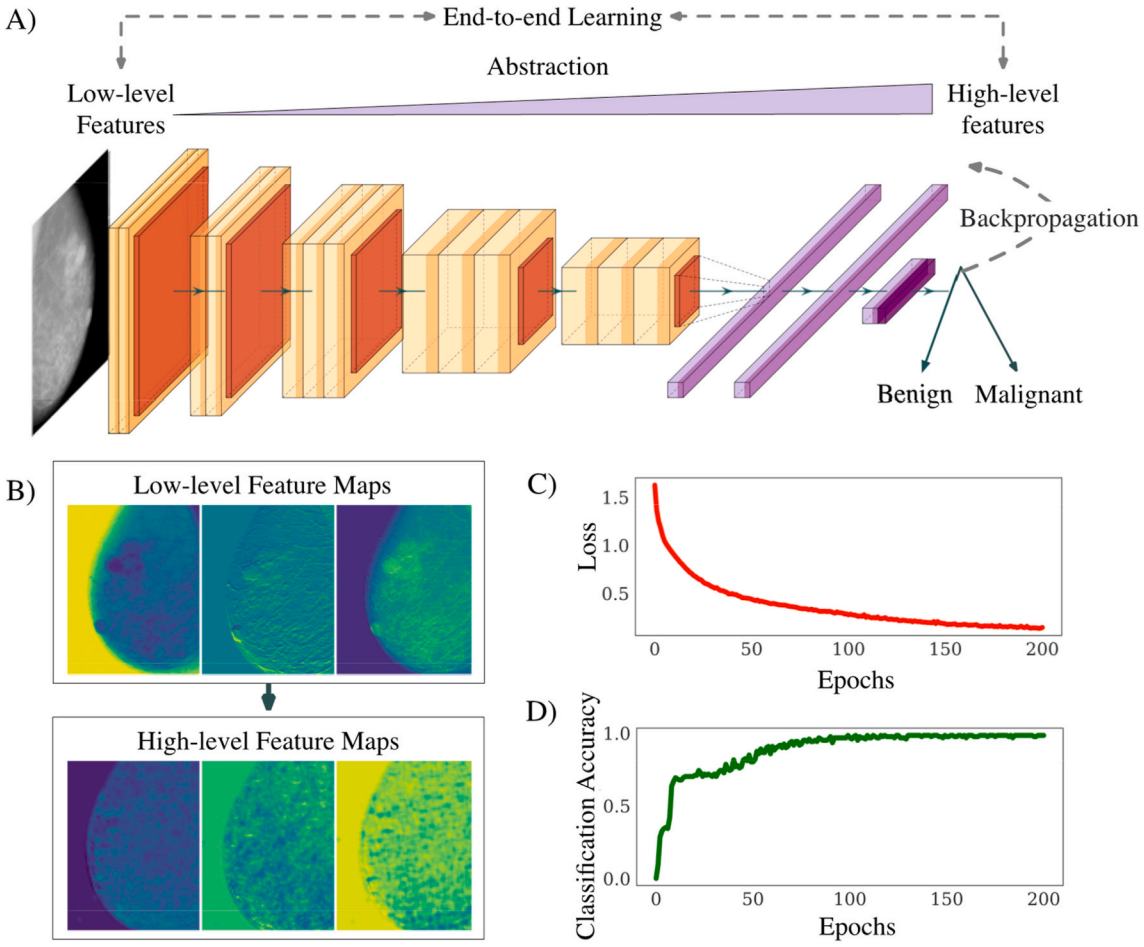

Fig. 2. A) CNN Architecture and Model Learning. Multiple filters generate different feature maps, shown as stacks of orange. The dark orange represents output after pooling. The architecture also flattens the layers into vectors (purple) before classification. The model updates the feature map parameters based on the loss function (which can be a measure for error) through backpropagation. The feature maps are shown growing in width as they are stacked after each convolution along the x-axis. This figure was partially generated with the PlotNeuralNet GitHub [150] repository. B) As the model deepens, it abstracts higher-level features from prior feature maps. While radiologists can often detect patterns in low-level features, convolutions can detect elusive patterns in the high-level features, one of the driving forces behind CAD. C) A sample loss function plot that shows how models minimize errors during each training iteration (epoch) as they learn. As the model learns, it updates the feature maps for the learning task. D) As the models learn by minimizing the loss, the classification accuracy increases. Best seen in color.

slow and computationally expensive training. To resolve this issue, researchers apply max and average pooling to the output of convolutional layers. Max-pooling takes a 2D window's maximum pixel value, whereas average pooling takes a 2D window's average of all pixel values. Pooling summarizes feature maps into encodings with reduced dimensions, as shown in Fig. 2 by the dark orange tensors. These operations reduce the number of trainable parameters to mitigate overfitting and computation time.

3.3. Activation functions

Activation functions are at the core of CNNs, as they introduce non-linearity to the neuronal output. Sigmoidal activation functions are typically used at the final classifier layers by researchers in mammography CAD [41–43], specifically for binary classification tasks, but also have applications in segmentation networks. In practice, researchers use the SoftMax [44] function for multiclass problems, and the final purple layer in Fig. 2 A) shows a typical SoftMax layer implementation. The ReLU activation function [45] is extremely efficient, as it involves

minimal mathematical operations, and thus researchers [42,46–49] often include it in their hidden layers. The leaky ReLU [50] improves upon the standard ReLU by introducing a small hyperparameter, and mammography researchers [47,51,52] have integrated it into their CNN hidden layers.

3.4. Dropout

Dropout [53] randomly pruning nodes in the model at each training epoch. The model learns to compensate for ablated neurons by randomly setting all the weights in nodes to zero. This random ablation reduces model overfitting, allowing models to generalize insights to the validation and test data with higher accuracy. In practice, researchers [30,31,33,48,54,55] show how this leads to robust CAD models that perform better on domain-specific or even cross-domain datasets.

3.5. Batch normalization

Batch normalization [56] reduces activation function reliance on the parameter scales or their initial values and induces self-regularization; it normalizes extremely high values within a smaller domain, and researchers [31,33,35,39,40,46,57] rely on it. Batch-normalized models work with higher learning rates, as each layer learns a more stable parameter distribution at a faster pace.

3.6. Deep CNNs with skip connections

Skip connections link network layers to provide alternative paths for backpropagation. These can generate efficient models with low space and time complexity. ResNets [58], U-Nets [59], DenseNets [60], and Squeeze-Excitation Networks [61] all rely on skip connections. Fig. 3 illustrates different examples of commonly used skip connection-based blocks and models. A block is a collection of repeating neural network layer motifs. Skip connections link earlier to later network layers, preserving important information propagation through deep networks.

3.6.1. ResNets

He et al. [58] introduced Residual networks (ResNets): They contain residual blocks connected by shortcut connections; these connections facilitate inputs to propagate information to future layers, skipping over weighted activations. The skip connection allows researchers to avoid costly computing gradients for these layers during backpropagation, drastically improving performance. These residual blocks often contain stacked convolutional layers at their core. The outputs of the feed-forward layers, $f(x)$, are added to the skip connection values, x, to yield a final prediction, $h(x)$. Framing the model weight optimization problem in this fashion, Equation (1) below provides a method to determine the $f(x)$ from x and $h(x)$.

$$f(x) = h(x) - x \tag{1}$$

3.6.2. U-Net

The U-Net has both CNN layers in the basic building blocks and skip connections. First proposed by Ronneberger et al. [59], the model architecture is different in that it contains both a contracting and expanding path. The contracting connections generate smaller and smaller feature maps before reaching a bottleneck. Upsampling then expands successive feature maps to an output map. The cross-connections in the U-Net concatenate information from the downsampled feature maps in the contracting path to the corresponding feature maps in the expanding path.

3.6.3. DenseNet

The DenseNet [60] connects each convolutional layer to every other layer in a feed-forward fashion. Dense blocks differ from traditional residual blocks in that they involve tensor concatenation rather than tensor addition. Not only are the previous layer's feature maps used as input for the subsequent layer in the DenseNet, but the feature maps of all preceding layers also serve as inputs. The increased number of connections has many advantages. Primarily, it promotes lower-level features detected earlier in the network to contribute to feature spaces deeper in the network.

3.7. Types of learning

3.7.1. Supervised methods

When there are large amounts of well-annotated data, supervised learning approaches are ideal. There are generally two subclasses of supervised learning: classifiers and regressions. Classifiers are appropriate when the output labels are discrete. For instance, the task of classifying and labeling a tumor by its type would require supervised methods. Most clinical tasks with labeled data worked with supervised learning methods where the ground truths are known.

3.7.1.1. Transfer learning. Transfer learning reduces the necessary training time on domain-specific data. Models are most commonly trained on large, annotated datasets such as MSCOCO [62] and ImageNet [63]. Through pretraining, CNNs can transfer learned high-level features from one domain to another [64]. Researchers often use a pretrained backbone like ResNet [65], VGGNet [58], or Inception [66], and fine-tune these backbones by removing hidden layers or unfreezing certain layers and further training on the domain dataset, among other

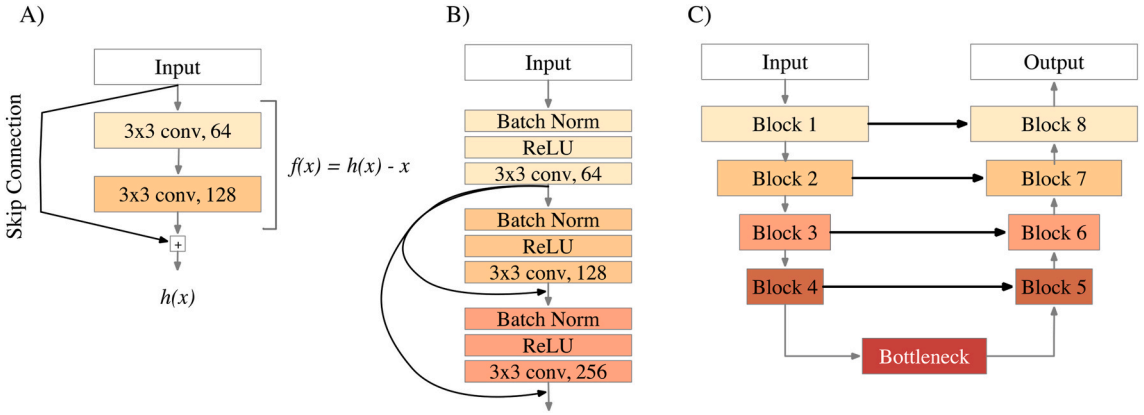

Fig. 3. Different types of skip connections. A) The basic residual block adds the skip connection outputs to the hidden layer output. B) The Dense Block concatenates the first block's output with each subsequent block's output. C) The U-Net concatenates downsampled layers with the parallel upsampled layers. Gray arrows represent sequential connections between layers. All skip connections are shown as black arrows. Best seen in color.

approaches. Pretrained backbones aid CNNs in mammography because datasets are often limited in size. Pretraining, followed by fine-tuning, allows the model to detect abstract features in the domain dataset while also learning new domain-specific features to succeed in tasks ranging from asymmetry detection to calcification detection in a mammogram.

3.7.2. Unsupervised methods

When datasets are entirely unlabeled unsupervised, CNNs can still achieve impactful results. Tasks optimal for unsupervised learning range from dimensionality reduction to segmentation [67]. Autoencoders [68] optimize low-dimensional codings by minimizing the reconstruction error. Researchers often use stacked autoencoders (SAEs) in end-to-end machine learning approaches for deep feature extraction and segmentation. Deep convolutional autoencoders consist of convolutional layers that encode feature maps into z. The decoder expands the bottleneck representation downstream to reconstruct the original image, as Fig. 4 shows. Unsupervised semantic segmentation is extremely powerful in breast cancer CAD, as well-annotated datasets are sparse.

3.7.3. Semi-supervised methods

Researchers use semi-supervised learning when there are many unlabeled sample instances but few labeled instances. With few "gold standard" heavily annotated datasets, the mammography domain is ripe for semi-supervised learning.

3.7.3.1. Generative networks. Models like variational autoencoders (VAEs) and generative adversarial networks (GANs) succeed at data augmentation and sample generation [69]. Researchers [69,70] investigated their role in data augmentation for generating robust mammography datasets. Fig. 5 demonstrates a GAN architecture outline, highlighting how this network succeeds in novel image generation.

Goodfellow et al. [71] first proposed GANs. The architecture consists of an autoencoder based generator trained to synthesize increasingly higher quality images. These images fool the discriminator network trained on ground-truth images. Fig. 2 highlights this sample generation method's key conceptual points. Another important advancement in semi-supervised GANs is Deep Convolutional GANs (DCGANs), first proposed by Radford et al. [72]. These models enhance original GAN architecture by using batch normalization and removing fully connected hidden layers.

3.7.3.2. Attention mechanism. Another increasingly popular semi-supervised method is incorporating attention into classifiers; attention is founded on the spatial transformer architecture, first proposed by Jaderberg et al. [73], and later refined by Wang et al. [74]. Researchers [75–78] are use attention modules in their architecture because it forces the model to focus on the necessary features in an image to make a decision. They choose attention-aware features based on their context to other feature maps in the network, greatly enhancing signal propagation through the network.

More recent advances such as self-attention mechanisms [79] highlight how input images can contextualize information with each other to provide additional context. Attention mechanisms are powerful in series data, such as slice sequences of digital breast tomosynthesis (DBT) or patient mammograms taken at multiple time points.

4. Public mammography datasets

Researchers apply techniques in our discussions on CNNs to many widely available mammography datasets. Table 1 summarizes the most commonly used public datasets, as well as critical details regarding the year they were published, the image format, image views (Mediolateral Oblique (MLO) and Craniocaudal (CC)), resolution, and annotations. The table includes datasets primarily with Full-Field Digital Mammograms (FFDM)s, which involve electronic processing of x-rays. Additionally, Fig. 6 shows selected sample images from these datasets and Fig. 7 highlights dataset usage over time.

4.1. MIAS

The Mammographic Image Analysis Society (MIAS) [80] has more than 300 screening mammograms. The dataset contains annotations for background tissue type (dense/fatty), the abnormality present in the breast (masses, asymmetry), and the abnormality's severity (benign/-malignant). Mammograms with lesions have recorded X and Y coordinates. It also contains labels regarding MCs, ADs, asymmetry, and healthy images.

4.2. The DDSM

The digital database of screening mammography (DDSM) [81] is an archive of over 2600 scanned film mammography studies. A subset of the DDSM is the curated breast imaging subset of the DDSM (CBIS-DDSM), and it includes well-annotated and labeled images. The dataset includes information related to bounding boxes for region of interests (ROIs), as well as detailed pathological information regarding breast mass type, tumor grade, and stage. The dataset consists primarily of scanned film-screen mammography, far behind most advanced imaging techniques like FFDM and DBT.

4.3. INBreast

The INBreast Database [82] is a collection of over 410 screening mammographs from more than 115 subjects. The data includes

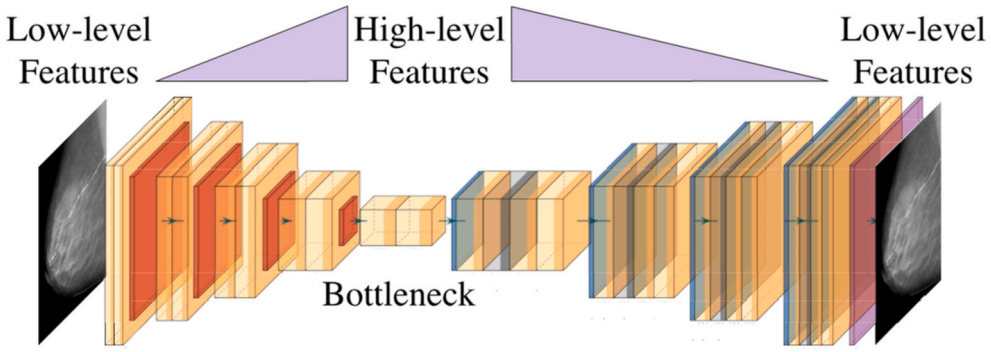

Fig. 4. The autoencoder's image reconstruction still preserves important feature embeddings and compresses low-level feature maps to high-level representations. The light orange represents convolutional layers, dark orange is max pooling, blue is the de-convolutional layers, and purple is the soft-max layer. We partially generate the figure with the PlotNeuralNet GitHub [147] repository. Best seen in color.

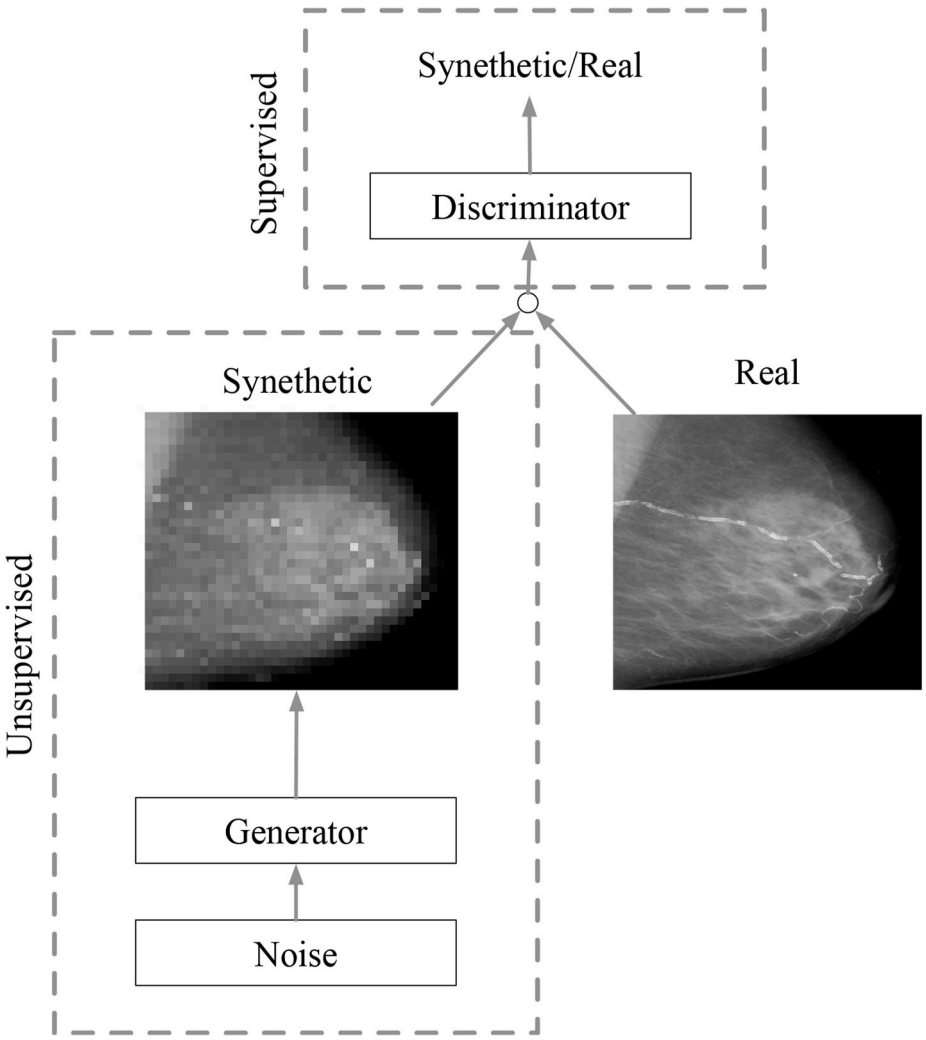

Fig. 5. GAN Architecture. The GAN has a generator (an autoencoder) that synthesizes images by initially sampling from a noisy distribution. This is the unsupervised kind of learning. In the supervised portion, a discriminator determines if this synthetic image is actually of the same class as the real, labeled image. As the discriminator penalizes for poor-quality synthesized images, the generator learns to produce more realistic images.

information regarding abnormality types and mass contour information. Although openly available for download, the Universidade do Porto has recently stopped supporting the image database, yet it is available upon request.

4.4. The BCDR

The Breast Cancer Digital Repository (BCDR) [83] is a public dataset of FFDMs. Arevalo et al. [84] further contributed new images to this public dataset, encouraging other researchers to contribute their own images to open repositories. Images are available both at the cranio-caudal view and the medio-oblique view. Researchers may access the images after registering on the database website.

4.5. BancoWeb LAPIMO

This image dataset [85] has over 1400 mammograms in the TIFF format. The images are collected from 320 subjects and contain a wide variety of lesions. The dataset contains images classified as benign, malignant, or healthy.

4.6. The VICTRE trial

The VICTRE Trial dataset [86] is entirely synthetic: A total of 2986 subjects, with breast sizes and densities representative of an entire screening population, were simulated and imaged on *in-silico* versions of digital breast tomosynthesis (DBT) as a replacement for digital mammography (DM). The authors of the dataset have also provided open-source software tools on GitHub [87] to identify ROIs, convert the data to other formats, and insert synthetic lesions. Although the dataset is entirely synthetic, the authors presented their findings from an *in-silico* trial at the FDA Grand Rounds [86].

4.7. OPTIMAM

The OPTIMAM [88] dataset is available upon request from the University of Surrey. The data is stored relationally, and researchers can download the data with an open-source Python package, easily plugging in the data into their systems for plug-and-play processing. The OPTIMAM database contains data from the first OPTIMAM1 project started in 2008, and the OPTIMAM2 project started in 2013: OPTIMAM2 is one of the few public datasets containing annotated 3D DBT imaging. Elangovan et al. [89] report how OPTIMAM2 also includes an image simulation toolbox to generate synthetic 2D-mammograms and 3D DBT images.

5. CNNs for four mammography tasks

With the datasets mentioned in Section 3 and others, researchers

Table 1
Detailed information of the most commonly cited available mammography datasets. '-' means not available.

Dataset	Year	# Imgs	Format	View	Resln. (bit/pxl)	Pros	Cons	Select Publications
mini-MIAS [80]	2003	322	.PGM	MLO	8	The data can be accessed with Unix commands and is easy to retrieve.	Outdated film-screen mammograms and lacks more modern imaging sources like 3D-mammography. Limited to the MLO view.	[33,100,145]
DDSM [81]	1999	10480	.LJPEG	MLO, CC	12, 16	Commonly cited by the literature. Includes both the MLO and CC views.	Lacks an API and researchers need to install a special tool to retrieve images. Consists of outdated film mammography scans.	[33,47,48,120, 123,124,126, 146]
INBreast [82]	2011	410	. DICOM	MLO, CC	14	Both the MLO and CC views are available, and the images are widely cited.	The database is now restricted; Researchers must contact the authors directly for access.	[33,34,36,42, 111,117,120, 123]
BCDR [83,84]	2012	7315	.TIFF	MLO, CC	8; 14	Standard format Precise lesion locations. Includes BI-RADS density annotations and precise mass coordinates, as well as detailed segmentation outlines. Auxiliary patient data (prior surgery, lesion characteristics, biopsy status) is also available.	Limited to only 2D FFDM data. Limited in size.	[25,99,121]
BancoWeb LAPIMO [85]	2010	1400	.TIFF	MLO, CC	12	BI-RADs density labeled. Contains auxiliary information (patient age, scanner brand, hormone replacement therapy status). Standard image format.	Researchers need to wait for administrator approval. Limited in size.	[145]
VICTRE [86]	2018	217,913	. DICOM	MLO, CC	–	Contains precise of mammographic lesions. Entirely synthetic.	Entirely synthetic.	[147] [148]
OPTIMAM [88]	2020	>1 M	. DICOM	MLO, CC	12, 16	Extremely Large Dataset. Open-source API for easy image retrieval in Python.	Researchers need administrator approval for access. Data only comes from patients in England.	[35,132]

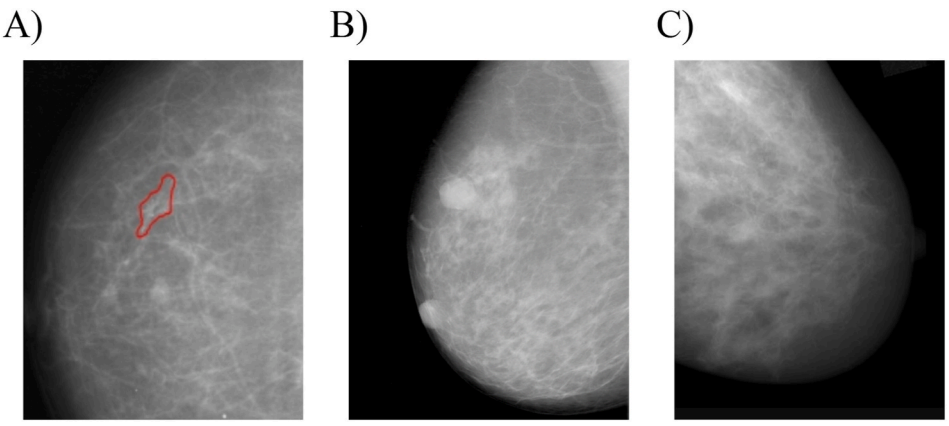

Fig. 6. Select sample images we downloaded from the databases: A) A BCDR FFDM with the ground truth outlined in red. B) The DDSM FFDM (note that the masks are included as separate files). C) A slice from the VICTRE trial FFDM (note that the VICTRE trial is comprised of entirely synthetic data).

work to address challenging mammography tasks, including classifying parenchymal tissue density and identifying and classifying asymmetries, calcifications, and masses. Fig. 8 illustrates a breakdown of the literature's distribution regarding these tasks since 2012.

Addressing these tasks, researchers resolve common clinical problems such as intra- and inter-observer variability [90], low sensitivity [6], and high false-positive rate [5]. This section highlights select CNN-based CAD methods, like supervised, unsupervised, semi-supervised, and transfer-learning that serve as vital decision-support tools to help resolve many of these challenges.

5.1. Breast density classification

The term breast density refers to the relative amount of radiopaque epithelial and stromal tissue elements compared with the amount of radiolucent fatty elements seen in mammograms. Physicians use the Breast Imaging Reporting and Data System (BI-RADS) to classify breast composition (density) into four categories as follows: A) almost entirely fatty, B) scattered areas of fibroglandular density, C) heterogeneously dense (which may be obscure small masses), and D) extremely dense (which lowers the sensitivity of mammography) [91]. Over half of women under age 50 have dense tissue [92]. It is well established that sensitivity decreases with increasing density, mainly due to the overlapping radiopaque's superimposition on dense, cancerous breast tissue. High breast density is a decisive, independent risk factor for breast cancer [93], and CAD models that can classify mammograms with their correct BI-RADs labels can improve early screening and risk assessment.

Current advances in CNNs have replaced traditional feature

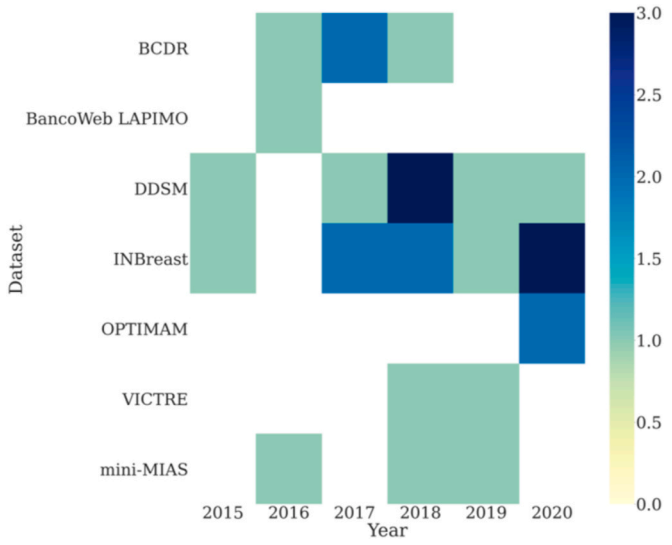

Fig. 7. The number of papers we cite that use different public datasets over time. White represents time instances where datasets are not referenced, either because the datasets were not published during that time, or the published dataset lacked published implementation. Best seen in color.

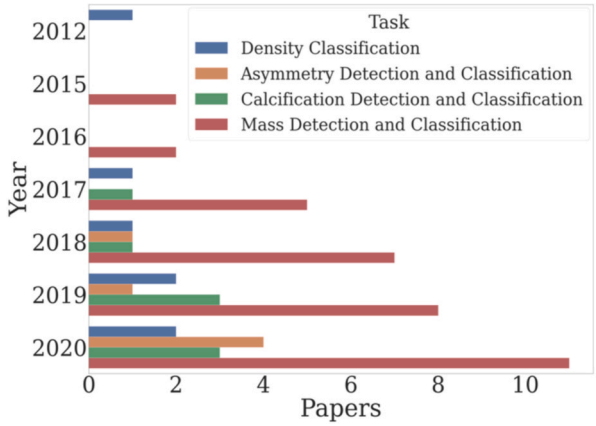

Fig. 8. Bar graph displaying the published paper distribution as a function of publication year and mammography task at the time of writing. Best seen in color.

extraction with network-based methods. Mohamed et al. [32] leverage supervised CNNs for classifying mammography density categories in a private dataset. The authors work with "scattered density" or "heterogeneously dense" labels. The CNN-based model trained on a subset of internally curated images (22,000) highlights supervised CNN's power to classify density. Their analysis aims to predict an image's specific nominal class, as a radiologist would, instead of a blanket "dense/not dense" statement. The model attains high classification accuracy, with an area under the curve (AUC) ranging from 0.94 to 0.98. Li et al. [75] use dilated and attention-guided residual learning to improve the accuracy of traditional CNNs. The dilated CNN increases a network's receptive field while preserving the image resolution. This approach relies on multimodal detection, as it fuses bilateral MLO and CC views to extract features before inputting them into the classifier for breast density classification. This multimodal approach highlights a growing trend in CNNs to use multiple input sources for improved model performance. These researchers apply this approach to a private dataset, and the public INBreast image repository [82]. Deng et al. [76] report the efficacy of SE-attention based networks with BI-RADS density classification. The authors train their model on a curated dataset of both MLO and CC views. Compared to the Inception-V4, ResNeXt, and DenseNet benchmarks, adding SE-attention blocks to these architectures leads to accuracy improvements from 89.97% to 92.17%, 89.64%–91.57%, and 89.20%–91.79%, respectively.

Instead of performing expensive and dedicated training, researchers use transfer learning to extract domain-specific features. Mohamed et al. [32] use a pretrained AlexNet for breast density classification. Wu et al. [94] fine-tune deep CNNs to classify breast density, reporting a pretrained BI-RADs classifier achieves the best performance.

Feature extraction with CNNs can be performed in an unsupervised fashion, ideally when the data is scarce or has minimal domain-specific knowledge. Petersen et al. [95] use denoising autoencoders (DAEs) for extracting features from a private dataset of mammograms and later compared automatic density segmentation scores with manual metrics such as the BI-RADS and Cumulus-like density scores.

5.2. Breast asymmetry detection & classification

Breast asymmetries in mammography represent unilateral deposits of fibroglandular tissue, not conforming to the definition of a mass. Physicians classify these deposits as asymmetry, focal asymmetry, global asymmetry, and developing asymmetry [91]. Researchers identifying asymmetry work to detect asymmetry's presence and quantify its extent. Mathematically, researchers have defined asymmetry by relying on different volume calculations. Fung et al.'s [96] method, representative of recent approaches [97], estimates the breast volume as a cone:

$$V = \frac{1}{3}\pi R_{CC} R_{MLO} H_{MLO} \tag{2}$$

here, R_{CC} is the radius of the CC view, R_{MLO} is the radius of the MLO view, and H_{MLO} is the breast height in the MLO view. Up to 20% of developing asymmetries are due to malignancy [98], representing a key point for researchers to investigate.

Thus, scientists are developing CNN-based methods for breast asymmetry detection and classification. Zhao et al. [77] devise cross-view attention networks that take in four different views (left/-right CC and left/right MLO) of a screening mammogram as input. The authors present bilateral attention mechanisms to fuse feature maps from both the right and left breast views on the DDSM [81], allowing for clear contrast determination independent of algebraic volume calculations. The model fuses and shares weights from all four views' feature maps, and their attention mechanism follows from this feature aggregation, allowing for enhanced context for bilateral symmetry detection.

Breast tissue density is dynamic, so temporal methods harness critical context in interpreting multiple time points. Kooi et al. [30], fusing temporal components, combines two mammographic images at different time points for classifying masses. This is similar to how radiologists make their own decisions. Adding temporal context and analysis through recurrent neural networks helps the researchers understand how patterns in the masses change over time. The authors combine their temporal analysis with multi-scale patches to add additional context to the detected breast symmetry changes.

Cai et al. [99] use transfer learning with a hybrid squeeze-excitation DenseNet model (termed SE-DenseNet). When comparing model performance with pretrained and randomly initialized weights, the authors found that transfer learning boosted accuracy by about eight percentage points with a final best accuracy of 0.982. The authors first pretrain models on the ImageNet and the MSCOCO datasets and fine-tune one version on the DDSM and BCDR datasets. They report how a fine-tuned model outperforms purely pretrained models. Fine-tuned models transfer low-level features from auxiliary domains while still optimizing domain-specific feature maps on the domain-specific mammograms.

Transfer learning plays an essential role in detecting asymmetry. Yu et al. [100] apply a fine-tuned DenseNet201 on the mini-MIAS dataset [80] to detect mammographic abnormalities, such as architectural

distortions and asymmetry. After augmenting the data with morphological opening, binarization, and ROI selection, the authors report a final mean accuracy, sensitivity, and specificity of 92.73%, 94.58%, and 91.67%, respectively. The DenseNet's strength lies in its skip connections: by transferring information from shallower to deeper layers, skip connections overcome the vanishing gradient problem and fuse early with late-stage feature maps to preserve signals, especially in extremely deep networks. In 2020, Shen et al. [101] use a weak, semi-supervised classifier on the NYU Breast Cancer Screening Dataset to aggregate local and global features to make predictions for mammographic breast cancer screening. The authors rely on weakly-supervised object detection via a low-capacity network to locate regions of interest on the full mammogram before using a higher capacity supervised classifier to identify abnormalities using only full-image level labels. Through its low-capacity network, the model extracted regions of breast asymmetry. The former network localizes the asymmetry in an unsupervised fashion; the latter successfully labels it as an asymmetry.

5.3. Calcification detection & classification

Calcifications are deposits of calcium oxalate, calcium phosphate, or magnesium salts within the breast tissue that appear as white specks on a mammogram. Calcifications form within the ductal system, the breast acini, stroma, and vessels [102]. However, the mechanisms by which calcifications occur are not clearly understood [103]. The diagnostic approach to breast calcifications detected on mammography analyzes the morphology, distribution, and, sometimes, change over time. Based on such analysis, radiologists label calcifications as either benign or suspicious, the latter class requiring biopsy [91]. Researchers in the literature work on segmenting and detecting MCs because they are present in approximately 55% of non-palpable breast malignancies. They account for the detection of 85–95% of cases of ductal carcinoma in situ (DCIS) by screening mammography [104] and may also present in invasive cancers [105].

A common theme of combining model sub-networks and features in sophisticated architectures runs throughout the literature. Wang and Yang et al. [39] develop a contextual architecture consisting of two CNN based subnetworks. The first subnetwork is specific to a macroscopic input image, and the second focuses on smaller, microscopic dimensions within the former subnetwork's macroscopic domain. The authors fuse the last hidden layers in the architecture to combine both networks, yielding a fully connected classification output layer dependent on both streams. When applied to a private mammography dataset, the model achieves a sensitivity of 0.874 compared to 0.84 from an SVM benchmark classifier; However, the authors omit specificity scores. Still, their findings show how fusing global and local features adds context to improve classification performance. Extending this trend in ensembling subnetworks, Savelli et al. [106] propose a model with four CNNs, each with training image scale dimensions of 12×12, 24×24, 48×48, and 96×96, respectively. Each input dimension enlargement provides a broader scope for context; soft-voting fuses the four input sources to yield an aggregate decision. When applied to the INBreast dataset, the model achieves a sensitivity of 0.83 on the task compared to 0.80 with SVM-based methods.

Many CNN-based algorithms for calcification detection in 2D FFDMs and recent developments in 3D DBT indicate a gap between 2D and 3D mammography data. To resolve this, Yin et al. [107] devise a scheme to transfer learned features from FFDMs calcification detection to DBTs. The authors use a hierarchical model for feature extraction with different input layer scales to detect calcification clusters. Before training, the authors compose a private dataset of 2D and 3D images, and later achieve comparable results when applying their feature extraction method to each 3D DBT frame. Samala et al. [54] feed ROI patches into a three-convolutional layer CNN to detect MCs in 3D DBT. The authors use 16×16 dimensional ROIs from the DBT slices to reduce memory constraints, yet because the ROI sample set is small, the authors rely on data augmentation to boost their results. Cai et al. [108] use transfer learning to enhance deep feature extraction for calcification detectors. The authors use an ImageNet-pretrained AlexNet with all layer weights preserved, except for the last two layers. After the authors fine-tune on two private datasets, the authors use the penultimate layer's features for visualization with the t-distributed stochastic neighbor embedding (t-SNE) [109] algorithm: they observe clear clustering patterns between benign and malignant samples. Interpreting latent coding distributions in 2-D space can help physicians and computer scientists identify latent embedding clustering patterns, increasing result interpretability.

Authors also implement unsupervised methods to address MC detection. For instance, Wang et al. [57] report how stacked autoencoders (SAE)s extract features as input sources for a deep classifier for detecting MCs. The authors first segment ROIs in an unsupervised manner before applying a hierarchy of SAEs to the segmented regions to yield latent feature maps. They then apply a SoftMax layer to the SAE to confirm if an anomalous region in the mammogram is a calcification. The model has a discriminative accuracy of 0.873 when applied to a private dataset from a retrospective study.

5.4. Breast mass detection & classification

Radiologists define a mass as a space-occupying 3D lesion observed in two different projections. The descriptors of a mass in mammography comprise shape, margins, and density [91]. For example, invasive ductal carcinoma (IDC) (not otherwise specified) is the most common type of invasive breast cancer: it typically presents as a spiculated, irregular mass or as a new focal asymmetry [110]. Researchers in the literature address classifying, segmenting, and locating masses in the image.

Ertosun et al. [111] report a CNN-based visual search algorithm for localizing masses within mammograms. The model consists of two sub-modules: an anomaly detector, followed by a mass localizer. The anomaly detector classifies a mammogram containing masses or no masses before feeding the resulting mass-containing images into the localizer. These modules, trained on over 2500 images from the DDSM dataset, are constructed as hierarchical CNN layers. Al-Masni et al. [47] use the You Only Look Once (YOLO) deep network Redmon et al. [112] for simultaneous mass detection and classification. The YOLO method is a type of end-to-end learning where successive convolutional layers first split the image into sub-regions; the algorithm then places bounding boxes around important objects and class labels. After 5-fold cross-validation, the authors reported an overall sensitivity score of 100% and a specificity of 94% on the INBreast dataset. Moreover, Moor et al. [113] apply the U-Net to segment and classify mammographic masses. The U-Net involves doubling the number of filters during each downsampling block and applying batch normalization layers for regularization and increasing the learning rate. The trained model achieves a sensitivity of 0.94 for mass segmentation and a sensitivity of 0.98 for overall image classification as either normal or containing malignant lesions. Sun et al. [78] implement an attention-guided dense-upsampling network for mass segmentation from whole mammograms. Models trained on whole mammograms are efficient because they avoid the preprocessing step of ROI generation. Sun et al.'s model consists of successive encoding, upsampling, and decoding blocks. The model achieves a sensitivity of 85% on the DDSM dataset and 79% on the INBreast dataset. In another supervised-learning approach, Li et al. [114] report a DenseNet-II based model that uses whole mammographic images as input. When trained on a dataset of 2042 cases provided by the Shanxi Medical University, the model outperforms the pretrained AlexNet, VGGNet, GoogLeNet, and DenseNet-I with an overall accuracy of 94.55%.

Introducing transfer learning to complex model architecture has yielded significant insights. Yemini et al. [43] use transfer learning with the Google Inception-V3 [115] architecture to generate an improved receiver operator curve (ROC) result when detecting masses in the

INBreast dataset. Lvy et al. [31] examine a shallow baseline model, AlexNet [26], and GoogLENet [116] to compare pretraining effects in combination with mass context and data augmentation for overall mammogram classification. The researchers pretrain AlexNet and GoogLENet on the ImageNet [63] dataset before fine-tuning on the DDSM mammographic dataset [81]. The study findings demonstrate how a fine-tuned AlexNet with an accuracy of 0.90 outperforms a shallow baseline network with an accuracy of 0.66. Context regions around the masses, including 50 pixels padded around the ROI, also boost AlexNet's accuracy. Chougrad et al. [33] show how fine-tuning the VGG16 [65], ResNet50 [58], and the Inception-V3 [115] leads to greater AUC improvements. The researchers perform an ablation study to iteratively remove the tuning layers from their networks to observe how each block affects model performance. After optimizing the model architecture, the authors achieve the greatest improvements on the DDSM and MIAS datasets, with AUCs of 0.99 and 0.98. Abdel Rahman et al. [117] demonstrate that pretrained CNNs such as modified Inception-V3 and ResNet50 classify tumors as either benign or cancerous in mammograms from the DDSM. Furthermore, Agarwal et al. [35] report a pretrained Faster R–CNN's [118] success in breast cancer mass detection and classification. The authors are the first to benchmark CNN model performance on the semi-private OPTIMAM dataset [119]. They also perform experiments where they pretrain on the INBreast dataset and then fine-tune on OPTIMAM FFDMs, highlighting how their methods generalize across dataset domains. The authors report an F1 score of 0.86 for malignant masses and 0.74 for benign samples. Similarly, Ribli et al. [120] implement a Faster R–CNN, achieving an AUC of 0.95 on the INBreast dataset. They are one of the few authors to open source their implementation on GitHub, increasing reproducibility. The authors strengthen their model's robustness by training on the DDSM dataset and testing on the INBreast dataset, showing domain generalizability.

Unsupervised learning involving autoencoders is also popular, especially with sparse datasets. Dhungel et al. [111] devise an unsupervised deep belief network that achieves a true positive rate (TPR) of 0.96 at a false positive per image (FPPI) rate of 1.2. In later work, the same authors introduce bounding boxes to increase the model's performance [48]. With a private dataset and the BCDR, Becker et al. [121] implement a multipurpose image analysis software (ViDi Suite Version 2.0; ViDi Systems Inc, Villaz-Saint-Pierre, Switzerland), which researchers use for anomaly detection in solar panel and textile manufacturing. Using this same model on mammograms, the researchers report how the anomaly detector achieves higher sensitivity scores than human readers. Moreover, the model excels at detecting lesions in low-density breast tissue as compared to higher density tissue. Al Antari et al. [122] devise a three-stage scheme to detect, segment, and later classify masses. The detection step is a YOLO algorithm similar to the one Al-Masni et al. [47] use. Later, a full-resolution Convolutional Network (FrCN) autoencoder trained on labeled ROIs returns the tumor's segmented mask images. Because the encoder network lacks subsampling and max pooling, the model preserves the original image resolution, only encoding it into high-level feature maps. Finally, the authors integrate a five-layer CNN for classifying the masses as benign or cancerous.

Zhu et al. [36] apply deep multi-instance learning approaches with CNNs, on 2D FFDMs, and later Yousefi et al. [37] apply them to 3D DBTs, overcoming memory constraints that arise when working with 3D tensors, as well as resolving sparse data annotation problems. Platania et al. [34] use pretrained CNN weights to initialize a binary classifier's weights in a semi-supervised fashion. In their two-module system, the authors use a YOLO [112] inspired CNN to detect ROIs. The researchers feed the initial detector's weights into a FFDM classifier before training the classifier on the entire mammogram image. After testing on the DDSM, the authors report an AUC score of about 92.3% and an accuracy of 93.5%. Shen et al. [123], apply encoders trained on the INBreast dataset [82], along with adversarial learning methods to detect masses in mammograms. The model adapts knowledge from well-annotated datasets to other unlabeled datasets. The adversarial learning helps align the latent target features from unlabeled datasets with labeled source domain latent features.

Jung et al. [124] use the RetinaNet [125] model as a one-stage mass detector in mammography. The RetinaNet uses a focal loss function rather than traditional cross-entropy, placing more weight on misclassified objects. It excels when there is a heavy class imbalance between positive and negative samples, as is the case for pixel-based mass segmentation and overall image classification. Jung applies a semi-supervised RetinaNet based model on the INBreast dataset images, later reporting true positive rates of up to 0.99. Li et al. [49] expand on the original U-Net in the DDSM by combining a dense U-Net with attention gates. The U-Net's encoding component is a dense CNN, and the decoder is a series of attention gates. Working with the DDSM and INBreast datasets, Yan et al. [126] apply auto-context [127] to the traditional U-Net: the auto-context makes long-range spatial context from lower resolution persist through training at a higher resolution. The researchers use two U-Nets at different image scales to determine how to fuse submodel outputs optimally. Using multiple resolutions, Lotter et al. [46] pass different-size patches along the mammogram to aggregate local and global features on the DDSM. The authors design a two-stage approach: 1) A ResNet determines if a mass or calcification is present in patches of different scales. 2) Then, global features are aggregated to determine if a mammogram is cancerous. The authors achieve an AUC of 0.92, and their results highlight how multi-scale patches, as well as global features, aggregate context and allow for accurate detection.

Researchers also implement unsupervised and semi-supervised techniques for mass detection and classification. Working with a private dataset, Sun et al. [128] design a weighted graph-based scheme to discriminate normal from abnormal mammograms. This approach relies on a semi-supervised neural structured learning scheme [129], where the graph outputs features from both labeled and unlabeled images to train a deep CNN classifier. Given the heterogeneous ROI annotation presence in the dataset, the model still discriminates benign from malignant masses. With GANs' popularity, Guan et al. [42] synthesize ROIs from DDSM mammograms with these models to boost traditional data augmentation, improving a CNN classifier accuracy by 3.6%. Training on synthetic images alone yields an overall stable performance score of about 64.52% compared to around 74% with organic and synthetic images. These preliminary results indicate GANs' promise in CAD for breast cancer.

Multimodal methods, involving fusing various modal input sources for a final decision, also assist in mass detection. Sharma and Preet [130] combine inputs from both the MLO and CC breast views for later fusion in their CNN-based classifier. The authors first display results from training on MLO and CC views separately and then demonstrate how fused information from both views helps the model perform accurately. Using 3D DBTs, Zhang et al. [131] use early and late tensor fusion with slices of the 3D DBT. The researchers use the entire DBT volume with slices, yet the model lacks 3D convolutional layers, admittedly because of their high computational costs; nevertheless, the authors fuse features from the 2D slices, still preserving information in the z axis while also reducing computational costs. Liang et al. [51] use a joint CNN that relies on both FFDM and 3D DBT, integrating both 2D and 3D CNNs for a classifier to classify benign and malignant mammograms. The authors use this 2D-3D ensemble method, composing a pretrained AlexNet as a backbone for feature extraction before input into their 2D and 3D convolutional layers. When trained on independent modalities, the model achieves 0.87 and 0.72 AUC on DM and DBT, respectively; when both mammographic images are ensembled, the new multimodal model, fusing data-specific features through weight-sharing, achieves an AUC of 0.97. McKinney et al. [132] further demonstrate how a four-model ensemble fusion method excels at detecting both MCs and masses. The researchers devise different individual model architectures, using various image resolutions as input. They also and add and remove layers

between the path classifier and the top layers. The researchers again compare the individual model performance to an ensembled version comprised of the four best performing individual models, showing how fusion methods outperform independent classifiers.

5.5. Challenges for the four mammography tasks

5.5.1. Density classification

Classifying breast density remains a challenging yet necessary task, as it is one of the first heuristics used by oncologists to determine risk status. Melnikow [133] shows that up to 22% of women have varying BI-RADS classifications based on the mammography machine used. Researchers, and more importantly, oncologists need to be confident in how mammograms are labeled. It is challenging to standardize CAD performance over varying machines; clinics now have higher-resolution imaging techniques, yet researchers should address device-level mammogram variations by generalizing their models across multiple datasets. Many of the papers also lack explainability: The Peterson et al. [95] autoencoder encodings are abstract representations of data, and very difficult for experienced clinicians to interpret; determining communication methods for CNNs in the form of clustering maps, ablation studies, and feature map analysis remains a challenge requiring further investigation.

5.5.2. Breast asymmetry detection and classification

Although our focus is on mammography, when investigating the literature, we find that most reported CNN-based asymmetry detection is on thermography, a clinical imaging tool absent from most clinics. We encourage more works specific to asymmetry detection in mammograms. Technically, we note that researchers need to report what they define as an asymmetry for ground truth measurements. Many of the papers using CNNs for asymmetry detection discuss asymmetry in broad terms, and these publications lack explicit mathematical formalism of what mathematically constitutes asymmetry as ground truth. We highlight the Zhao et al. [77] paper for fusing different imaging views to add more context to identifying asymmetry, and we challenge more authors to explore more multi-view methods. We note that most public datasets contain both the MLO and CC views from our discussion on datasets. However, we also acknowledge how most public datasets lack asymmetry (or even breast volume) annotations. We call for papers that either a) provide these necessary ground-truth annotations to the datasets or b) present new datasets containing asymmetry annotations. Moreover, the literature lacks time-series analysis. Kooi et al. 's [30] analysis is one of the few that examines two time points. Researchers need a richer analysis of time-series mammography, as oncologists themselves contextualize their analysis with prior imaging.

5.5.3. Calcification detection and classification

CAD systems first began using calcifications for breast cancer detection. As demonstrated by Cai et al. [108], researchers are improving their ability to explain convolutions' latent representations. However, the community needs more interpretability in AI algorithms to add patient and provider trust in their models. Another challenge specific to MCs is a lack of publicly available 3D DBT datasets with explicit annotations. For example, although the DDSM has FFDM images annotated with calcifications, it lacks annotated 3D DBT data annotated for calcifications. The VICTRE trial does contain DBT images with labeled calcification data.

5.5.4. Mass detection and classification

Ensemble methods are increasingly popular, and within the past year, using 3D CNNs for analyzing DBTs [51]. Although 3D CNNs are trendy, they require massive computational power, and many GPUs lack enough RAM to handle so many floating-point operations (FLOPs). Many researchers focused on DBT overcome this RAM limitation by slicing the 3D tensors and applying 2D CNNs to each slice [131]. However, 3D CNN methods show promise because they allow for 3D spatial analysis, allowing for full contextual feature maps that can capture an additional dimension of data and draw insights into how the breast tissue is in real-life.

When reviewing the literature, we note that few researchers [36, 120] open source their code to GitHub. Leaders both in the AI and medical community call for more AI papers to publish their source code on public repositories [134,135], and this a challenge found in the breast mass detection and every other mammography task.

6. Quantitative comparison

The CNN for mammographic CAD scope is vast; contextualizing and comparing currently published methods is necessary to understand current gold-standards in the field and specific tasks open for improvement. This section compares selected authors' reported results on a per-task basis.

6.1. Density classification results

The literature on CNNs applied to breast density classification is limited to datasets with the BI-RADS label. The majority of published literature is limited to privately collected mammogram repositories. The first group of rows in Table 2 detail and compare the different techniques' reported results on various data sources. The results indicate a shift in the past three years towards attention-based mechanisms combined with CNNs, leading to improved model performance. Trends in increasing ROC-AUC values highlight how transfer learning improves attention-based models.

6.2. Breast asymmetry detection and classification results

Breast asymmetry detection is limited to either privately curated datasets, the DDSM, or the miniMIAS dataset. Table 2 shows different CNN-based results for this task. Examining the quantitative results from various methods, we note that recent papers have primarily shifted to reporting the ROC-AUC curves. The ROC-AUC is a more informative metric, as it allows readers to compare the true-positive and false-positive rates. Researchers have recently implemented complex CNN block combinations like squeeze-excitation and attention networks to enhance model performance.

6.3. Calcification detection and classification results

Calcification detection with CNNs has a more extended history than other tasks, with the first CNN based article published in 1995. Since then, researchers have devised sophisticated strategies involving multi-context methods and autoencoders.

6.4. Mass detection and classification results

Table 2 's results show how researchers have made improvements to increase mass detection sensitivity, with the highest reported sensitivity being 0.99. There is, however, room to make more extensive improvements in the specificity of these models. The GAN-based data augmentation model's lower performance is understandable: the researchers test the DDSM-based synthetic images model, suggesting that future work should improve the GANs' performance in data augmentation.

With breast cancer detection, the primary interest is in the sensitivity (also known as the true-positive rate). Mass detection typically has a sensitivity rate of 0.90, and mass classification as either benign or malignant has comparable performance. Recent applications of autoencoders for mass segmentation merit more investigation to improve reported sensitivity rates. The table highlights a need for investigations into cross-domain transfer learning.

Table 2
Quantitative comparison of results reported by authors for density classification, and asymmetry, calcification, and mass detection and classification. '-' indicates the data point is not available.

Mammography Task	Method	Spec.	Sens.	Acc.	AUC	F1
Density Classification	Unsupervised Deep Autoencoders [95]	–	–	–	0.68	–
	Supervised Pretrained CNN on Predicting BIRADS Classes [32]	–	–	–	0.94–0.98	–
	Transfer Learning with Deep CNN [94]	–	–	0.86	0.9	–
	Dilated Attention Guided CNN [75]	–	–	–	0.97	0.87
Asymmetry Detection and Classification	SE-Attention based Classifier [76]	–	0.9	0.92	–	0.9
	SE-DenseNet [99]	–	–	0.98	0.98	–
	DenseNet201 Transfer Learning [100]	0.92	0.95	0.93	–	–
	ResNet50 with Cross-view Attention Network [77]	–	–	–	0.86	–
	Weakly Supervised Localization [101]	–	–	–	0.93	–
Calcification Detection and Classification	SE-DenseNet [99]	–	–	0.98	0.98	–
	DenseNet201 Transfer Learning [100]	0.92	0.95	0.93	–	–
	Multi-scale patches [46]	–	–	–	0.92	–
	ResNet50 with Cross-view Attention Network [77]	–	–	–	0.86	–
	Weakly Supervised Localization [101]	–	–	–	0.93	–
Mass Detection and Classification	Whole-image Mass Localization [111]	–	0.85	0.85	–	–
	GoogLeNet pretrained classifier [31] YOLO based ROI Classifier [34]	–	0.93	0.93	–	–
	YOLO based ROI Classifier [34]	0.93	0.94	0.95	0.92	–
	Pretrained InceptionV3 Classifier [33]	–	–	0.97	0.97	–
	Attention-guided Dense Upsampling segmentation [78]	–	0.79	–	–	–
	GAN based Data Augmentation [42]	–	–	0.75	–	–
	ResNet50-like Pretrained Tumor Classifier [117]	–	0.87	0.86	–	0.87
	Deep Belief-RF Detection [146]	–	0.96	–	–	–
	Bounding Box CNN Detection [48]	–	0.95	0.95	0.91	–
	YOLO based Classifier [47]	0.94	1	0.95	0.96	–
	RetinaNet based Detector [124]	–	0.95	–	–	–
	Pretrained InceptionV3 Classifier [33]	–	–	0.96	0.98	–
	Multiscale Autoencoder Segmentation [126]	0.99	0.72	–	–	–
	Attention-guided Upsampling Segmentation [78]	–	0.85	–	–	–
	Adversarial Domain Transfer Learning [123]	–	0.88	–	0.9	–
	DM Dream Ensemble [149]	–	0.88	–	0.86, 0.90	–
	ROI patches [54]	–	–	–	0.92	–
	Combining Handcrafted Features with CNNs [84]	–	–	–	0.82	–
	ROI patches [54]	–	–	–	0.92	–
	Multi-instance Learning [36]	–	–	0.92	0.86	–
	International Mass and Calcification classification [132]	–	–	–	0.90, 0.74	–
	Multi-scale patches [46]	–	–	–	0.92	–
	Faster-RCNN Detection [35]	–	–	–	–	0.74–0.84
	Faster-RCNN Detection with Cross-Domain Training [120]	–	–	–	0.95	–
	Multi-instance Learning [36]	–	–	0.92	0.86	–
	Transpara™TM-Assisted Reading [141]	0.79	0.86	–	0.89	–
	Cascade Trees and CNNs [25]	–	0.78	–	0.79	–

7. Approach pros and cons

We strengthen our analysis by comparing the pros and cons of various reported techniques. Table 3 evaluates different approaches used in the literature, citing their pros and cons; it also provides references for papers using these approaches, in case future authors want to benchmark their methods against published works.

8. FDA approved devices

Table 4 highlights all the main features of FDA-approved devices with underlying CNN methods for interpreting mammographs.

8.1. PowerLook® tomo detection V2

The FDA approved the PowerLook® software [136] for its proven ability to detect soft tissue densities and calcifications in 3D DBTs. The model segments the identified lesions for radiologists to review. This algorithm is one of the first commercial products to use CNNs on 3D rather than 2D mammography. It provides algorithmic confidence scores for radiologists to know which lesions merit rigorous review.

8.2. DM-density®

The DM-Density® tool [137] calculates percent breast density as the ratio of fibroglandular tissue to total breast area estimates; the model outputs predicted BI-RADS classifications of the tissue. This software is specific to classifying breast density and provides quantitative values related to density for radiologists to interpret before they make difficult diagnostic decisions.

8.3. ProFound™ AI software V2.1

The FDA approved the ProFound™ AI device in 2019. The program processes features and uses pattern recognition to identify suspicious breast lesions with CNNs, focusing on breast density classification and calcification detection tasks.

8.4. cmTriage™

The FDA approved cmTriage™ [138,139] to analyze mammograms of at-risk patients for triage. The model uses CNN-based algorithms to alert radiologists on 2D FFDMs passively. It specifically classifies tissue density and identifies calcifications and masses.

8.5. Transpara™ 1.6.0

The FDA approved Transpara™ 1.6.0 [140] in 2020 as a mammographic CAD system for locating calcification groups and soft-tissue regions. The model outputs scores indicating the likelihood that cancer is present. The software has CNNs to detect calcifications and soft tissue lesions: it specifically locates densities, masses, architectural distortions,

Table 3
Comparing pros and cons of select approaches.

Approach	Pros	Cons	Select References
Multi-View Fusion	Different mammographic imaging angles provide more context.	Challenges arise with explainability, as its difficult to pinpoint what view contributes to the classifier.	[38,77,130]
Multi-Scale CNNs	Aggregates context. Larger receptive field.	Choosing image scales is often arbitrary.	[46,106, 107,124, 131]
Transfer Learning	Reduces training time. Increases sample space to transfer high-level features across domains.	Models lacking fine tuning tend to underperform.	[31,32,43, 94,100,108, 123]
Network Ensembling	Ensembling forces each individual model to compensate for the others' weaknesses.	Complex model combinations makes it challenging to explain where a model derives its decision from.	[51,132, 146]
Generative Models	Greatly resolves data augmentation problems. Assists with class balancing datasets.	Computationally expensive. May misrepresent true FFDM and DBT data. Needs more validation and experimentation.	[42,52,69, 70]
Autoencoders	Great for dimensionality reduction. Can detect salient features necessary for image reconstruction.	Encodings lead to abstraction that requires explainable AI.	[57,95,122]
ResNets	Skip connections reduce the number of trainable parameters.	Non-linearity leads to lack of explainability.	[46,51]
UNet	Perform well with small ($n < 100$) datasets. Skip connections preserve information from the encoder.	It is challenging to understand how the UNet image transformations and traditional signal processing approaches (wavelets, nonlocal processing, etc.).	[49,78,113, 126]
Recurrent CNNs	Allows for time-series context analysis.	Vanishing gradients. Accuracy diminishes with long sequences.	[30]
Deep Multi-Instance	Useful for sparsely labeled data.	Fails with imprecise instance classes.	[36,37]
Attention	Selectively focuses on segments of sequential data. Enhances model interpretability.	Significantly adds to model spatial complexity.	[75–78]
3D CNNs	Allows for spatial context and feature extraction in the z dimension.	Computationally expensive.	[51]

and asymmetries. This device is specific to FFDMs. In 2019, Rodriguez et al. [141] measured radiologist reading time as a function of the CNN-based Transpara™ CAD tool presence; the radiologist classification AUC was higher with the tool, yet reading time remained stagnant. Interestingly, the Transpara™ tool alone performs at the sole radiologist's level, with only the technology and physician reading yielding significant performance gains.

8.6. *MammoScreen*™

The MammoScreen™ software [142] uses CNN models trained on an extensive database of biopsy-proven breast cancer mammograms and healthy tissue. The device also works on FFDMs in DICOM format, outputting predicted lesion locations and model certainty scores.

Table 4
List of FDA approved CNN-based CAD algorithms for mammography between 2018 and 2020.

Trade Name	Year Approved	CAD Task	FFDM	DBT
PowerLook®Tomo Detection	2018	Density Classification; Calcification Detection	✓	
DM-Density®	2018	Density Classification		
ProFound™	2019	Density Classification; Mass Detection	✓	
cmTriage™	2019	Density Classification; Calcification Detection; Mass Detection		✓
Transpara™ 1.6.0	2020	Density Classification; Asymmetry Detection; Calcification Detection; Mass Detection		
MammoScreen™	2020	Mass Detection		✓
densitas densityai™	2020	Density Detection		✓

8.7. *Densitas densityai*™

The densitas densityai™ software [143] predicts breast density in FFDMs and 3D DBTs. It provides an ACR BI-RADS Atlas 5th Edition breast density category to assist radiologists in their diagnostic decisions.

9. Open research areas & conclusion

Throughout the literature, we note critical open areas for research. Transfer learning has helped to apply knowledge from larger, well-annotated datasets to sparser ones. However, for the breast asymmetry detection task, this needs more in-depth exploration. This approach requires additional annotated datasets, and thus future works should focus on generating extensive corpora. Our survey tabulates existing public datasets of breast images and highlights a need for public data. Moreover, practitioners should apply and test their methods across multiple datasets to validate a model's robustness across domains.

Throughout this paper, there is a skew in the current corpus towards mass detection tasks. Researchers should further apply CNN-based techniques to density classification and asymmetry detection and classification, which have relatively lower numbers of related publications. Literature related to breast mass localization and classification is limited to discriminating between benign and malignant masses; researchers should explore predicting fine-grained mass types. For instance, models distinguishing pre-invasive DCIS from IDC and ILC can discriminate between pre-cancerous and more critical lesions. Moreover, models tend to excel in fatty breast tissue but struggle with dense tissue; research that specifically addresses this gap may significantly improve CAD systems.

GANs have revolutionized the major class imbalance problem that exists in mammography datasets. They generate new synthetic images of masses, typically less than 150 pixels in area. However, researchers should generate large-scale synthetic mammograms on the 1024×1024-pixel level to target both macro- and micro-features. Across many papers, authors cite incomplete or very unbalanced datasets. GANs' ability to generate large synthetic images may help solve this problem.

We encourage researchers to open-source their source code or develop APIs for working with their models. Researchers [134,135] recommend open-sourcing repositories for other researchers to reproduce prior findings. GitHub helps disseminate software for reproducibility. Some researchers go one step further to develop application programming interfaces (API)s with easy access: The OPTIMAM [88] dataset includes a Python API, and the VICTRE [87] dataset also provides a Python API to retrieve and preprocess the data.

Researchers should focus on their findings' clinical implications and how they can communicate their findings to a medical and computational audience. Translating research insights into clinical curricula and

training is necessary for the next generation of physicians [144]. For instance, radiologists rely on medical imaging's temporal changes to make definitive diagnostic and prognostic decisions, and computational models and curricula on using these new tools should integrate with clinical expertise for optimal performance. CAD systems can assist radiologists to accurately detect and diagnose breast cancer [141]. We need more studies demonstrating how CNN-based CAD systems perform in sync with a trained clinician. Radiologists can help interpret a system's decision to deliver better care.

Overall, the survey highlights current trends in CNNs for mammography by exploring concepts, datasets, and the literature. CNN-based CAD can work in harmony with physicians, and there are many opportunities open for clinical and computational innovation.

Author contributions

Abdelrahman L contributed to data acquisition and drafting of the article. Al Ghamdi M contributed revising the manuscript critically for important intellectual content. Collado-Messa F contributed to data acquisition and analysis and interpretation, and Abdel-Mottaleb M completed final approval of the version submitted.

Declaration of competing interest

The authors have no disclosures or competing interests.

References

[1] National Cancer Institute, Cancer of the breast (female) - cancer stat facts, SEER (2020), 1–1.

[2] E. Heer, A. Harper, N. Escandor, H. Sung, V. McCormack, M.M. Fidler-Benaoudia, Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study, The Lancet Global Health 8 (no. 8) (2020) e1027–e1037, https://doi.org/10.1016/S2214-109X(20)30215-1.

[3] C. Nickson, K.E. Mason, D.R. English, A.M. Kavanagh, Mammographic screening and breast cancer mortality: a case-control study and meta-analysis, Cancer Epidemiol. Biomark. Prev. 21 (9) (Sep. 2012) 1479–1488, https://doi.org/10.1158/1055-9965.epi-12-0468.

[4] M. Broeders, et al., The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies, J. Med. Screen 19 (1) (Sep. 2012) 14–25, https://doi.org/10.1258/jms.2012.012078.

[5] W.A. Berg, R.E.H. PhD, D.B. Kopans, P. Robert, A. Smith, Frequently asked questions about mammography and the uspstf recommendations: a guide for practitioners, Soc. Breast Imag. (2020) 1–17.

[6] C.K. Kuhl, The changing world of breast cancer, Invest. Radiol. 50 (9) (Sep. 2015) 615–628, https://doi.org/10.1097/rli.0000000000000166.

[7] S. Destounis, A. Santacroce, Age to begin and intervals for breast cancer screening: balancing benefits and harms, Am. J. Roentgenol. 210 (2) (2018) 279–284, https://doi.org/10.2214/AJR.17.18730.

[8] E. Sickles, W. Weber, H. Galvin, S. Ominsky, R. Sollitto, Baseline screening mammography: one vs two views per breast, Am. J. Roentgenol. 147 (6) (1986) 1149–1153, https://doi.org/10.2214/ajr.147.6.1149.

[9] A.A. Roelofs, et al., Importance of comparison of current and prior mammograms in breast cancer screening, Radiology 242 (1) (2007) 70–77, https://doi.org/10.1148/radiol.2421050684.

[10] A. Rimmer, Radiologist shortage leaves patient care at risk, warns royal college, BMJ Br. Med. J. (Clin. Res. Ed.) 359 (2017), https://doi.org/10.1136/bmj.j4683.

[11] J.J. Fenton, et al., Influence of computer-aided detection on performance of screening mammography, N. Engl. J. Med. 356 (14) (Apr. 2007) 1399–1409, https://doi.org/10.1056/nejmoa066099.

[12] C.D. Lehman, R.D. Wellman, D.S.M. Buist, K. Kerlikowske, A.N.A. Tosteson, D. L. Miglioretti, Diagnostic accuracy of digital screening mammography with and without computer-aided detection, JAMA Internal Med. 175 (11) (Nov. 2015) 1828, https://doi.org/10.1001/jamainternmed.2015.5231.

[13] A.D. Trister, D.S.M. Buist, C.I. Lee, Will machine learning tip the balance in breast cancer screening? JAMA Oncol. 3 (11) (Nov. 2017) 1463, https://doi.org/10.1001/jamaoncol.2017.0473.

[14] American College of Radiology, Data science Institute,, FDA Clear. AI algorith. (2020), 1–1.

[15] T.G. Debelee, F. Schwenker, A. Ibenthal, D. Yohannes, Survey of deep learning in breast cancer image analysis, Evolv. Syst. 11 (1) (Mar. 2020) 143–163, https://doi.org/10.1007/s12530-019-09297-2.

[16] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, R. Zwiggelaar, Deep learning in mammography and breast histology, an overview and future trends, Med. Image Anal. 47 (2018) 45–67, https://doi.org/10.1016/j.media.2018.03.006.

[17] D. Houfani, S. Slatnia, O. Kazar, N. Zerhouni, A. Merizig, H. Saouli, Machine learning techniques for breast cancer diagnosis: literature review, in: Advanced intelligent systems for sustainable development (ai2sd'2019), 2020, pp. 247–254, https://doi.org/10.1007/978-3-030-36664-3_28.

[18] G. Litjens, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, https://doi.org/10.1016/j.media.2017.07.005.

[19] V.M. Rao, D.C. Levin, L. Parker, B. Cavanaugh, A.J. Frangos, J.H. Sunshine, How widely is computer-aided detection used in screening and diagnostic mammography? J. Am. Coll. Radiol. 7 (10) (2010) 802–805, https://doi.org/10.1016/j.jacr.2010.05.019.

[20] A. White, et al., Cancer screening test use - United States, 2015, MMWR Morb. Mortal. Wkly. Rep 66 (8) (Mar. 2017) 201–206, https://doi.org/10.15585/mmwr.mm6608a1.

[21] A. Kolak, et al., Primary and secondary prevention of breast cancer, Ann. Agric. Environ. Med. *24* (4) (Dec. 2017) 549–553, https://doi.org/10.26444/aaem/75943.

[22] T. Kooi, et al., Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312, https://doi.org/10.1016/j.media.2016.07.007.

[23] S. Singh, V. Kumar, H. K. Verma, and D. Singh, "SVM based system for classification of microcalcifications in digital mammograms," in 2006 international conference of the ieee engineering in medicine and biology society, 2006, pp. 4747–4750, doi: 10.1109/IEMBS.2004.1403481.

[24] W. Chiracharit, Y. Sun, P. Kumhom, K. Chamnongthai, C. Babbs, E.J. Delp, Normal mammogram classification based on a support vector machine utilizing crossed distribution features, in: In the 26th Annual International Conference of the Ieee Engineering in Medicine and Biology Society vol. 1, 2004, pp. 1581–1584, https://doi.org/10.1109/IEMBS.2004.1403481.

[25] A. Bria, C. Marrocco, F. Tortorella, Addressing class imbalance in deep learning for small lesion detection on medical images, Comput. Biol. Med. (2020) 103735, https://doi.org/10.1016/j.compbiomed.2020.103735.

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, 2012, pp. 1097–1105, https://doi.org/10.1145/3065386.

[27] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. (1951) 400–407.

[28] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (7) (2011).

[29] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, 2014 arXiv preprint arXiv:1412.6980.

[30] T. Kooi, N. Karssemeijer, Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks, J. Med. Imag. 4 (4) (2017), https://doi.org/10.1117/1.JMI.4.4.044501, 044501.

[31] D. Lévy, A. Jain, Breast mass classification from mammograms using deep convolutional neural networks, 2016 arXiv preprint arXiv:1612.00542.

[32] A.A. Mohamed, W.A. Berg, H. Peng, Y. Luo, R.C. Jankowitz, S. Wu, A deep learning method for classifying mammographic breast density categories, Med. Phys. 45 (1) (2017) 314–321, https://doi.org/10.1002/mp.12683.

[33] H. Chougrad, H. Zouaki, O. Alheyane, Deep convolutional neural networks for breast cancer screening, Comput. Methods Progr. Biomed. 157 (2018) 19–30, https://doi.org/10.1016/j.cmpb.2018.01.011.

[34] R. Platania, S. Shams, S. Yang, J. Zhang, K. Lee, S.-J. Park, Automated breast cancer diagnosis using deep learning and region of interest detection (bc-droid), in: Proceedings of the 8th acm international conference on bioinformatics, computational biology, and health informatics, 2017, pp. 536–543, https://doi.org/10.1145/3107411.3107484.

[35] R. Agarwal, O. Diaz, M.H. Yap, X. Llado, R. Marti, Deep learning for mass detection in full field digital mammograms, Comput. Biol. Med. 121 (2020) 103774, https://doi.org/10.1016/j.compbiomed.2020.103774.

[36] W. Zhu, Q. Lou, Y.S. Vang, X. Xie, Deep multi-instance networks with sparse label assignment for whole mammogram classification, in: In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2017, pp. 603–611, https://doi.org/10.1007/978-3-319-66179-7_69.

[37] M. Yousefi, A. Krzyżak, C.Y. Suen, Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning, Comput. Biol. Med. 96 (2018) 283–293.

[38] L. Sun, J. Wang, Z. Hu, Y. Xu, and Z. Cui, "Multi-view convolutional neural networks for mammographic image classification," *IEEE Access*, vol. 7, pp. 126273–126282, 2019, doi: 10.1109/ACCESS.2019.2939167.

[39] J. Wang, Y. Yang, A context-sensitive deep learning approach for microcalcification detection in mammograms, Pattern Recogn. 78 (2018) 12–22, https://doi.org/10.1016/j.patcog.2018.01.009.

[40] L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. Mcbride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography, Sci. Rep. 9 (no. 1) (2019), https://doi.org/10.1038/s41598-019-48995-4.

[41] J. Ge, et al., Computer aided detection of clusters of microcalcifications on full field digital mammograms, Med. Phys. 33 (8) (2006) 2975–2988, https://doi.org/10.1118/1.2211710.

[42] S. Guan, M. Loew, Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks, J. Med. Imag. 6 (3) (2019) 1–10, https://doi.org/10.1117/1.JMI.6.3.031411.

[43] M. Yemini, Y. Zigel, and D. Lederman, "Detecting masses in mammograms using convolutional neural networks and transfer learning," in 2018 ieee international conference on the science of electrical engineering in israel (icsee), 2018, pp. 1–4, doi: 10.1109/ICSEE.2018.8646252.

[44] J.S. Bridle, Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition, in *Neurocomputing*, Springer, 1990, pp. 227–236.

[45] V. Nair, G.E. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines, 2010.

[46] W. Lotter, G. Sorensen, D. Cox, A multi-scale cnn and curriculum learning strategy for mammogram classification, in: In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, 2017, pp. 169–177.

[47] M.A. Al-Masni, et al., Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system, Comput. Methods Progr. Biomed. 157 (2018) 85–94, https://doi.org/10.1016/j.cmpb.2018.01.017.

[48] N. Dhungel, G. Carneiro, A.P. Bradley, A deep learning approach for the analysis of masses in mammograms with minimal user intervention, Med. Image Anal. 37 (2017) 114–128, https://doi.org/10.1016/j.media.2017.01.009.

[49] S. Li, M. Dong, G. Du, and X. Mu, "Attention dense-u-net for automatic breast mass segmentation in digital mammogram," *IEEE Access*, vol. 7, pp. 59037–59047, 2019, doi: 10.1109/ACCESS.2019.2914873.

[50] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: In Proc. Icml vol. 30, 2013, p. 3.

[51] G. Liang, et al., Joint 2D-3D breast cancer classification, in: In 2019 Ieee International Conference on Bioinformatics and Biomedicine (Bibm), 2019, pp. 692–696, https://doi.org/10.1109/BIBM47256.2019.8983048.

[52] V.K. Singh, et al., Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network, in: Expert Systems With Applications vol. 139, 2020, p. 112855, https://doi.org/10.1016/j.eswa.2019.112855.

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[54] R.K. Samala, H.-P. Chan, L.M. Hadjiiski, K. Cha, M.A. Helvie, Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis, in: In Medical Imaging 2016: Computer-Aided Diagnosis vol. 9785, 2016, https://doi.org/10.1117/12.2217092, 97850Y.

[55] T. Kooi, N. Karssemeijer, Deep learning of symmetrical discrepancies for computer-aided detection of mammographic masses, in: In Medical Imaging 2017: Computer-Aided Diagnosis vol. 10134, 2017, https://doi.org/10.1117/12.2254586, 101341J.

[56] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, 2015 arXiv preprint arXiv:1502.03167.

[57] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, L. Li, Discrimination of breast cancer with microcalcifications on mammography by deep learning, Sci. Rep. 6 (1) (Jun. 2016) 27327, https://doi.org/10.1038/srep27327.

[58] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: In 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr), 2016, pp. 770–778.

[59] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, Lect. Notes Comput. Sci. Med. Image Comput. Comput. Assisted Interv. – MICCAI 2015 (2015) 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[60] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks [Online]. Available, http://arxiv.org/abs/1608.06993, 2016.

[61] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," In 2018 Ieee/cvf Conference on Computer Vision and Pattern Recognition, vol. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[62] T.-Y. Lin, et al., Microsoft coco: common objects in context, in: In European Conference on Computer Vision, 2014, pp. 740–755.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," In 2009 Ieee Conference on Computer Vision and Pattern Recognition, vol. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[64] S.J. Pan, Q. Yang, A survey on transfer learning, in: IEEE Transactions On Knowledge and Data Engineering vol. 22, 2009, pp. 1345–1359, https://doi.org/10.1109/TKDE.2009.191, 10.

[65] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014 arXiv preprint arXiv:1409.1556.

[66] C. Szegedy, et al., Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[67] A. Geron, Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts, tools, and techniques to build intelligent systems, OReilly Media, Inc, 2019.

[68] P. Baldi, Autoencoders, unsupervised learning, and deep architectures, in: In Proceedings of Icml Workshop on Unsupervised and Transfer Learning vol. 27, 2012, pp. 37–49, https://doi.org/10.5555/3045796.3045801.

[69] A.J. Ratner, H.R. Ehrenberg, Z. Hussain, J. Dunnmon, C. Ré, Learning to compose domain-specific transformations for data augmentation, in: Advances In Neural Information Processing Systems vol. 30, 2017, pp. 3239–3249.

[70] Z. Hussain, F. Gimenez, D. Yi, D. Rubin, Differential data augmentation techniques for medical imaging classification tasks, in: AMIA... Annual Symposium Proceedings. AMIA Symposium vol. 2017, Apr. 2018, pp. 979–984.

[71] I. Goodfellow, et al., Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), In Advances in Neural Information Processing Systems vol. 27, Curran Associates, Inc, 2014, pp. 2672–2680.

[72] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015 arXiv preprint arXiv:1511.06434.

[73] M. Jaderberg, K. Simonyan, A. Zisserman, Spatial transformer networks, in: In Advances in Neural Information Processing Systems, 2015, pp. 2017–2025, others.

[74] F. Wang et al., "Residual attention network for image classification," In Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition, vol. 2017, pp. 3156–3164, doi: 10.1109/CVPR.2017.683.

[75] C. Li, et al., Multi-view mammographic density classification by dilated and attention-guided residual learning, in: IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, https://doi.org/10.1109/TCBB.2020.2970713, 1–1.

[76] J. Deng, Y. Ma, D.-a. Li, J. Zhao, Y. Liu, H. Zhang, Classification of breast density categories based on se-attention neural networks, in: Computer Methods And Programs In Biomedicine vol. 193, 2020, p. 105489, https://doi.org/10.1016/j.cmpb.2020.105489.

[77] X. Zhao, L. Yu, X. Wang, Cross-view attention network for breast cancer screening from multi-view mammograms, in: In ICASSP 2020-2020 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp), 2020, pp. 1050–1054, https://doi.org/10.1109/ICASSP40776.2020.9054612.

[78] H. Sun, C. Li, B. Liu, H. Zheng, D.D. Feng, S. Wang, AUNet: Attention-Guided Dense-Upsampling Networks for Breast Mass Segmentation in Whole Mammograms, 2018, https://doi.org/10.1088/1361-6560/ab5745.

[79] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 5998–6008.

[80] J. Suckling, et al., Mammographic Image Analysis Society (Mias) Database v1.21, 2015.

[81] M. Heath, et al., Current status of the digital database for screening mammography, Comput. Imag. Vision Digital Mammogr. (1998) 457–460.

[82] I. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. Cardoso, J. Cardoso, INbreast: toward a full-field digital mammographic database, Acad. Radiol. 19 (Nov. 2011) 236–248, https://doi.org/10.1016/j.acra.2011.09.014.

[83] M.A. Guevara Lopez, et al., BCDR: a breast cancer digital repository, in: In 15th International Conference on Experimental Mechanics, Jan. 2012, pp. 1065–1066.

[84] J. Arevalo, F.A. González, R. Ramos-Pollán, J.L. Oliveira, M.A.G. Lopez, Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Methods Progr. Biomed. 127 (2016) 248–257, https://doi.org/10.1016/j.cmpb.2015.12.014.

[85] B.R.N. Matheus, H. Schiabel, Online mammographic images database for development and comparison of cad schemes, J. Digit. Imag. 24 (3) (Jun. 2011) 500–506, https://doi.org/10.1007/s10278-010-9297-2.

[86] A. Badano, et al., Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial, JAMA Network Open 1 (7) (Nov. 2018), https://doi.org/10.1001/jamanetworkopen.2018.5474 p. e185474.

[87] A. Badano, et al., GitHub [Online]. Available, https://github.com/DIDSR/VICTRE, 2018.

[88] M.D. Halling-Brown, et al., OPTIMAM mammography image database: a large scale resource of mammography images and clinical data, *arXiv preprint arXiv*, 2004, 04742, 2020.

[89] P. Elangovan, A. Hadjipanteli, A. Mackenzie, D.R. Dance, K.C. Young, K. Wells, OPTIMAM image simulation toolbox-recent developments and ongoing studies, in: International workshop on breast imaging, 2016, pp. 668–675, https://doi.org/10.1007/978-3-319-41546-8_83.

[90] A. Alikhassi, H. Esmaili Gourabi, M. Baikpour, Comparison of inter- and intra-observer variability of breast density assessments using the fourth and fifth editions of breast imaging reporting and data system, Eur. J. Radiol. Open 5 (Apr. 2018) 67–72, https://doi.org/10.1016/j.ejro.2018.04.002.

[91] E. Sickles, D. Cj, B. Lw, et al., ACR Bi-rads Atlas : Breast Imaging Reporting and Data System, American College of Radiology, Reston, VA, 2013.

[92] K. Kerlikowske, et al., Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer risk, JNCI J. Natl. Canc. Inst. 99 (5) (Mar. 2007) 386–395, https://doi.org/10.1093/jnci/djk066.

[93] K. Kerlikowske, et al., Strategies to identify women at high risk of advanced breast cancer during routine screening for discussion of supplemental imaging, JAMA Internal Med. 179 (9) (2019) 1230–1239, https://doi.org/10.1001/jamainternmed.2019.1758.

[94] N. Wu et al., "Breast density classification with deep convolutional neural networks," In 2018 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp), vol. 2018, pp. 6682–6686, doi: 10.1109/ICASSP.2018.8462671.

[95] K. Petersen, K. Chernoff, M. Nielsen, A.Y. Ng, Breast density scoring with multiscale denoising autoencoders, in: In Lecture Notes in Computer Science: Authors' Instructions, 2012, pp. 1–8.

[96] J.T. Fung, S.W. Chan, A.N. Chiu, P.S. Cheung, S. Lam, Mammographic determination of breast volume by elliptical cone estimation, World J. Surg. 34 (7) (2010) 1442–1445, https://doi.org/10.1007/s00268-009-0283-0.

[97] J.W. Rostas, et al., Calculation of breast volumes from mammogram: comparison of four separate equations relative to mastectomy specimen volumes, J. Surg. Oncol. 117 (8) (2018) 1848–1853, https://doi.org/10.1002/jso.25076.

[98] W. Berg, Diagnostic Imaging, Elsevier, Philadelphia, 2019.

[99] Q. Cai, X. Liu, Z. Guo, Identifying architectural distortion in mammogram images via a se-densenet model and twice transfer learning, in: 11th International Congress on Image and Signal Processing, Biomedical Engineering and

Informatics (Cisp-bmei), vol. 2018, 2018, pp. 1–6, https://doi.org/10.1109/CISP-BMEI.2018.8633197.
[100] X. Yu, N. Zeng, S. Liu, Y.-D. Zhang, Utilization of densenet201 for diagnosis of breast abnormality, Mach. Vis. Appl. 30 (7) (Oct. 2019) 1135–1144, https://doi.org/10.1007/s00138-019-01042-8.
[101] Y. Shen, et al., An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization, 2020. *ArXiv*, vol. abs/2002.07613.
[102] P. Henrot, A. Leroux, C. Barlier, P. Génin, Breast microcalcifications: the lesions in anatomical pathology, Diagnos. Interv. Imag. 95 (2) (Feb. 2014) 141–152, https://doi.org/10.1016/j.diii.2013.12.011.
[103] L. Wilkinson, V. Thomas, N. Sharma, Microcalcification on mammography: approaches to interpretation and biopsy, Br. J. Radiol. 90 (1069) (Jan. 2017), https://doi.org/10.1259/bjr.20160594, 20160594.
[104] C.K. Bent, L.W. Bassett, C.J.D. Orsi, J.W. Sayre, The positive predictive value of BI-RADS microcalcification descriptors and final assessment categories, Am. J. Roentgenol. 194 (5) (May 2010) 1378–1383, https://doi.org/10.2214/ajr.09.3423.
[105] P.C. Stomper, J. Geradts, S.B. Edge, E.G. Levine, Mammographic predictors of the presence and size of invasive carcinomas associated with malignant microcalcification lesions without a mass, Am. J. Roentgenol. 181 (6) (Dec. 2003) 1679–1684, https://doi.org/10.2214/ajr.181.6.1811679.
[106] B. Savelli, A. Bria, M. Molinara, C. Marrocco, F. Tortorella, A multi-context cnn ensemble for small lesion detection, Artif. Intell. Med. 103 (2020) 101749, https://doi.org/10.1016/j.artmed.2019.101749.
[107] Y. Yin, S.V. Fotin, H. Haldankar, J.W. Hoffmeister, S. Periaswamy, Transferring learned microcalcification group detection from 2D mammography to 3D digital breast tomosynthesis using a hierarchical model and scope-based normalization features [Online]. Available, http://arxiv.org/abs/1603.05955, 2016.
[108] H. Cai, et al., Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms, Computat. Math. Methods Med. 2019 (Mar. 2019) 2717454, https://doi.org/10.1155/2019/2717454.
[109] L. van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605.
[110] J.A. Harvey, Unusual breast cancers: useful clues to expanding the differential diagnosis, Radiology 242 (3) (Mar. 2007) 683–694, https://doi.org/10.1148/radiol.2423051631.
[111] M.G. Ertosun, D.L. Rubin, Probabilistic visual search for masses within mammography images using deep learning, in: Ieee International Conference on Bioinformatics and Biomedicine (Bibm) vol. 2015, 2015, pp. 1310–1315, https://doi.org/10.1109/BIBM.2015.7359868.
[112] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection [Online]. Available, http://arxiv.org/abs/1506.02640, 2015.
[113] T. de Moor, A. Rodriguez-Ruiz, A.G. Merida, R. Mann, J. Teuwen, Automated soft tissue lesion detection and segmentation in digital mammography using a u-net deep learning network [Online]. Available, http://arxiv.org/abs/1802.06865, 2018.
[114] H. Li, S. Zhuang, D.-a. Li, J. Zhao, Y. Ma, Benign and malignant classification of mammogram images based on deep learning, Biomed. Signal Process Contr. 51 (2019) 347–354, https://doi.org/10.1016/j.bspc.2019.02.017.
[115] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision [Online]. Available, http://arxiv.org/abs/1512.00567, 2015.
[116] C. Szegedy, et al., Going deeper with convolutions, in: In 2015 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr), 2015, pp. 1–9, https://doi.org/10.1109/CVPR.2015.7298594.
[117] A.S. Abdel Rahman, S.B. Belhaouari, A. Bouzerdoum, H. Baali, T. Alam, A. M. Eldaraa, Breast mass tumor classification using deep learning, in: In 2020 Ieee International Conference on Informatics, Iot, and Enabling Technologies (Iciot), 2020, pp. 271–276, https://doi.org/10.1109/ICIoT48696.2020.9089535.
[118] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, in: IEEE Transactions On Pattern Analysis And Machine Intelligence vol. 39, 2017, pp. 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031, 6.
[119] M.D. Halling-Brown, P.T. Looney, M.N. Patel, L.M. Warren, A. Mackenzie, K. C. Young, The oncology medical image database (omi-db), Med. Imag. 2014: PACS Imag. Inform. : Next Gener. Innov. (2014), https://doi.org/10.1117/12.2041674.
[120] D. Ribli, A. Horváth, Z. Unger, P. Pollner, I. Csabai, Detecting and classifying lesions in mammograms with deep learning, Sci. Rep. 8 (1) (2018) 1–7, https://doi.org/10.1038/s41598-018-22437-z.
[121] A.S. Becker, M. Marcon, S. Ghafoor, M.C. Wurnig, T. Frauenfelder, A. Boss, Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer, Invest. Radiol. 52 (7) (Jul. 2017) 434–440, https://doi.org/10.1097/RLI.0000000000000358.
[122] M.A. Al-Antari, M.A. Al-Masni, M.-T. Choi, S.-M. Han, T.-S. Kim, A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification, Int. J. Med. Inf. 117 (2018) 44–54, https://doi.org/10.1016/j.ijmedinf.2018.06.003.
[123] R. Shen, J. Yao, K. Yan, K. Tian, C. Jiang, K. Zhou, Unsupervised domain adaptation with adversarial learning for mass detection in mammogram, Neurocomputing 393 (2020) 27–37, https://doi.org/10.1016/j.neucom.2020.01.099.
[124] H. Jung, et al., Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network, PloS One 13 (9) (2018) p. e0203355.
[125] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection [Online]. Available, http://arxiv.org/abs/1708.02002, 2017.
[126] Y. Yan, et al., Cascaded multi-scale convolutional encoder-decoders for breast mass segmentation in high-resolution mammograms, in: In 2019 41st Annual International Conference of the Ieee Engineering in Medicine and Biology Society (Embc), 2019, pp. 6738–6741, https://doi.org/10.1109/EMBC.2019.8857167.
[127] Z. Tu, X. Bai, Auto-context and its application to high-level vision tasks and 3D brain image segmentation, in: IEEE Transactions On Pattern Analysis And Machine Intelligence vol. 32, 2010, pp. 1744–1757, https://doi.org/10.1109/TPAMI.2009.186, 10.
[128] W. Sun, T.-L. Tseng, J. Zhang, W. Qian, Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data, Comput. Med. Imag. Graph. 57 (2017) 4–9, https://doi.org/10.1016/j.compmedimag.2016.07.004.
[129] T.D. Bui, S. Ravi, V. Ramavajjala, Neural graph machines: learning neural networks using graphs, vol. abs/1703.04818, CoRR (2017) 3–4 [Online]. Available, http://arxiv.org/abs/1703.04818.
[130] K. Sharma, B. Preet, Classification of mammogram images by using cnn classifier, in: In 2016 International Conference on Advances in Computing, Communications and Informatics (Icacci), 2016, pp. 2743–2749, https://doi.org/10.1109/ICACCI.2016.7732477.
[131] Y. Zhang, X. Wang, H. Blanton, G. Liang, X. Xing, N. Jacobs, 2D convolutional neural networks for 3D digital breast tomosynthesis classification, in: In *2019 Ieee International Conference on Bioinformatics and Biomedicine (Bibm)*, 2019, pp. 1013–1017, https://doi.org/10.1109/BIBM47256.2019.8983097.
[132] S.M. McKinney, et al., International evaluation of an ai system for breast cancer screening, Nature 577 (7788) (2020) 89–94, https://doi.org/10.1038/s41586-019-1799-6.
[133] J. Melnikow, et al., Supplemental screening for breast cancer in women with dense breasts: a systematic review for the us preventive services task force, Ann. Intern. Med. 164 (4) (2016) 268–278, https://doi.org/10.7326/M15-1789.
[134] F.C. Kitamura, I. Pan, T.L. Kline, Reproducible artificial intelligence research requires open communication of complete source code, Radiology: Artif. Intell. 2 (4) (2020), https://doi.org/10.1148/ryai.2020200060 p. e200060.
[135] D.C. Ince, L. Hatton, J. Graham-Cumming, The case for open computer programs, Nature 482 (7386) (2012) 485–488, https://doi.org/10.1038/nature10836.
[136] PowerLook®V2.1 fda approval letter., U.S. Food; Drug administration, center for Drug evaluation, Research (2018) 5–8.
[137] DM-Density®FDA approval letter, U.S. Food; Drug administration, center for Drug evaluation, Research (2018) 1–6.
[138] CmTriageTMFDA approval letter, U.S. Food; Drug administration, center for Drug evaluation, Research (2019) 2–5.
[139] ProFoundTMV2.1 fda approval letter, U.S. Food; Drug administration, center for Drug evaluation, Research (2019) 4–8.
[140] TransparaTM.0 fda approval letter, U.S. Food; Drug administration, center for Drug evaluation, Research (2020) 1–9.
[141] A. Rodríguez-Ruiz, et al., Detection of breast cancer with mammography: effect of an artificial intelligence support system, Radiology 290 (2) (2019) 305–314, https://doi.org/10.1148/radiol.2018181371.
[142] MammoScreenTMFDA approval letter, U.S. Food; Drug administration, center for Drug evaluation, Research (2020) 1–21.
[143] DensityaiTMFDA approval letter, U.S. Food; Drug administration, center for Drug evaluation, Research (2020) 1–6.
[144] F. Collado-Mesa, E. Alvarez, K. Arheart, The role of artificial intelligence in diagnostic radiology: a survey at a single radiology residency training program, J. Am. Coll. Radiol. 15 (12) (2018) 1753–1757, https://doi.org/10.1016/j.jacr.2017.12.021.
[145] N. Dhungel, G. Carneiro, A.P. Bradley, Automated mass detection in mammograms using cascaded deep learning and random forests, in: In 2015 International Conference on Digital Image Computing: Techniques and Applications (Dicta), 2015, pp. 1–8.
[146] T. Schaffter, et al., Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms, JAMA Netw. open 3 (3) (2020), https://doi.org/10.1001/jamanetworkopen.2020.0265 e200265–e200265.
[147] H. Iqbal, PlotNeuralNet. https://github.com/HarisIqbal88/PlotNeuralNet, 2019.